UNIVERSITÄT BONN

Generative Adversarial Networks for Semantic Image Synthesis and Unconditional Synthesis with Limited Data

DISSERTATION

zur Erlangung des Doktorgrades (*Dr. rer. nat.*)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich–Wilhelms–Universität, Bonn

vorgelegt von

VADIM SUSHKO

aus

Schukowski, Russland

Bonn, 2024

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich–Wilhelms–Universität Bonn

1. Gutachter / 1st Advisor: Prof. Dr. Juergen Gall
2. Gutachter / 2nd Advisor: Prof. Dr. Nicu Sebe
Tag der Promotion / Day of Promotion: 06.02.2024
Erscheinungsjahr / Year of Publication: 2024

Abstract

by Vadim Sushko

for the degree of

Doctor rerum naturalium

Generative modeling of images is an important task aimed at synthesizing new images that are indistinguishable from real samples. The ability to generate diverse, realistic-looking images holds numerous applications, ranging from artistic content creation to data augmentation for other computer vision tasks. In this thesis, we study generative adversarial networks (GANs), which have gained popularity as the leading image synthesis framework due to their exceptional performance. GANs consist of a generator and a discriminator that are engaged in an adversarial game. In this game, the discriminator learns to correctly classify incoming real and fake images, while the generator is trained to fool the discriminator by producing more realistic samples. Although GANs have made significant advancements in recent years, more research is needed to further enhance their scalability across diverse datasets, as well as the quality and diversity of their synthesis. To this end, this thesis presents several new GAN methods that expand the literature on GANs in various aspects.

Firstly, we address the task of semantic image synthesis, generating realistic images from semantic label maps. For this task, our work introduces a new segmentation-based discriminator that provides a strong training signal for the generator, eliminating the need for additional losses and tricks used in prior models. Compared to previous GAN approaches, our proposed method achieves a synthesis with higher image quality and diversity, while showing much better scalability to datasets with severe class imbalances.

Secondly, we explore the training of unconditional GAN models in low data regimes. Previously, under limited data, GAN models suffered from training instabilities and memorization issues, which limited their application in restricted image domains. To address this, we propose new GAN approaches for various limited-data scenarios, including traditional one-shot and few-shot learning regimes. The advancements of our methods include new training schemes, improved architectures of discriminators, and novel regularization terms for generators. Our proposed methods enable high-quality and diverse synthesis from extremely small datasets, on which prior GAN models could not be trained successfully.

Overall, this thesis advances the field of generative adversarial networks by introducing new GAN models that improve over prior work in the image synthesis quality, diversity, and applicability across various image domains. The proposed approaches demonstrate superior performance in semantic image synthesis and unconditional training with limited data, making GANs more powerful and effective for a wide range of computer vision applications.

Keywords: generative adversarial networks (GANs), image synthesis

Acknowledgements

I would like to express my heartfelt gratitude to Prof. Juergen Gall for his warm support at all stages of my PhD journey. I cannot thank him enough for his exemplary supervision, openness in exploring new research ideas, and invaluable help with paper submissions. I am especially grateful to him for organizing the research visit to the University during the third year of my PhD.

I also owe sincere gratitude to my industry supervisors Anna Khoreva and Dan Zhang. I want to thank them for their rigorous supervision and highest professional standards in all aspects of research. Under Anna's and Dan's guidance, I learned to work with more diligence and attention to detail, and to reflect critically on my own ideas. I could not be more lucky with the amount of time and effort they invested in our discussions, the development of our projects, and paper writing.

Very special thanks goes to all my colleagues at Bosch Center for Artificial Intelligence. I very much enjoyed the company of my fellow PhD students: Edgar, Nadine, Yuxuan, Alex, Yumeng, Zhak, Kanil. I would like to especially thank Edgar and Nadine for their unwavering humor, great collaborations, and for being great source of encouragement during most difficult PhD times. I also thank Ruyu for collaborations, David for code reviews, and Bill for proofreading our submissions. In addition, I also thank the BCAI Frisbee team for the great afterwork matches.

I had a pleasure of working with my fellow PhD students at the CVG group of the University of Bonn. Thank you, Andreas, Emad, Federico, Hakam, Hamid, Jinhui, Julian, Laura, Olga, Shijie, Shuai, for interesting discussions, fun memories, and great time spent together in Bonn.

Finally, I am deeply grateful to my family and friends. PhD can be a stressful endeavour, especially at the times of a global pandemic. Their unconditional love and support made it possible for me to go through the most stressful PhD times and approach the final PhD stage of the thesis writing.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

An alphabetically sorted list of abbreviations used in the thesis:

CNN	Convolutional Neural Network
DA	Differentiable Image Augmentation
DR	Diversity Regularization
DM	Diffusion Model
FA	Feature Augmentation
FID	Frechet Inception Distance
GAN	Generative Adversarial Network
IoU	Intersetion-over-Union
IS	Inception Score
LPIPS	Learned Perceptual Image Patch Similarity
mAP	Mean Average Precision
mIoU	Mean Intersetion-over-Union
MS-SSIM	Multi-Scale Structural Similarity
PPL	Perceptual Path Length
ResNet	Deep Residual Network
SIFID	Single-Image Frechet Inception Distance
VAE	Variational Autoencoder

CHAPTER 1

Introduction

Contents

1.1 Motivation

Can you imagine a giant city floating in the clouds, entirely made of glass, without any visible support structures? Although this concept obviously violates the laws of physics, it is not surprising that you can imagine such a structure and reproduce it on a piece of paper. Indeed, as humans, we often hallucinate objects we have never seen before or scenarios that have never occurred to us. We use such ability for a variety of purposes, including artistic expression, problem-solving, teaching, innovation, or simply for entertainment. In the field of artificial intelligence, researchers are therefore trying to develop algorithms that can replicate the human ability to create new images. This problem, known as generative modeling of images, has gained significant attention from the research community in recent years. The objective of this task is to train a machine learning model, usually a neural network, to generate novel images that follow the distribution of a given training dataset. The potential applications of such models are numerous, ranging from generating scenery for video games or movies and new designs for products or buildings, to converting images to new artistic styles and producing synthetic data augmentation for other computer vision applications.

Training powerful generative models is, unfortunately, not straightforward. Unlike discriminative tasks such as image classification or segmentation, generative models are typically more challenging to train in the paradigm of supervised learning, mainly because the concept of "realism" in generated images is difficult to formalize. For example, the most obvious approach – forcing the generated

Figure 1.1: These people are not real, they were generated by StyleGANv2 (*Karras et al.*, 2020b). GANs excel with human faces due to their simplicity and ample data, but struggle with more complex, imbalanced, or small datasets. This thesis aims to enhance GANs' capabilities for image generation in diverse image domains, as well as their applicability to new downstream applications.

images to be similar to training examples via a reconstruction loss – can encourage the model to produce only the images that it has already seen during training, which is undesirable, since the ultimate goal of generative models is to be "creative".

To overcome these issues, the computer vision community has been working to develop alternative approaches. In this thesis, we focus on a special class of generative models – generative adversarial networks (GANs) (*Goodfellow et al.*, 2014), a very powerful class of models that have demonstrated state-of-the-art performance across various image generation tasks. GANs consist of two neural networks, a generator and a discriminator, that are trained simultaneously in a game-like setup. The goal of the generator is to transform provided input noise vectors into realistic images that the discriminator would judge as real. In turn, the discriminator tries to distinguish between the synthesized fake images and the real images from the dataset. Through this game, the generator optimizes the objective that directly forces the output to be "indistinguishable from reality", without an explicit goal to reconstruct any of the training images. After training, the generator allows to sample many new images by varying the input noise. This training scheme has allowed to generate diverse and highly-realistic images in many image domains. A famous example of this ability is the domain of human faces, which is shown in Fig. 1.1. In this example, all the presented human faces are not real, but instead were generated by a GAN model called StyleGANv2 (*Karras et al.*, 2020b). The ability to generate new, diverse, and highly-realistic human faces at high resolutions has enormous potential for a variety of applications, including in fashion, cosmetics, photography enhancement, privacy protection, or virtual reality. Remarkably, this has already led to the creation of several popular online services, such as *this-person-does-not-exist.com*, which provide synthetic human faces that can be used for downstream applications without concerns for the privacy of the original images.

While such success of GANs is impressive, it is important to note that it heavily relies on the availability of high-quality, large-scale datasets, which require careful curation. For instance, in order to build the FFHQ[1] dataset used to generate the human faces shown in Fig. 1.1, researchers had to crawl Flickr[2] to find images containing human faces, apply a series of filters to remove poor-quality images, and manually check every image for inappropriate content or copyright violations using Amazon Mechanical Turk[3]. The resulting dataset contains 70,000 high-quality images of centered human faces at resolution of 1024×1024 pixels.

However, it is evident that not all applications allow to collect well-balanced datasets of such

[1] https://github.com/NVlabs/ffhq-dataset
[2] https://www.flickr.com/about
[3] https://www.mturk.com/

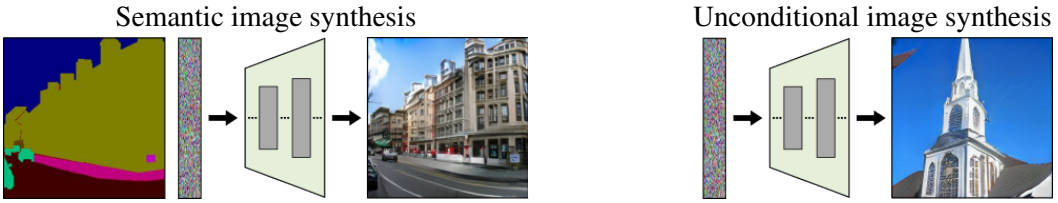

Figure 1.2: Two main tasks of interest in this thesis. In semantic image synthesis, a GAN generator is conditioned on input noise and semantic label maps. In unconditional image synthesis, the generator is conditioned only on noise. For the latter, we study specific scenarios where training data is limited.

size, resolution, and quality. In most cases, real-world datasets are constrained by factors such as privacy concerns, application design, or the costs associated with data collection and annotation. For example, in industrial applications such as the detection of manufacturing defects, only a few images of defective details may be available, and each of these images may bear a striking similarity to others. Alternatively, in large-scale image collections encompassing many object categories, it is natural to observe a small number of frequently occurring classes alongside a long tail of exceedingly rare classes (*Gupta et al.*, 2019). As demonstrated throughout this thesis, irregular data conditions impede the effectiveness of GANs, and achieving a synthesis with satisfactory image quality and diversity becomes challenging. Consequently, the applicability of GANs has been severely hindered, relegating them primarily to "good" types of images like in the domain of human faces.

Therefore, in this thesis we aim to broaden the scope of potential applications for GANs while enhancing their performance in terms of image synthesis quality and diversity. We make contributions to several important image generation tasks, which can be categorized into two main areas (see Fig. 1.2). The first task focuses on semantic image synthesis, which aims to generate realistic and diverse images from provided input noise vectors and semantic label maps (Fig. 1.2, left). The second area of our study is unconditional image synthesis (Fig. 1.2, right), with a specific focus on scenarios where training data is limited. In both domains, we introduce novel GAN models that surpass the existing state of the art in the GAN literature. These models not only enable new applications of image synthesis but also overcome several challenges faced by previous GAN methods. The two subsequent sections provide a detailed discussion of these challenges and our corresponding contributions.

1.2 Challenges

The goal of generative adversarial networks is to learn the distribution of provided images and train a generator with several desirable properties. **(1):** The first objective is to generate realistic images that belong to the distribution of real images, without any distortions that would visually differentiate them from the images in the training dataset. **(2):** The second objective is to ensure that the generated images exhibit diversity, covering all the different modes of the training distribution rather than concentrating on a limited types of images. **(3):** Finally, in numerous applications, it is essential that the generated images are distinct enough from the training examples to avoid mere repetition. Achieving these goals involves several challenges, many of which are characteristic to all GANs due to their architecture and optimization procedures.

General Challenges of Training GANs

(1): Image quality. From the optimization perspective, training a GAN model involves solving a minimax problem, where the discriminator maximizes its objective in distinguishing between real and fake images, while the generator attempts to minimize it. Due to the non-convex nature of this loss function and the dynamic adversarial process, the optimization of GANs is notoriously unstable. Usually, it requires a careful balance between the two players. For example, if the GAN discriminator becomes too strong, its gradient updates may become too small or vanish, causing the generator to stop learning. In this case, the training saturates at the point when the quality of produced images is still far from optimal. On the other hand, a very strong generator leads to difficulties in the learning of the discriminator, which in turn impedes the overall training and often collapses the generator. Therefore, achieving realistic image synthesis requires a delicate balancing of the adversarial process, which requires extensive experimentation to find optimal hyperparameters such as learning rates, batch size, or network architecture. These parameters often depend on the dataset type, its size, and the image resolution used during training, which makes achieving good image quality with GANs both challenging and time-consuming in practice.

(2): Synthesis diversity. Another problem of GANs is that their training procedure generally discourages the diversity of synthesis. As the GAN generator is rewarded for producing samples that the discriminator cannot distinguish from real samples, it naturally focuses on the samples that were successful in fooling the discriminator in the past. In contrast, less successful modes of generated images receive less attention and disappear from the generated distribution. In the worst-case scenario, the generator may become so restricted that it generates only a single image that it believes is the most realistic, resulting in mode collapse. Mitigating this problem requires the tuning of hyperparameters to reach more optimal training dynamics and employing additional regularization losses to encourage diversity in generated images.

(3): Memorization of training data. Training a GAN typically takes hundreds or thousands epochs until convergence. Having seen the training images many times, the GAN discriminator becomes prone to overfitting, which leads it to judge about the realism of generated images based only on their similarity to any of the training images. Naturally, this forces the generator to produce only the exact copies of training examples rather than create new images. This is undesirable, since the ultimate goal of generative models is to be "creative" and to produce novel images, which is especially important in applications in which generated images are used as synthetic data augmentation. The problem of memorization is naturally amplified in situations when the available training datasets are very small, which severely limits the applications of GANs in restricted image domains.

The contributions of this thesis are concerned with overcoming all the above challenges in different image generation tasks. In addition to these general GAN challenges, each studied task has its own characteristic challenges. They are summarized in the next sections.

1.2.1 Semantic Image Synthesis

GAN models offer the flexibility to condition their output on specific information, allowing greater control over the generation process. A prominent example of this is the task of semantic image synthesis, which aims to generate images that are aligned with provided semantic label maps (see Fig. 1.2, left). The limitations of previous semantic image synthesis models are as follows.

Reliance on the perceptual loss. In order to reach good image quality, previous models for

semantic image synthesis have to be trained with a perceptual loss (*Wang et al.*, 2018a) in addition to the standard adversarial loss from the discriminator. The perceptual loss is computed by passing images through a deep neural network pre-trained on ImageNet (*Deng et al.*, 2009), with the aim to bring the extracted features of real and fake images closer together. The perceptual loss is critical for existing methods, and poor results are obtained without it. Although this loss improves the performance of previous methods, unfortunately, it has several disadvantages, as we demonstrate in Chapter 4. Firstly, the perceptual loss can bias the training signal with the color and texture statistics learned by the perceptual network on ImageNet, limiting the GAN's ability to learn the target dataset's colors and textures. Secondly, the perceptual loss imposes additional constraints on the feature space of the GAN generator, significantly limiting the diversity of its synthesis. Lastly, as the the perceptual loss utilizes an extra network, it introduces a noticeable computational overhead to the training process.

Insensitivity to input noise. In semantic image synthesis, one label map can correspond to multiple plausible generated images. It is therefore desired for the generator to produce diverse images from the same label map by varying its input noise vector. However, previous models struggle to achieve this goal by simply resampling noise, as the generator often disregarded the noise and instead concentrated solely on the provided semantic label maps. As a result, previous models resorted to using additional image encoders to change the style of produced images based on reference images. This is an expensive solution as it requires to train an additional network and necessitates a new reference image for every generated sample. Moreover, the diversity of this sampling method is limited, partially owing to the perceptual loss that is needed to train prior methods.

Imbalanced datasets. GAN models for semantic image synthesis are trained with datasets that were originally introduced for semantic segmentation. A well-known challenge for semantic segmentation applications is class imbalance, which leads to suboptimal performance of models (*Sudre et al.*, 2017). As shown in Chapter 4, this problem becomes similarly important in semantic image synthesis, where GAN models perform best for well-represented classes, while the quality of colors and textures of rare semantic classes remains poor. Overcoming the problem of learning from class-imbalanced datasets is a crucial step to enable the application of GAN-based data augmentation for semantic segmentation models.

1.2.2 Training GANs in Extremely Low Data Regimes

The quality of GAN-based image synthesis strongly depends on the amount and quality of data provided for training. GANs achieve impressive results in many image domains due to the availability of large, diverse datasets with thousands of images. However, in some applications, collecting even a small dataset can be challenging due to privacy, copyright, or rare nature of objects or events. Therefore, improving the training procedure of GANs in very low data regimes can enhance their utility for applications. This requires overcoming several limitations of prior GAN models.

Training instabilities. Even when training data is sufficient, GAN optimization is notoriously unstable, frequently displaying oscillatory behaviour and escalated gradients of the discriminator (*Arjovsky and Bottou*, 2017; *Mescheder et al.*, 2018). In much lower data regimes, this problem is naturally amplified due to a quick overfitting of the discriminator to limited data. In fact, even training a large-scale GAN model with only 5000 images often results in the early generator's collapse. When the number of training images is 2-3 orders of magnitude lower, successfully training such

models becomes almost impossible. Solving this problem requires a careful redesign of existing GAN architectures and training losses.

Memorization. Even in case when the GAN training on limited data is stabilized, it remains challenging to address the issue of the discriminator's overfitting. Typically, after several training epochs, the discriminator memorizes the limited dataset and starts to penalize even minor deviations from the training data in generated images. As a result, the generator converges to a state when it reproduces exact copies of the training samples. One potential solution to this issue is to weaken the discriminator by limiting its receptive field, learning capacity, or providing additional tasks beyond evaluating the proximity between real and fake images, which is commonly done in few-shot GAN models. However, this can negatively impact the quality of the generated images, and does not solve the issue completely when the lack of data is extreme, as is shown in Chapter 5.

Distortions in patch-based GANs. Patch-based single-image GANs (*Shaham et al.*, 2019; *Hinz et al.*, 2021) appeared as a remedy to overcome memorization in the extreme case of training models on a single image. These models employ a cascade of multi-scale GANs trained in multiple stages, each having a different resolution and receptive field. Although they have been demonstrated to overcome memorization and generate diverse versions of the provided single image, they still cannot learn the high-level semantic features of image scenes. Consequently, they often suffer from incoherent shuffling of image patches, resulting in distortions in the appearance of objects and scene layouts. Moreover, the design of this solution prevents patch-based GANs to learn from multiple images. Therefore, alternative solutions have to be designed to enable effective training of GANs with extremely limited data.

1.2.3 Generating Data Augmentation for One-Shot Segmentation Applications

Even though GANs often excel at generating images that look very realistic to a human eye, it is not straightforward whether these images can be used as additional data for other computer vision applications. Firstly, data augmentation is most valuable when the applications themselves have insufficient data for training the models. Apart from providing only limited data available for training a GAN model, many applications require annotations for each image during training. In Chapter 6, we will explore the application of GANs to one-shot segmentation, a class of limited-data tasks that requires segmentation masks for each training image. To generate valuable data augmentation for such applications, it is necessary not only to overcome the challenges related to training on limited data, but also to learn to generate accurate segmentation masks for synthesized images. To achieve this goal, the challenges are as follows.

Training a GAN on a single image-mask pair. One-shot segmentation models are commonly trained using the meta-learning paradigm, which involves pre-training on a large dataset, and the fine-tuning stage which uses a single provided image-mask pair depicting previously unseen objects. As the training data at inference stage is severely limited, synthetic data augmentation has a significant potential to improve the performance. However, achieving this requires training a GAN to generate not only images, but also their accurate segmentation masks, using only a single labelled example for training. As shown in Chapter 6, training a GAN on a single image-mask pair is even more difficult than on a single image, as models become more susceptible to memorization due to a smaller variation in the content of segmentation masks compared to RGB images. In fact, due to the difficulty of the task for prior image-mask GANs, this problem remained unaddressed prior to our work.

Need for additional supervision. In order to generate segmentation masks alongside images, the GAN generator needs guidance on how to align the masks with the images. However, as images generated by GANs do not have ground truth annotations, finding the source of such supervision is not straightforward. To address this issue, existing approaches use external pre-trained segmentation networks, employ additional discriminators that judge the image-mask realism jointly, or manually annotate a small set of generated images. Although these solutions do allow to supplement generated images with segmentation masks, they introduce significant overhead to the adversarial GAN training process and are not effective for learning new classes with limited data. As such, they are not well-suited for one-shot segmentation applications as a data augmentation generation tool.

1.2.4 Usage of Pre-Trained GANs for Few-Shot Image Synthesis

As training GAN models with limited data from scratch suffers from several drawbacks, it is natural to mitigate them with the help of transfer learning. This line of research, referred to as few-shot GAN adaptation, pre-trains GAN models on large datasets and fine-tunes them on smaller datasets of interest. This approach commonly improves the quality and diversity of images generated by GANs in restricted image domains, but also comes with a major limitation.

Requirement on source-target proximity. The difficulty of fine-tuning a GAN on a small dataset are similar to the ones described in Sec. 1.2.2: training instabilities and memorization issues. For example, the model can forget about its initial knowledge very quickly, falling into simply repeating the training samples from the small target dataset. To avoid this issue, existing approaches for few-shot GAN adaptation commonly employ additional regularization losses to maintain the realism and diversity of images from the source dataset. While this method works well when the target domain closely resembles the source dataset, its performance drastically degrades when the source and target domains are not so restrictively similar. For example, in Chapter 7 it will be demonstrated that prior methods struggle to generate high-quality images when the shapes of objects in the two domains do not match. This poses a significant challenge to the use of GANs in restricted image domains, where large pre-training datasets are practically impossible to find.

1.3 Contributions

This thesis focuses on solving the challenges highlighted in Sec. 1.2. We provide contributions to several GAN-based image generation tasks, mainly focusing on improving the quality and diversity of semantic image synthesis, as well as unconditional image synthesis in various extremely low data regimes. Following the order in which the challenges in Sec. 1.2.1-1.2.4 were introduced, in Sec. 1.3.1-1.3.4 we present our contributions and approaches to the challenges.

1.3.1 OASIS Model for Semantic Image Synthesis

In Chapter 4, we introduce our OASIS model, which effectively addresses several issues of previous semantic image synthesis GANs, such as overreliance on the perceptual loss, limited diversity of multi-modal synthesis, and inadequate synthesis of rare classes in datasets with class imbalances. Parts of this chapter have been published in (*Schönfeld et al.*, 2021) and (*Sushko et al.*, 2022).

The main contribution of the OASIS model is a segmentation-based discriminator, which segments each pixel of a given image into one of the real classes or an additional fake class, instead of

making one global real/fake decision for the whole image. This discriminator brings much stronger supervision about image realism than previous discriminators, which makes the use of the perceptual loss superfluous. In addition, the OASIS discriminator allows to introduce a new LabelMix regularization, which mixes semantic areas from both real and fake images. This regularization forces the discriminator to learn stronger and more local representations. Overall, these innovations significantly improve the quality of semantic image synthesis.

The second major contribution of OASIS is a novel 3D noise injection scheme. In this scheme, we sample a 3D noise tensor and use it to modulate the intermediate generator features at different layers. The benefit of using 3D noise is that it allows not only global image resampling, but also changing only the areas belonging to selected semantic classes. Apart from improving the controllability of image resampling, this scheme significantly improves sensitivity of the generator to input noise and, therefore, overall diversity of multi-modal synthesis.

With the proposed modifications in the discriminator and generator design, our OASIS model outperforms the prior state of the art on several benchmark datasets. Omitting the necessity of the VGG perceptual loss, our model generates samples of higher quality and diversity, and follows the color and texture distributions of real images more closely.

Another contribution of our work considers the LVIS dataset (*Gupta et al.*, 2019), which had not been considered before our work in the context of semantic image synthesis. LVIS has a large set of more than 1000 semantic classes, most of which are severely underrepresented. For comparison, the largest dataset among standard benchmark datasets, COCO-Stuff (*Caesar et al.*, 2018), has only 184 classes, which are relatively well balanced. As demonstrated in our experiments on LVIS, prior work performs poorly in case of severe class imbalances and suffers from mode collapse. In contrast, our OASIS model overcomes these issues and outperforms prior work by a large margin.

Finally, we introduce a new evaluation protocol for semantic image synthesis, which considers the performance of generated images as synthetic data augmentation in semantic segmentation applications. This performance depends on many important synthesis aspects, including image quality, diversity, label map alignment, and degree of memorization. OASIS outperforms prior work in this measure, showing higher utility for downstream semantic segmentation applications.

1.3.2 SIV-GAN Model for Image Synthesis in Extremely Low Data Regimes

In Chapter 5, we introduce a new limited-data GAN training regime and a new SIV-GAN model that overcomes training instabilities, memorization, and the issue of global scene incoherence when trained in extremely low data regimes. Parts of this chapter have been published in (*Sushko et al.*, 2021a).

Previous studies on GAN training in low data regimes focused on using single-image datasets or few-shot datasets consisting of at least 100 diverse images. Our first contribution is a new setup where a GAN is trained on approximately 100 frames extracted from a single video. The goal in this setup is to generate diverse images, but not necessarily coherent videos. This setup is interesting for training unconditional GANs because short videos are easy to capture and offer a practical solution for collecting data required for successful GAN training. Additionally, compared to few-shot datasets, the frames in a video exhibit lower diversity due to high correlation between adjacent video frames. Our experiments show that previous single-image and few-shot GAN models are ineffective in this scenario, making it an interesting benchmark for evaluation.

The next two contributions are concerned with our new SIV-GAN model. One of them is the SIV-GAN two-branch discriminator architecture. This discriminator has two branches which separately judge the realism of image content and layout. The content branch evaluates the fidelity of scene objects irrespective of their spatial arrangement, while the layout branch looks only at the global scene coherency. With this disentanglement, we solve the problem of memorization of the entire image, providing the generator with informative signals even when training data is extremely limited.

The third contribution is a diversity regularization for the generator developed for extremely low data regimes. This regularization forces the images produced from different latent codes to be different. In contrast to diversity losses from previous methods, our approach is not dependent on the distance between generated images in the latent space and operates in the feature space of various generator layers, rather than in the image space. As a result, while prior regularizations prove ineffective for extremely low data regimes, our proposed diversity regularization ensures a high diversity among generated samples.

SIV-GAN is the first model that successfully overcomes the challenges of both the single-image and extreme few-shot settings. Despite the low data regimes, the design of our model allows it to avoid memorization and training instabilities, yet to preserve the realism objects and layouts. The exceptional diversity of synthesis of SIV-GAN from one or few images makes it well-suited for data augmentation tasks in various restricted image domains.

1.3.3 New Task: One-Shot Synthesis of Images and Segmentation Masks

Chapter 6 is devoted to generating data augmentation for one-shot segmentation applications. We introduce a novel task of one-shot joint image-mask synthesis and propose our OSMIS model as an approach for this task. Parts of this chapter have been published in (*Sushko et al.*, 2023b).

The first contribution of our work is a new evaluation protocol for GANs, in which a model is required to generate diverse and accurate paired image-mask data, given only a single image-mask pair for training. This setup had not been addressed in the literature, as previous image-mask GAN methods could not be trained in such low data regimes. Nonetheless, considering the recent advances of training GANs on single images, such as our SIV-GAN model, we demonstrate that this task can be accomplished by extending successful single-image GAN models to segmentation masks.

Thus, our second contribution is the OSMIS model, which enables the accurate and diverse synthesis of image-mask paired data in a one-shot regime. The main novelty of the model is a masked content attention mechanism, which integrates the learning of the image-mask alignment into the discriminator's objective. Considering the provided segmentation masks, this mechanism allows to compare the image areas belonging to the same objects in real and fake images. As a result, this allows the discriminator not only to prevent the memorization of the whole given image, but also to provide supervision for the generator to label all objects correctly. An advantage of this approach over previous image-mask methods is the purely adversarial training, without the need for manual annotation of generated images, external segmentation networks, or additional discriminators.

Finally, the third contribution of our work is the successful application of OSMIS in one-shot segmentation applications. Despite the low data regime, our model can generate accurately labeled data of sufficient diversity to provide useful data augmentation for one-shot video object segmentation (*Caelles et al.*, 2017) and one-shot semantic image segmentation (*Boudiaf et al.*, 2021). OSMIS is the first model to demonstrate the potential of synthetic data augmentation in such low-data appli-

cations.

1.3.4 Smoothness Similarity Regularization for Few-Shot GAN Adaptation

In Chapter 7, we introduce a new method for few-shot GAN adaptation, which, in contrast to previous approaches, does not require a strong similarity between the source and target domains to achieve good performance. Parts of this chapter have been published in (*Sushko et al.*, 2023a).

The main contribution of the proposed method is a new regularization term for the fine-tuning of GANs. This regularization encourages the target generator to preserve the smoothness properties of the generator that was obtained during pre-training on a larger source dataset. By enforcing similar smoothness between generators, our loss encourages the target generator to build a well-structured latent space, in which different latent space directions can lead to smooth and interpretable image transitions. The advantage of this approach is that the nature of transferred image transitions is remarkably general, making it unnecessary for the source and target image domains to be very similar. In effect, the proposed model can be applied to new few-shot image domains that were previously unreachable by previous techniques.

The second contribution of the method is a new way to compute the adversarial discriminator's loss. While conventional GAN discriminators compute the loss after the last discriminator block, our method allows the computation of the loss after *each* block. Interestingly, the discriminator learns to utilize this freedom to automatically adjust the contribution of different layers depending on the similarity between the source and target datasets. This adaptiveness is a great advantage that further enhances the ability of our method to be trained successfully in a wide range of source and target domain pairs.

1.4 Publications

Parts of this thesis are based on the following publications:

- **You Only Need Adversarial Supervision for Semantic Image Synthesis**
  Edgar Schönfeld*, Vadim Sushko*, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva
  International Conference on Learning Representations (ICLR), 2021.

- **One-Shot GAN: Learning to Generate Samples from Single Images and Videos**
  Vadim Sushko, Juergen Gall, Anna Khoreva
  IEEE Computer Vision and Pattern Recognition Conference (CVPR) Workshops, 2021.
  DOI: 10.1109/CVPRW53098.2021.00293

- **OASIS: Only Adversarial Supervision for Semantic Image Synthesis**
  Vadim Sushko*, Edgar Schönfeld*, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva
  International Journal of Computer Vision (IJCV), 2022.
  DOI: 10.1007/s11263-022-01673-x

- **One-Shot Synthesis of Images and Segmentation Masks**
  Vadim Sushko, Dan Zhang, Juergen Gall, Anna Khoreva
  IEEE Winter Conference on Applications of Computer Vision (WACV), 2023.
  DOI: 10.1109/WACV56688.2023.00622

- **Smoothness Similarity Regularization for Few-Shot GAN Adaptation**
  Vadim Sushko, Ruyu Wang, Juergen Gall
  IEEE International Conference on Computer Vision (ICCV), 2023.
  DOI: 10.1109/ICCV51070.2023.00651

(* denotes equal contribution)

1.5 Thesis Structure

The rest of this thesis is organized as follows:

- In **Chapter 2**, we provide an overview of the literature related to this thesis. We begin with the general review of works on generative adversarial networks, followed by surveys on existing approaches to our studied challenges. The chapter is concluded with the discussion of alternative image generation methods to GANs.

- **Chapter 3** formally defines the task of image generation, its relation to other machine learning tasks, and existing evaluation methods. Then, it explains the basic working principles of GANs, and introduces the building blocks of GAN architectures and training losses that are used throughout the thesis. Finally, it provides a description of applications of image synthesis studied in subsequent chapters.

- **Chapters 4-7** are based on the publications listed in Sec. 1.4 and constitute the main body of this thesis. They present in detail our contributions introduced in Sec. 1.3, aimed at resolving the challenges outlined in Sec. 1.2. In particular, **Chapter 4** is devoted to semantic image synthesis. There we introduce our OASIS model that improves over existing approaches in the quality and diversity of synthesis, while enabling new capabilities that were not explored by prior work. The subsequent three chapters study unconditional GAN training in low data regimes. **Chapter 5** introduces a new training setup of learning from frames of a single video and the SIV-GAN model, capable of learning successfully in extremely low data regimes. In **Chapter 6**, the impressive capability of SIV-GAN is extended to one-shot joint image-mask synthesis. With our proposed OSMIS model, we generate useful data augmentation that helps to improve the performance of one-shot segmentation applications. Lastly, **Chapter 7** studies few-shot adaptation of GANs. There we introduce a smoothness similarity regularization that enables effective GAN fine-tuning even between dissimilar source and target domains.

- Finally, **Chapter 8** concludes the thesis with discussions on our main contributions, remaining challenges, and the outlook of future research directions on generative modelling of images.

CHAPTER 2

Related Work

In this chapter, we discuss the relevant literature that directly relates to this thesis. Given the primary focus of our thesis on GANs, the initial section is dedicated to the discussion of the overall evolution of GANs, encompassing the development of their training procedures and architectures. The subsequent sections provide the review of the previous approaches to the specific tasks and challenges addressed in this thesis. These include semantic image synthesis, training GANs under various limited data conditions, and the joint synthesis of images and segmentation masks. Lastly, we briefly discuss alternative image generation models that are not based on adversarial training.

Contents

2.1 Development of GANs

Generative adversarial networks were first introduced in (*Goodfellow et al.*, 2014). In their pioneering work, *Goodfellow et al.* (2014) trained a generator and a discriminator, both initialized as multilayer perceptron networks, on datasets with relatively small resolutions (32×32), such as MNIST (*LeCun et al.*, 1998) and CIFAR-10 (*Krizhevsky and Hinton*, 2009). Subsequent works on GANs have primarily concentrated on scaling these models to larger, more complex datasets, and on improving the quality and diversity of produced images (see Fig. 2.1). This research mainly focused on two directions: improving the stability of training through new training schemes and losses, and developing more sophisticated network architectures. In the following, we review both these directions.

Figure 2.1: Illustration of the rapid image quality development of GAN-based face generation methods, as demonstrated in (*Hermosilla et al.*, 2021).

2.1.1 GAN Training Procedure and Losses

Adversarial losses. In the original GAN formulation (*Goodfellow et al.*, 2014), the loss function of the generator is to maximize the loss function of the discriminator, which is a simple binary cross entropy loss of its real/fake classification task. However, this loss tended to produce insufficient gradients when the predictions of the discriminator were too confident, which lead to the saturation of the training process. Thus, already in the first GAN work, the loss function was replaced by a simple negative log-likelihood of the discriminator being wrong. This loss function helped to stabilize the original GAN and solved the saturation problem, and the community has thus been referring to this formulation as non-saturating GAN loss, or NS-GAN. Since then, several alternative GAN loss functions have been proposed. For example, one proposal was to replace the binary cross-entropy loss of the discriminator with a least squares loss (LSGAN) (*Mao et al.*, 2017), which was shown to improve image quality and training stability. Another popular GAN loss function was the Wasserstein GAN (WGAN) loss (*Arjovsky et al.*, 2017). The WGAN loss uses the earth-mover Wasserstein distance between the real and generated data distributions. In contrast to the JS-Divergence minimized by NS-GAN, the Wasserstein distance yields strong and useful gradients even when the real and synthetic data have only a small overlap, e.g., at the beginning of training. Other variants of GAN losses include f-GAN (*Nowozin et al.*, 2016), EBGAN (*Zhao et al.*, 2017), Fisher-GAN (*Mroueh and Sercu*, 2017), MCGAN (*Mroueh et al.*, 2017b), Sobolev-GAN (*Mroueh et al.*, 2017a), and many other. Although the research on new GAN loss formulations was extensive and solved many issues of early GAN models, there is still no consensus in the community regarding which GAN objective performs best with most modern GAN architectures (*Mescheder et al.*, 2018; *Shannon et al.*, 2021). Interestingly, in a large study conducted by *Kurach et al.* (2019), it was concluded that the original NS-GAN should be used as the default choice for new GAN architectures. Our thesis goes in line with this study, as we use the standard non-saturating adversarial GAN loss in our models that will be introduced in Chapters 4-7.

Regularization losses. To improve the quality and diversity of generated images, as well as to address the issues of overfitting, memorization, and training instabilities, various regularizations losses were introduced for GAN training. Compared to the development of adversarial losses, regularization techniques played a bigger role in the evolution of GANs. Particularly notable among them is the class of Lipschitz regularizations, which ensures that the gradients in both the generator and discriminator do not exceed a certain threshold. This regularization technique restricts the weights of the networks from making sudden jumps between training epochs, which significantly stabilizes the overall adversarial training and minimizes the risk of training collapse. The Lipschitz regularization can be implemented in several ways, including gradient clipping (*Arjovsky and Bottou*, 2017), gradi-

ent penalties (*Gulrajani et al.*, 2017; *Mescheder et al.*, 2018), or by rescaling the weights and biases of the trained networks (*Miyato et al.*, 2018). Like many other state-of-the-art GANs, our models from Chapters 4-7 employ spectral normalization to achieve stable training.

While restricting gradient updates to enforce the Lipschitz property can help achieve optimization stability, it is not a complete solution to the problems of overfitting and mode collapse. Previous research has developed a new category of regularizations that specifically encourage high diversity among generated images to tackle these issues. These regularizations typically involve generating a batch of fake images and penalizing the generator if the produced images are too similar. For instance, *Yang et al.* (2019) proposed a new loss term to encourage the generator to produce distinct outputs depending on their input latent codes, in a way that the generated samples with closer latent codes look more similar to each other, and vice versa. Later works experimented with diversity regularizations using small perturbations in the latent space (*Zhao et al.*, 2021) or extended it to conditional tasks, such as image-to-image translation (*Choi et al.*, 2020). In Chapter 5, we demonstrate the crucial role of diversity regularization in GAN training with extremely limited data and introduce a new diversity regularization approach that outperforms prior regularizations significantly.

Another method to avoid mode collapse in GANs is by looking at image transitions while walking in the generator's space. Ideally, small shifts in the latent space of the generator should result in small (but non-zero) changes in the output images. In this case, the generator avoids quick jumps and has to generate smooth and realistic transitions between all generated images. Thus, to enforce the smoothness and good conditioning of the generator's mapping, prior work proposed several regularizations. *Odena et al.* (2018) proposed to penalize the generator against having a large conditioning number, which is defined as the ratio between the maximal and minimal spectral values of its Jacobian matrix. More recently, StyleGANv2 (*Karras et al.*, 2020b) introduced a regularization based on the perceptual path length measure (PPL) (*Karras et al.*, 2019), which encourages that a fixed-size step in the latent space results in a fixed-magnitude change in the image space. A smoothness regularization was also introduced for training GANs in low data regimes (*Kong et al.*, 2022). With the help of the above regularizations, modern GAN models often display smooth and realistic latent space interpolations. In Chapter 7, we explore the methods to transfer this property during fine-tuning of GANs on extremely small datasets.

2.1.2 GAN Architectures

Unconditional GANs. The development of GAN training losses and regularizations has been accompanied by the emergence of novel GAN architectures. The original GAN formulation from *Goodfellow et al.* (2014) used multilayer perceptron networks for both the generator and discriminator, which was suitable for modeling simple grayscale datasets of digits or faces (see leftmost image in Fig. 2.1), but not complex datasets. DCGAN (*Radford et al.*, 2016) was the first fully-convolutional GAN architecture that employed convolutions and transposed convolutions in the discriminator and generator, BatchNorm layers, and LeakyReLu activations. This allowed for the synthesis of still low-resolution but more complex scenes, such as bedrooms. DCGAN became a building block for many other GAN models (*Mao et al.*, 2017; *Gulrajani et al.*, 2017; *Miyato et al.*, 2018). However, scaling these models to high image resolutions was prone to instabilities, as at higher resolutions the discriminator could easier distinguish between real and fake images. One way to bypass this issue was to use progressive growing (*Karras et al.*, 2018), gradually increasing the generated image size

starting from a low-resolution image and adding more layers as the training progresses. Alternatively, MSG-GAN (*Karnewar and Wang*, 2020) used skip-connections between intermediate generator and discriminator blocks to facilitate gradient flow to lower image resolutions. High-resolution image synthesis also posed the challenge of balancing global and local realism, as generated images looked realistic at first glance but had artifacts at the local level or incoherent details in different parts of images. To address this problem, some researchers introduced a self-attention mechanism to the generator (*Zhang et al.*, 2019) or a segmentation-based discriminator (*Schönfeld et al.*, 2020), which helped to achieve high-quality synthesis on complex datasets (e.g., photos of different animals) at 128×128 or even 256×256 resolution.

The field of generative adversarial networks was greatly influenced by a series of works on StyleGANs (*Karras et al.*, 2019, 2020b, 2021). The first StyleGAN (*Karras et al.*, 2019) introduced a style-based generator that separates the generation of image content from image style. This generator introduces an additional latent space of style codes that are used to modulate its intermediate features. In effect, this architecture enabled stable training at high resolutions, allowing for high-quality and diverse synthesis with a high degree of control over the style of images. The next model StyleGANv2 (*Karras et al.*, 2020b) redesigned the normalization inside the convolutional blocks and revisited the progressive growing mechanism. Lastly, *Karras et al.* (2021) observed a strong aliasing in the first two StyleGANs and designed their alias-free StyleGANv3 generator by treating images as continuous signals, accordingly redesigning convolution layers, nonlinearities, and upsampling layers. Overall, all StyleGANs showed state-of-the-art performance on numerous image synthesis benchmarks at the time of their publishing, so their architectures are a popular choice for researchers experimenting with new GAN features. We follow this trend in Chapter 7, where the StyleGANv2 backbone is used to test our regularization loss for few-shot adaptation of pre-trained GAN models.

While StyleGANs showed the usefulness of style-content separation for unconditional image generation, it is noteworthy that similar separations were also explored in other contexts with GANs (*Huang et al.*, 2018; *Wu et al.*, 2019; *Park et al.*, 2020a). For example, such works attempt to disentangle the parts of latent space responsible for the appearance of objects in the scene (e.g., their shape) and their style (e.g., colors and textures). This was shown useful in representation learning (*Mathieu et al.*, 2016), improving diversity in unconditional image generation (*Wu et al.*, 2019), or in image editing, when a generator is provided with style and content vectors of two different images (*Park et al.*, 2020a). In Chapter 5, we also explore how to disentangle the learning of content and layouts of images. In contrast to these related works, we study this separation in order to mitigate memorization of training data in low data regimes, and do so in a GAN discriminator, not generator.

Conditional GANs. Our thesis is also related to conditional GAN models. Unlike unconditional GANs, that take only a noise vector as input (see Fig. 2.2), conditional GANs (*Mirza and Osindero*, 2014) can produce images that adhere to provided class labels (*Brock et al.*, 2019; *Casanova et al.*, 2021), text descriptions (*Zhang et al.*, 2018a, 2021a; *Sauer et al.*, 2023), semantic label maps (*Park et al.*, 2019b; *Wang et al.*, 2021b; *Shijie et al.*, 2022), or other images (*Isola et al.*, 2017; *Park et al.*, 2020b; *Choi et al.*, 2020). Among these, our thesis is most related to two conditional tasks: GANs conditioned on semantic label maps and class labels. We provide a review on GAN synthesis from semantic label maps in Sec. 2.2, while the discussion on class-conditional GANs follows next.

The earliest GAN model conditioned on class labels is the Conditional GAN (cGAN) (*Mirza and Osindero*, 2014), which is an extension of the original GAN framework from *Goodfellow et al.* (2014), simply taking a class label as input for both the generator and discriminator. Since this work,

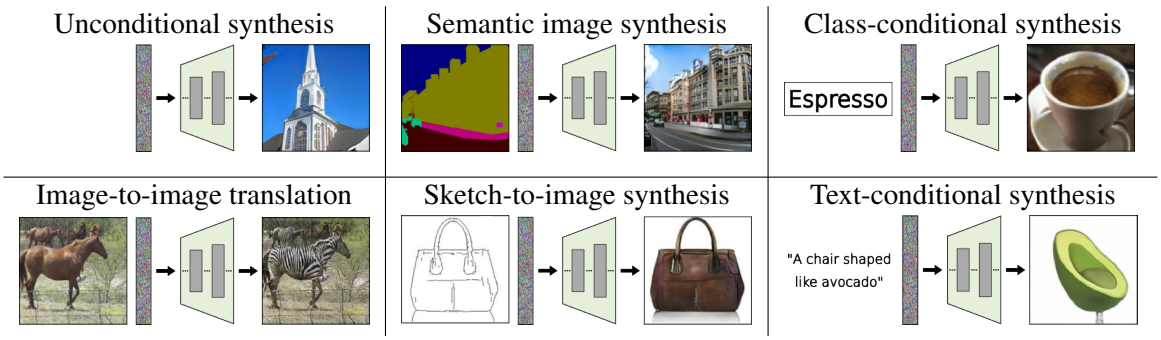

Figure 2.2: Overview of unconditional and conditional image synthesis tasks solved with GANs. In this thesis, we cover unconditional, class-conditional, and GANs for semantic image synthesis.

cGANs have quickly improved in quality. On the one hand, this was enabled by borrowing all the advances in architectures and training losses from literature on unconditional GANs. On the other hand, this required designing architectures that made use of the provided labels more effectively. *Odena et al.* (2017) proposed an Auxiliary Classifier GAN (ACGAN) that extends the cGAN framework by adding an auxiliary classifier in the discriminator network (see Fig. 2.3). The auxiliary classifier is trained to predict the class label of the generated sample, in addition to its realism. This additional supervision helps to improve the quality of generated samples and ensures that they are consistent with the provided class label. Later, *Miyato and Koyama* (2018) proposed to input class labels to the discriminator via a linear embedding layer, which projects the class information onto the discriminator's features before the final classification layer. While these works concentrated on discriminator's conditioning, BigGAN (*Brock et al.*, 2019) introduced an effective mechanism for the conditioning of the generator, in which a class label is first mapped into an embedding space via a learnable mapping, concatenated to the input noise, and injected in the generator at every layer. Thanks to the efficient usage of class labels and large-scale architecture, BigGAN became the first model to enable high-quality synthesis from class-conditional ImageNet (*Deng et al.*, 2009) at resolutions 256×256, 512×512, and 1024×1024. Since then, BigGAN became the most popular model for class-conditional synthesis, inspiring many later models (*Casanova et al.*, 2021; *Sauer et al.*, 2022; *Hou et al.*, 2022; *Zhou et al.*, 2021; *Zhao et al.*, 2020b). In this thesis, we use the projection-based discriminator from *Miyato and Koyama* (2018) as a baseline for assessing our proposed approaches in Chapters 4 and 6, and explore our proposed regularization term for few-shot GAN adaptation (Chapter 7) with the help of the BigGAN (*Brock et al.*, 2019) backbone architecture.

2.2 GANs for Semantic Image Synthesis

The task of semantic image synthesis corresponds to conditional image generation, where a model is conditined on semantic label maps. Since the introduction of the first GAN model for image-to-image translation (*Isola et al.*, 2017), GANs have emerged as dominant approach for this task. To achieve the synthesis of images that are realistic, diverse, and well-aligned to semantic label maps, researchers designed specialized generator and discriminator architectures that incorporate label maps into training. A specific feature of semantic image synthesis models is the perceptual loss, which is used in addition to the discriminator loss to train the generator. Next, we review these

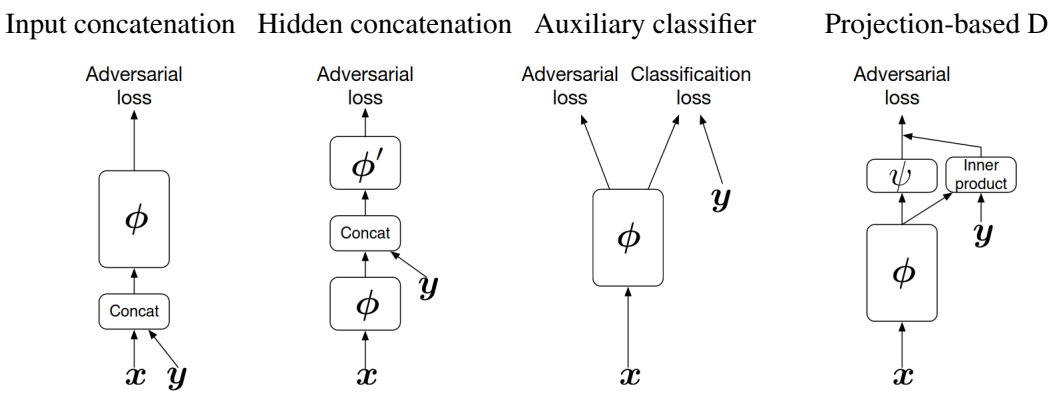

Figure 2.3: Different ways to condition a GAN discriminator on labels y, as presented in (*Miyato and Koyama*, 2018).

previous approaches, which constitute the related work for our OASIS model introduced in Chapter 4.

2.2.1 Generator Architectures

To enforce the alignment between the generated images and the conditioning label maps, previous methods for semantic image synthesis explored different ways to incorporate the label maps into the generator. In many early approaches (*Isola et al.*, 2017; *Wang et al.*, 2018a; *Tang et al.*, 2020c,b; *Ntavelis et al.*, 2020; *Richardson et al.*, 2021), label maps are provided to the generator via an additional encoder network. However, this solution has been shown to be suboptimal at preserving the semantic information until the later stages of image generation. For this reason, SPADE (*Park et al.*, 2019b) introduced a spatially-adaptive normalization layer that directly modulates the label map onto the generator's hidden layer outputs at various scales. Alternatively, CC-FPSE (*Liu et al.*, 2019) proposed to use spatially-varying convolution kernels conditioned on the label map. After the publication of our OASIS, SC-GAN (*Wang et al.*, 2021b) proposed to utilize label maps as input to generate class-specific semantic vectors at different scales, which are used as conditioning at different layers of the image rendering network; and CollageGAN (*Li et al.*, 2021b) proposed to extract a label map representation via feature pyramid encoder and inject it as spatial style tensor to a StyleGAN2 generator.

Noise injection. While improving the quality of generated images, the models published before OASIS struggled to achieve diverse results by sampling the input noise, as the generator tended to become insensitive to noise or achieved only poor quality, as first observed by *Isola et al.* (2017). Thus, these previous approaches resorted to having an image encoder in the generator design to enable multi-modal synthesis. The generator then combined the extracted image style with the label map to reconstruct the original image. By alternating the style vector, one can generate multiple outputs conditioned on the same label map. However, using an image encoder is a resource-demanding solution. In our OASIS model, we enable multi-modal synthesis directly through sampling of a 3D noise tensor which is injected at every layer of the network. Differently from the structured noise injection of *Alharbi and Wonka* (2020) and class-specific latent codes of *Zhu et al.* (2020), OASIS injects the 3D noise along with label maps and adjust it to image resolution, which also enables re-sampling of selected semantic segments.

2.2.2 Discriminator Architectures

To provide a powerful guiding signal to the generator, a GAN discriminator for semantic image synthesis should evaluate both the image realism and its alignment to the provided semantic label map. Thus, a fundamental question is to find the most efficient way for the discriminator to utilize the given semantic label maps. To this end, Pix2pix (*Isola et al.*, 2017), Pix2pixHD (*Wang et al.*, 2018a) and SPADE (*Park et al.*, 2019b) rely on concatenating the label maps directly to the input image, which is fed to a multi-scale PatchGAN discriminator. Alternatively, SESAME (*Ntavelis et al.*, 2020) employed a projection-based discriminator (*Miyato and Koyama*, 2018), applying an additional branch to process semantic label maps separately from images, and merging the two streams before the last convolutional layer via a pixel-wise multiplication. CC-FPSE (*Liu et al.*, 2019) proposed a feature-pyramid discriminator, embedding both images and label maps into a joint feature map, and then consecutively upsampling it in order to classify it as real/fake at multiple scales. LGGAN (*Tang et al.*, 2020c) introduced a classification-based feature learning module to learn more discriminative and class-specific features. In Chapter 4, we propose to use a simple pixel-wise semantic segmentation network as a discriminator instead of multi-scale image classifiers as in the above approaches, and to directly exploit the semantic label maps for its supervision. Segmentation-based discriminators have been shown to improve semantic segmentation (*Souly et al.*, 2017) and unconditional image synthesis (*Schönfeld et al.*, 2020), but have not been explored for semantic image synthesis. OASIS is therefore the first work to apply an adversarial semantic segmentation loss for this task. Noteworthy, the idea to use a segmentation-based discriminator has been adopted in several subsequent works (*Wang et al.*, 2021b; *Jain et al.*, 2022; *Jeong et al.*, 2021; *Musat et al.*, 2022).

2.2.3 Perceptual Losses

Gatys et al. (2015); *Gatys et al.* (2016); *Johnson et al.* (2016) and *Bruna et al.* (2016) were pioneers at exploiting perceptual losses to produce high-quality images for super-resolution and style transfer using convolutional networks. Such a loss extracts deep features from real and generated images by an external classification network, and minimizes their L1-distance to bring fake images closer to the real data. For semantic image synthesis, the VGG-based perceptual loss was first introduced by CRN (*Chen and Koltun*, 2017), and later adopted by Pix2pixHD (*Isola et al.*, 2017). Since then, it has become a default for training the generator (*Park et al.*, 2019b; *Liu et al.*, 2019; *Tan et al.*, 2020; *Tang et al.*, 2020a; *Richardson et al.*, 2021; *Wang et al.*, 2021b; *Li et al.*, 2021b; *Musat et al.*, 2022). As the perceptual loss is based on a VGG network pre-trained on ImageNet (*Deng et al.*, 2009), methods relying on it are constrained by the ImageNet domain and the representational power of VGG. With the recent progress on GAN training, e.g., by architecture designs and regularization techniques, the actual necessity of the perceptual loss requires a reassessment. In Chapter 4, we experimentally show that such loss imposes unnecessary constraints on the generator, significantly limiting the diversity among samples. We find that without the VGG loss it is possible to achieve higher diversity, at the same time not compromising the synthesis quality.

2.2.4 Semantic Image Synthesis not with GANs

In addition to GANs, the literature suggested alternative models for semantic image synthesis. Concurrently to Pix2Pix, several works trained generators without the adversarial components, using only

the VGG perceptual loss for supervision. For example, *Chen and Koltun* (2017) proposed to train a cascaded refinement network (CRN) with a combination of the VGG loss and a diversity regularization, outperforming Pix2Pix in diversity. SIMS (*Qi et al.*, 2018) extended CRN with a non-parametric component, serving as a memory bank of segments of real images to further improve the synthesis quality. However, all the subsequent non-GAN approaches were significantly underperforming in image quality compared to subsequent state-of-the-art GANs.

Diffusion models. More recently, diffusion models appeared as an alternative class of powerful image generation models (*Dhariwal and Nichol*, 2021; *Rombach et al.*, 2022). The first diffusion model for semantic image synthesis was SDM (*Wang et al.*, 2022b), which incorporates label maps using spatially-adaptive normalizations, similarly to SPADE (*Park et al.*, 2019b). Other models aimed to leverage pre-trained diffusion models, introducing extra label map encoders (PITI) (*Wang et al.*, 2022a) or spatial cross-attention layers (FreestyleNet) (*Xue et al.*, 2023). These diffusion models were published significantly later than OASIS and generally outperform our model in image quality and diversity. However, they still do not achieve a good alignment between images and label maps, achieving lower scores in the mIoU measure compared to OASIS.

2.3 Unconditional GANs in Low Data Regimes

To achieve impressive performance using the models reviewed in previous sections, it is typically necessary to provide a large dataset of training images, consisting of at least several thousand images. In contrast, the performance of GAN models severely degrades when the available training data is insufficient. For this reason, improving GANs in low data regimes has become its own research direction. In this section, we discuss existing approaches for training GANs in various data-limited regimes, which constitute related work for our models introduced in Chapters 5, 6, and 7. We categorize previous models based on the data regimes they were designed for (see Table 2.1), such as limited-data setups (100-5000 training images), few-shot learning (5-100 images), and one-shot training regime (learning from a single image). Additionally, we provide a separate overview of existing methods for the fine-tuning of pre-trained GAN models.

2.3.1 GANs Learning from Limited Data

Training Generative Adversarial Networks (GANs) can be highly challenging, especially when working with smaller datasets. The inherent imbalance between the generator and discriminator is amplified in such cases, as the discriminator quickly overfits to the limited training data, overtaking the generator. In fact, *Zhao et al.* (2020a) demonstrated that this issue is already prominent when training on a mere 10% subset of the CIFAR-10 dataset, which contains as much as 5000 images. Thus, many prior works attempted to address this problem and enable successful GAN training on smaller datasets. One line of works focused on leveraging data augmentations, a widely used technique to combat overfitting in discriminative deep learning (*Simard et al.*, 2003; *Wan et al.*, 2013). While augmenting real images is the obvious choice for GANs (*Zhang et al.*, 2020a), it has been shown that augmenting both real and fake images through differentiable image augmentations yields superior results (*Zhao et al.*, 2020c,a). *Karras et al.* (2020a) introduced an adaptive approach called Adaptive Differentiable Augmentation (ADA), which dynamically adjusts the strength of differentiable augmentations based on the current degree of overfitting in the discriminator. These data

Training regime	Dataset size	Chapters of this thesis
Training GANs from limited data	100 - 5000	-
Few-shot image synthesis	10-100	6, 7
One-shot image synthesis	1	5, 6

Table 2.1: Different low data regimes for training GANs.

augmentations have significantly reduced the data requirements of state-of-the-art GANs. For instance, StyleGANv2-ADA achieved successful training with as few as 1000 images from various high-resolution datasets, becoming the standard baseline for many subsequent studies. In line with these advancements, this thesis adopts differentiable augmentations during the training of our proposed models in Chapters 6-7.

Several other works have proposed alternative methods that complement differentiable augmentations. *Tseng et al.* (2021) introduced an anchors-based regularization term that encourages the discriminator to blend the predictions of real and generated images. *Yang et al.* (2021) made the discriminator's task more complex by requiring it to differentiate between individual images, thereby replacing the conventional real-fake binary classification. *Chen et al.* (2021a) proposed to train only a small subset of the networks' weights, resulting in reduced training time and improved synthesis quality. *Li et al.* (2022) identified the problem of latent space discontinuity in data-efficient GANs and solved it with a new contrastive loss. These works have significantly enhanced GAN synthesis when working with limited data. Still, their performance has primarily been demonstrated with datasets containing several hundreds of images, a scenario commonly referred to as "learning from limited data" in the GAN literature. In Chapters 6-7, we do not directly compare our proposed approaches to these models, as our methods operate in even lower data regimes, commonly denoted as "few-shot" or "one-shot" regimes.

2.3.2 Few-Shot GANs

Despite the considerable progress made in limited-data GANs thanks to data augmentation techniques and regularizations, these models still struggle with addressing the issue of overfitting in few-shot data regimes. This problem arises when the number of training images is severely limited, e.g., does not exceed a hundred. Stabilizing the training in this challenging regime required developing new architectures. For example, FastGAN (*Liu et al.*, 2021) proposed a new generator architecture using a channel-wise excitation module, and a self-supervised discriminator trained as a feature-encoder. This resulted in a lightweight architecture capable of generating high-quality and diverse images even when provided with as few as 100 training images. Building upon the advancements of FastGAN, FreGAN (*Wang et al.*, 2022c) introduced a series of techniques that enhance the spectral properties of the generated images. Additionally, GenCo (*Cui et al.*, 2022) introduced an approach involving multiple complementary discriminators that provide diverse supervision from multiple distinctive image views, further mitigating overfitting. These improvements contributed to further improving the quality and diversity of few-shot image generation.

In Chapter 5, we extend previous few-shot image generation benchmarks by introducing a new setup, in which the training datasets are composed of 60-100 frames taken from a short video clip. These datasets, characterized by a strong correlation among adjacent video frames, exhibit reduced

overall diversity, making them particularly challenging for existing few-shot GANs. As we will demonstrate in Chapter 5, our SIV-GAN model enables GAN training in this new challenging few-shot regime, while FastGAN, a strong few-shot GAN baseline, is not able to overcome memorization.

2.3.3 Single-Image GANs

Another line of works investigated generative models from a single image. Before the emergence of first single-image GANs, *Ulyanov et al.* (2018); *Shocher et al.* (2018) showed that when trained on a single image, a deep convolutional network can learn a useful representation that captures the internal statistics of that image. These learned representations can be employed to synthesize textures from a sample texture image (*Bergmann et al.*, 2017; *Zhou et al.*, 2018; *Ulyanov et al.*, 2017), to "blindly" super-resolve (*Shocher et al.*, 2018), or to inpaint the image (*Ulyanov et al.*, 2018).

Motivated by these advances, several single-image GAN models (*Shocher et al.*, 2019; *Shaham et al.*, 2019) have been proposed, revealing the power of image priors learned from a single natural image for synthesis and manipulation tasks. For example, InGAN (*Shocher et al.*, 2019) introduced a GAN model conditioned on a single natural image, which can remap the input to any size or shape while preserving its internal patch-based distribution. However, this model is limited to input images with highly repetitive image content (e.g., used for texture synthesis) and performs poorly with more natural images. SinGAN (*Shaham et al.*, 2019) proposed a new GAN architecture, in which a cascade of networks at different resolutions is trained in different stages. This allowed to learn multi-scale patch distribution of a given image and produce images of new size from noise. Later, ConSinGAN (*Hinz et al.*, 2021) improved SinGAN by rescaling the multi-stage training and training several stages concurrently, which enabled reducing the model size and made the training more efficient.

In Chapters 5 and 6, our SIV-GAN and OSMIS models demonstrate the capability to excel in the one-shot learning regime. In contrast to previous multi-stage single-image SinGAN and ConSinGAN, our models are not only designed to capture distributions of image patches, but also to learn more high-level image features like scene layouts and appearance of objects. Consequently, our models not only surpass prior methods in the one-shot regime but can also learn from multiple images, e.g., in few-shot setups. Furthermore, in addition to new versions of the provided single image, our model OSMIS from Chapter 6 not only provides new versions of the original single image, but also generates segmentation masks for synthesized images, which was not considered in prior works.

2.3.4 Fine-Tuning of Pre-Trained GANs

While the models discussed in Sections 2.3.1-2.3.3 are trained from scratch, there is a line of research aimed at avoiding overfitting to limited data through the use of transfer learning. These methods usually start from GANs pre-trained on large datasets, and adapt them on a few samples in the target domain by fine-tuning the generator and discriminator weights, e.g., as first proposed in TGAN (*Wang et al.*, 2018b). However, if the number of training images is not large enough (e.g. just several hundreds), naive fine-tuning still often suffers from overfitting and results in poor performance. To mitigate this issue, researchers proposed several techniques, such as mining suitable parts of the latent space before fine-tuning (*Wang et al.*, 2020) or restricting weight updates, for example, freezing the singular vectors of all layers (*Robb et al.*, 2021), updating only the BatchNorm parameters of the generator (*Noguchi and Harada*, 2019), penalizing drastic changes in important weights (*Li et al.*, 2020), or freezing the earliest layers of the discriminator, a technique referred to as FreezeD

(*Mo et al.*, 2020). While the above techniques help to mitigate overfitting during adaptation, their effectiveness is limited by the number of images required for successful training. They are unable to overcome the challenge of memorization when dealing with datasets consisting of fewer than 100 images.

In Chapter 7, we are interested to adapt pre-trained GANs on extremely small datasets, such as target domains containing only 10 images. This regime could not be successfully tackled with the above fine-tuning approaches, neither with prior few-shot or single-image GANs trained from scratch. Thus, recent works advocated the necessity of strong regularizations that preserve specific knowledge from the pre-trained model and prevent the degradation of diversity of the initial generators (*Zhao et al.*, 2022). For example, CDC (*Ojha et al.*, 2021) proposed to preserve the pair-wise perceptual similarity between samples from the source domain and transfer it to the target domain, while RSSA (*Xiao et al.*, 2022) designed a novel consistency term to align the structural information between source and target domains. These two methods achieved impressive performance on the task of 10-shot GAN adaptation, however, their assumptions impose strong constraints on the structure of the few-shot target domain. In particular, they fail in the more challenging regime when the source and target domains are not restrictively similar, as will be shown in Chapter 7. This issue was addressed by AdAM (*Yunqing et al.*, 2022), which replaced knowledge preservation criterias with adaptation-aware kernel modulation (AdAM), relaxing the source-target proximity requirement to some extent. Nonetheless, the issue of source-target proximity is not solved until today, as the performance of methods still strongly depends on the semantic consistency between the target domain and the pre-trained model, and their incompatibility can make the use of pre-training meaningless (*Zhao et al.*, 2020a). In Chapter 7, we introduce a new regularization term to preserve the generator's smoothness properties that are not limited to a specific domain, enabling successful adaptation between image domains of unprecedented structural dissimilarity.

2.4 GANs for Joint Synthesis of Images and Segmentation Masks

So far, we have discussed only single-modality GANs, that only generate new images from provided image datasets. Naturally, many computer vision applications require paired data, such as segmentation applications that require images and their pixel-level label maps. Therefore, several works concentrated on GANs to generate segmentation masks along with images.

Most works on image-mask GANs are based on the observation that a GAN generator, trained on a large dataset, implicitly learns discriminative pixel-wise features of the generated scene objects (*Tritrong et al.*, 2021). Thus, it is possible to extract these features from different generator layers and transform them into a segmentation mask of objects using a small decoder. As the ground truth segmentation masks for generated images are typically not available, prior works proposed several methods to train such a decoder. For example, RepurposeGAN (*Tritrong et al.*, 2021) and DatasetGAN (*Zhang et al.*, 2021b) proposed to train the decoder using a handful of manually annotated generated images. LinearGAN (*Xu and Zheng*, 2021) replaced manual annotations by the predictions of an external segmentation network. Alternatively, SemanticGAN (*Li et al.*, 2021a) and EditGAN (*Ling et al.*, 2021) enforced the alignment between generated images and masks with the loss from an additional discriminator, which takes both images and masks as inputs.

Although the above models require only a few masks to achieve high-quality image-mask synthesis, they are not successful when the number of training images is not sufficient. This significantly

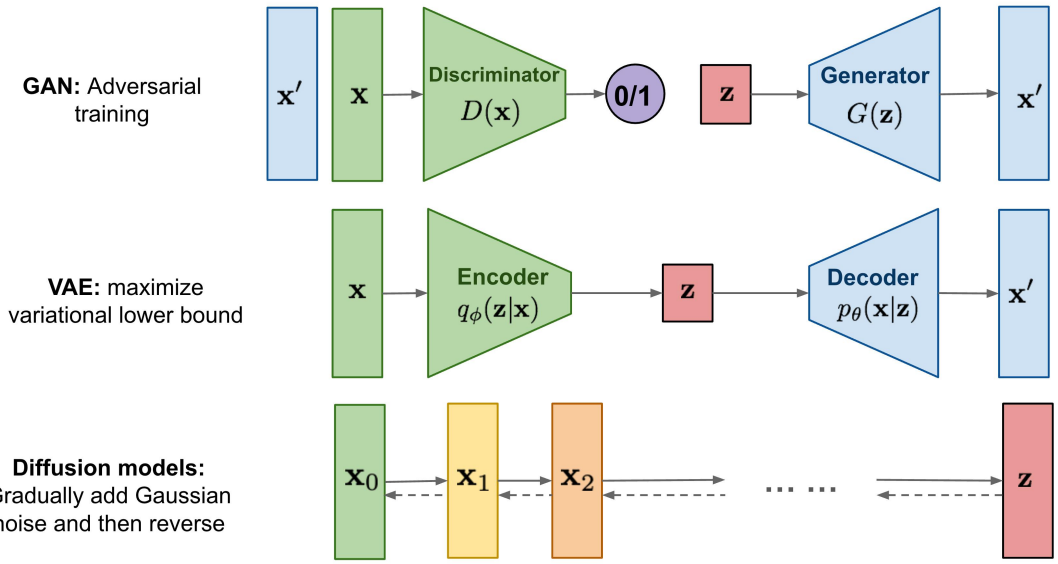

Figure 2.4: Comparison of GANs and two alternative image synthesis paradigms: variational autoenvoders (VAE) and diffusion models. The image source is (*Weng*, 2021).

limits their application in restricted image domains. For example, these models do not allow to generate diverse data augmentation that can be used in data-limited setups, for example, in one-shot segmentation applications. In Chapter 6, we introduce our OSMIS model, that, in contrast to prior image-mask GANs, learns high-quality image-mask synthesis just from a single image-mask pair, without requiring pre-training. A specific advantage of our model is that it is trained in a purely adversarial fashion without any additional overhead, e.g., not requiring manual annotations of generated images, external segmentation networks, or additional discriminators.

2.5 Alternatives to GANs

Lastly, we discuss alternative image generation models that are based on likelihood maximization rather than on adversarial training. The two most popular alternative paradigms are variational autoencoders (VAE) and diffusion models (their schemes are shown in Fig. 2.4). In the following, we briefly explain their working principles and their advantages and disadvantages compared to GANs.

Variational autoencoders. VAEs are constructed as encoder-decoder structures, with an encoder mapping an input image to a low-dimensional latent space, and a decoder reconstructing the image based on its latent input code. The loss function used in VAEs consists of two components: a reconstruction loss and a regularization term. The reconstruction loss measures the dissimilarity between the original input and the reconstructed output. The regularization term, based on the Kullback-Leibler (KL) divergence, encourages the distribution in the latent space to match a predefined prior distribution, typically a standard Gaussian. Once a VAE is trained, the decoder network can be used to generate new data samples by sampling from the latent space.

The first VAE was proposed in (*Kingma and Welling*, 2014). The most recent VAE-based image synthesis models (*Child*, 2022; *Vahdat and Kautz*, 2020; *Hazami et al.*, 2022) excel at generating

images from datasets of simple structure, such as human faces. However, modeling more complex datasets, such as ImageNet, is still challenging with VAEs and so far has led to unsatisfying image quality (*Child*, 2022). Overall, compared to GANs, VAEs have an advantage of having more stable training process, as well as a lack of mode collapse. However, they allow to achieve only reduced image quality and cover less complex datasets. Another disadvantage of VAEs is the tendency to produce blurry outputs, since using the reconstruction loss incentivizes the decoder to tend towards the mean image across all plausible outputs from a given latent vector. Among the tasks studied in Chapters 4-7, VAEs do not achieve comparable performance to GANs. We therefore do not compare our proposed models to VAEs in this thesis.

Diffusion models. Diffusion models (*Ho et al.*, 2020) build on the concept of the diffusion process, which is a stochastic process that describes the spread or diffusion of particles over time. In the context of image generation, the core idea behind diffusion models is to transform a simple Gaussian distribution into distribution of realistic images through a sequence of intermediate steps. In the forward diffusion step, the image is updated by adding noise and applying a learnable transformation. In the backward steps, the input Gassian noise is gradually reduced in multiple steps, in a way that the the final denoised image looks realistic. During training, the model learns the parameters of the diffusion steps, such as the noise level and the transformation functions, which allows it to generate high-quality samples during backward process.

Unlike GANs, diffusion models do not suffer from training instabilities and frequently achieve higher synthesis diversity, avoiding the issues related to mode collapse. Diffusion models are evolving rapidly, and have already outperformed GANs on several tasks and datasets (*Nichol and Dhariwal*, 2021; *Dhariwal and Nichol*, 2021; *Rombach et al.*, 2022). The main disadvantage of diffusion models is their slow sampling time due to the iterative generation procedure, which may require thousands of network evaluations. For example, it can take up to several days to generate 50000 images with a diffusion model, which is commonly required to compare to other generative models from the literature. One way to make diffusion models more efficient is to apply them only in an image generator's latent space, referred to as a latent diffusion model (*Rombach et al.*, 2022).

Diffusion models, as a relatively recent type of models, have mostly emerged after the development of our models presented in Chapters 4-7. They have been successfully applied in semantic image synthesis, but so far have not been widely adopted in image generation tasks related to low data regimes. Therefore, in this thesis, we will compare our semantic image synthesis model OASIS to diffusion models in Chapter 4. In contrast, models from Chapters 5-7, designed for low data regimes, will not be compared to diffusion models.

CHAPTER 3

Preliminaries

In this chapter, we present the formalism and notations related to the key concepts addressed in this thesis, including the studied tasks, algorithms, and applications. As the main focus of this thesis revolves around the generation of realistic images, we start by defining the tasks of generative modeling and image synthesis and highlight their connection to other machine learning tasks. We also discuss different methods for their evaluation. The subsequent sections focus on the central tool studied in this thesis: generative adversarial networks (GANs). Here, we discuss the operational principles of GANs and explain the optimization procedures used for their training. As highlighted in the preceding chapter, the effectiveness of GANs heavily relies on the specific architectures of the generator and discriminator networks, as well as on the applied regularizations and other training techniques. Hence, we provide an overview of the architectures, regularizations, and training strategies upon which our models in Chapters 4-7 are constructed. Finally, apart from synthesizing realistic images across various scenarios, our thesis also explores the usage of image synthesis in downstream applications. Therefore, we conclude this chapter by providing a description of these downstream tasks and their relationship to our proposed models.

Contents

3.1 The Task of Image Generation

3.1.1 Unsupervised Learning and Generative Modeling

From the perspective of machine learning, generative modeling can be considered as a form of unsupervised learning. The general goal of unsupervised learning is to learn the distribution of provided

data and extract meaningful patterns, structures, and relationships inside a given dataset, without any explicit guidance or labeled examples. Mathematically, we consider an input space $\mathcal{X} \subseteq \mathbb{R}^d$, a fixed unknown probability distribution $\mathcal{D}_{data}$, and a dataset $S = (\mathbf{x}_i)_{i=1}^m$ consisting of m examples that are i.i.d. sampled from $\mathcal{D}_{data}$. There are several tasks that fall into the category of unsupervised learning, for instance, the following examples:

- **Clustering:** given a similarity measure $d : \mathcal{X}^2 \to \mathbb{R}^+$ between data points, find a function $F(\mathbf{x}) : \mathcal{X} \to \mathcal{Y}$, which assigns a given data point $\mathbf{x}$ to one of n clusters $\{1, ..., n\} \in \mathcal{Y}$, in a way that similar data points fall into the same clusters, while dissimilar points are separated.
- **Dimensionality reduction:** learn a mapping from input space to a much lower-dimensional space $F(\mathbf{x}) : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}^l$, where $l << d$, in a way that the transformed distribution $F(\mathcal{D}_{data})$ preserves most of the information about original data distribution $\mathcal{D}_{data}$.
- **Density estimation:** learn the underlying probability density function $p(\mathbf{x})$ (PDF) of the data distribution $\mathcal{D}_{data}$.

Our task of generative modeling is similar to the above examples, as it also aims to learn the characteristics of data distribution $\mathcal{D}_{data}$ by using unstructured collections of provided data points $S = (\mathbf{x}_i)_{i=1}^m$, while also assuming no access to labels and any other external guidance. In contrast to the above examples, our task operates directly in the input data space $\mathcal{X}$:

- **Generative modeling:** train a generator function $G : \mathcal{Z} \to \mathcal{X}$, in a way that it transforms a given prior probability distribution $\mathcal{Z}$ into a generated distribution $\mathcal{D}_{gen} = G(\mathcal{Z})$ that is close to the real data distribution $\mathcal{D}_{data}$.

This way, the trained generator G should allow to generate individual synthetic samples $x_{gen} = G(z), z \sim \mathcal{Z}$ that follow the distribution of the dataset S consisting of real data samples.

The concept of data generation is closely associated with other unsupervised learning tasks. On one hand, the prior distribution $\mathcal{Z}$ typically encompasses low-dimensional choices, e.g., $\mathcal{Z} : \mathbb{R}^k \to \mathbb{R}$, where k is significantly smaller than the input data dimensionality d. Consequently, when training a generative model, the real data distribution can be approximated by a lower-dimensional manifold $G(\mathcal{Z})$, which remains within k dimensions. In this manner, generative modeling provides a form of dimensionality reduction for representing real data.

On the other hand, in the context of density estimation, many existing approaches not only learn the density function $p(x)$, but also offer effective algorithms for sampling from this distribution. For a well-estimated $p(x)$, this sampling provides samples that follow $\mathcal{D}_{data}$, which can be regarded as a solution for the data generation task. However, it is important to note that these methods may not be suitable for sampling all types of data $\mathcal{X}$, as will be discussed further.

3.1.2 Image Synthesis

Image synthesis is a branch of generative modeling that focuses specifically on generating new images. Mathematically, an image can be represented as a set of C×H×W numbers, where H and W are the height and width of images in pixels, and the third dimension C is usually set to 3, representing the intensities of each pixel in the RGB color space. Thus, the task of image synthesis can be formulated as:

- **Unconditional image synthesis:** given a prior probability distribution (latent space) $\mathcal{Z} : \mathbb{R}^k \rightarrow \mathbb{R}$, a dataset of images $S = (\mathbf{x}_i)_{i=1}^m, \subseteq \mathcal{X} = \mathbb{R}^{C \times H \times W}$ sampled i.i.d. from $\mathcal{D}_{data}$, and a pre-defined similarity measure between data distributions $d(.,.)$, train a generator function $G : \mathcal{Z} \rightarrow \mathcal{X}$ that minimizes $d(G(\mathcal{Z}), \mathcal{D}_{data})$.

Optionally, the dataset S can contain conditioning information $y \in \mathcal{Y}$ about each image, such as a class label or semantic label map. In this case, the task of image synthesis becomes conditional, and the goal is to model a joint distribution of images and labels:

- **Conditional image synthesis:** given a prior probability distribution (latent space) $\mathcal{Z} : \mathbb{R}^k \rightarrow \mathbb{R}$, a dataset of images with labels $S = (\mathbf{x}_i, y_i)_{i=1}^m, \subseteq \mathcal{X} \times \mathcal{Y} = \mathbb{R}^{C \times H \times W} \times \mathcal{Y}$ sampled i.i.d. from $\mathcal{D}_{data}$, and a pre-defined similarity measure between data distributions $d(.,.)$, train a generator function $G : \mathcal{Z} \times \mathcal{Y} \rightarrow \mathcal{X}$ that minimizes $d(G(\mathcal{Z}, \mathcal{Y}), \mathcal{D}_{data})$.

While the definitions of unconditional and conditional image synthesis are similar to those of other generative modeling tasks, working with images as the modeled space $\mathcal{X}$ presents distinct challenges compared to other data types. These challenges arise primarily due to the intricate structure of natural images. The summary of their specific structure and their challenges is provided below:

High dimensionality. Images are typically represented as very high-dimensional data. For example, most images considered in Chapters 4-7 have resolution 256×256, so they consist of 3×256×256=196608 dimensions. Modeling complex dependencies in $\mathcal{X} \subseteq \mathbb{R}^{196608}$ requires very powerful models catching non-linear dependencies between millions of pairs of dimensions. The high dimensionality of the problem also makes the problem very difficult computationally.

Scarcity of the target manifold. The manifold of realistic images occupies an extremely sparse region within the vast space of all possible images $\mathcal{X}$ (*Arjovsky et al.*, 2017). For this reason, during the initial stages of training a generator, the distributions $\mathcal{D}_{data}$ and $G(\mathcal{Z})$, representing real and generated images, often lack any significant overlap. This lack of overlap poses challenges for training the generator, as numerous statistical similarity measures rely on non-zero support overlap between the compared distributions in order to yield useful training signals.

Spatial structure. Images have inherent dependencies along spatial dimensions W and H, meaning that the arrangement of pixels and their relationships within a local neighborhood or global context is crucial for generating coherent images. Capturing spatial structure requires models that can effectively learn both short-range and long-range dependencies and preserve spatial consistency.

Diversity. At a high level, images can exhibit a vast range of variations, including different objects, backgrounds, lighting conditions, orientations, and scales. Additionally, while certain datasets encompass a wide range of object categories and scene types, there are also datasets specifically curated for industrial applications, featuring grayscale images showcasing only one type of manufacturing detail. Such diversity in images, as well as diversity in datasets themselves, requires careful generator designs that have to be different for each specific task at hand.

Interpretability and ambiguity. Images can be subject to subjective interpretation by people, and their meaning may vary across different observers. For example, in class-conditional datasets, it is a known issue that the annotation of classes can vary between individuals (e.g., what one person perceives as a "dog" may be labelled as a "wolf" or a "fox" by other annotators). In such cases, a

successful generator needs to capture the inherent ambiguity and subjectivity in image data, which is hard to formalize.

3.1.3 Evaluation of Image Synthesis

Similar to other machine learning tasks, the development of image synthesis models requires quantitative measures which allow objective comparisons across different models. Designing such measures is, however, not straightforward. At first glance, one might consider deriving a measure of a model's quality from the task's definition itself: a similarity measure $d(,)$ already provides an estimation of how close the distributions of generated and real images are. Nonetheless, the problem with such a metric is that it would not be universal, as different models employ drastically different approaches for modeling $d(,)$. For example, while some approaches like autoregressive models (*Van Den Oord et al.*, 2016) allow density estimation and computing likelihood of images in the validation set, other models (e.g., GANs) utilize learnable discriminator critiques that are unique to each trained model. Thus, these measures are not directly comparable to each other, which makes objective comparisons between models difficult.

Another difficulty in the evaluation of image synthesis is subjectivity. Unlike other types of data, generated images in many applications are primarily intended for visual perception by humans. Therefore, any attempt to quantitatively assess image synthesis models must consider the correlation to human preferences and expectations about how generated images should look like (*Heusel et al.*, 2017). This is difficult due to the highly non-linear and subjective nature of human perception of image quality, which also varies from person to person (*Goetschalckx et al.*, 2019). In addition, human perception is highly sensitive to minor visual cues, and even slightest deviations from realism can lead to images that appear unnatural or unconvincing to a human eye.

One more challenge of evaluating image synthesis is to detect overfitting and mode collapse. In principle, an image generator that learned to replicate all training samples $(\mathbf{x}_i, y_i)_{i=1}^m$ naturally exhibits low $d(G(\mathcal{Z}, \mathcal{Y}), \mathcal{D}_{data})$, and such "generated" images would also look highly realistic from the perspective of human perception. Therefore, a good evaluation protocol for image synthesis should encompass both the quality and diversity aspects.

Although the search for reliable evaluation metrics for image synthesis remains an active area of research (*Borji*, 2022; *Alaa et al.*, 2022), several evaluation approaches have gained widespread usage. In this thesis, we mostly use two metrics: FID (*Heusel et al.*, 2017) and LPIPS (*Zhang et al.*, 2018b), separately measuring the quality and diversity of generated images. For completeness, below we present an overview of these metrics along with other commonly used evaluation methods.

- **Human studies.** Evaluating the realism of generated images often begins with the simplest approach of seeking human assessment. Numerous studies employ large-scale evaluations by asking users to provide feedback on the generated images through various types of questions. For instance, one common test involves presenting a mixture of real and fake images and asking users to identify the fake ones. A higher confusion rate among users in this test would suggest a higher level of realism in the generated images. Alternatively, researchers may present batches of images generated by different models and request users to rank them. Overall, human studies allow the researchers to capture subjective perceptions and preferences of human participants. However, conducting extensive user studies with a significant number of participants can be both time-consuming and expensive, often requiring external platforms

such as Amazon Mechanical Turk. These logistical constraints, coupled with the potential for biased preparation of samples and the subjective nature of user responses, make human studies challenging to reproduce and susceptible to misinterpretation.

- **Negative log-likelihood (NLL).** To assess the generalization ability of generative models that allow the computation of the image likelihood, a common approach is to calculate the negative log-likelihood (NLL) of real images. This typically requires a held-out validation set of images that were not used during training the model. For evaluation, the NLL is then averaged over all images in the validation set, and lower NLL values indicate better performance. It is worth noting that NLL does not always align with the perceptual quality of generated images, and it is not applicable in case of GANs, which are the main focus of this thesis.

- **Inception score (IS).** Other proposed approaches to evaluate image synthesis involve external pre-trained neural networks. One of them is Inception Score (IS) (*Salimans et al.*, 2016), which is based on the Inceptionv3 image classification network (*Szegedy et al.*, 2016), pre-trained on ImageNet (*Deng et al.*, 2009). Computing IS requires passing a set of generated images $(\mathbf{x}_{gen,i})_{i=1}^{m}$ (typically, with $m = 50$k samples) through the Inceptionv3 network to estimate conditional label distributions $p(y|\mathbf{x}_{gen,i})$, where y denotes classes from ImageNet. The two assumptions of IS are that a well trained generator should have images with $p(y|\mathbf{x}_{gen,i})$ of low entropy, while the entropy of $\int p(y|\mathbf{x} = G(z))dz$ should be high. The first assumption corresponds to generating "sharp" (realistic) images that can be easily categorized by the InceptionV3 network, while the second assumption forces the images to be categorized as different classes, corresponding to high diversity. Mathematically, IS is formulated as:

$$IS = \exp\left(\mathbb{E}_{x\sim\mathcal{D}_{gen}}\left[D_{KL}\left(p(y|\mathbf{x})||\int p(y|\mathbf{x} = G(z))dz\right)\right]\right). \tag{3.1}$$

A high Inception Score indicates that the generated images are both of high quality and exhibit diversity. However, it is worth noting that the Inception Score has some limitations. The major drawback of IS is that it is computed irrespective of $\mathcal{D}_{real}$ and does not consider the similarity of the generated distribution to the target domain of real images. This way, IS cannot provide a reliable metric for the image domains that are significantly different from ImageNet.

- **Frechet Inception distance (FID).** Unlike IS, which evaluates only the distribution of generated images, the Frechet Inception distance (FID) (*Heusel et al.*, 2017) compares the distribution of generated images with the distribution of a set of real images. Computing FID involves calculating the feature representations of both the real images and the synthetic images using a pre-trained InceptionV3 network (excluding the last classification layer). This is followed by the computation of the mean and covariance matrices of the obtained representations. The final step is to fit multivariate Gaussian distributions for the real and fake mean and covariance matrices and to compute the Frechet distance between them. Mathematically, for the two estimated Gaussian distributions $\mathcal{N}(\mu_{gen}, \Sigma_{gen})$ and $\mathcal{N}(\mu_{real}, \Sigma_{real})$, the FID equals:

$$FID = ||\mu_{gen} - \mu_{real}||_2^2 + \text{Tr}\left(\Sigma_{gen} + \Sigma_{real} - 2\left(\Sigma_{gen}\Sigma_{real}\right)^{1/2}\right). \tag{3.2}$$

The FID metric considers both the quality and diversity of generated images. A lower FID score indicates a closer similarity between the real and synthetic image distributions, implying

better image quality and more similar diversity to the real images. Similarly to IS, the FID metric has its own set of advantages and limitations. In general, it addresses some of the limitations of the Inception Score by directly comparing feature representations, rather than relying solely on predicted class probabilities. More importantly, it was shown to correlate very well with the human perceptual judgement. On the other hand, FID is still computed based on a network pre-trained on ImageNet, which introduces biases. In addition, computing FID requires multiple forward passes through the InceptionV3 network and computing inverse matrices, which can be computationally expensive, especially when dealing with large-scale datasets. It was also shown to be statistically biased, in the sense that the expected value of FID over finite data is not the same as when a number of samples is infinite (*Chong and Forsyth*, 2020). Despite these considerations, FID remains the standard metric for evaluating image synthesis models, and is used in Chapters 4-7 of this thesis.

- **Single-Image Frechet Inception distance (SIFID).** In the special task of generating new images from a single training image, computing FID becomes impossible due to the absence of a defined covariance matrix Σ_{real} (this matrix cannot be derived from a single image alone). To address this limitation, *Shaham et al.* (2019) introduced a new evaluation approach, which involves computing statistics from feature representations at much earlier layers of the InceptionV3 network, rather than before the final classification layer. As the receptive field of these layers is relatively small, this way of computation allows to evaluate the statistics of different small image patches (e.g., $\frac{1}{16}$ of the size of the original image), and compute the covariance matrices using them. After that, when $\mathcal{N}(\mu_{gen}, \Sigma_{gen})$ and $\mathcal{N}(\mu_{real}, \Sigma_{real})$ are calculated based on different patches of generated and fake images, the SIFID is computed according to Eq. 3.2. In effect, SIFID allows to measure the quality of images generated by models trained on a single image. In this thesis, SIFID at various InceptionV3 layers is used in Chapters 5-6.

- **Learned Perceptual Image Patch Similarity (LPIPS).** While FID has been widely adopted in the community for evaluating the quality of generative models, it has faced criticism due to its limitations in detecting significant failure cases. One fundamental drawback is its one-dimensional nature, which prevents it from distinguishing between models with high precision but poor recall and those with poor precision but high recall (*Sajjadi et al.*, 2018). Additionally, FID fails to identify overfitting, as a model that simply replicates the training set can achieve an FID score close to zero. Consequently, a comprehensive evaluation of image synthesis should include another measure that isolates image diversity from fidelity.

  An example of a purely diversity-based metric is LPIPS. It is based on the idea that human perception of image diversity is not purely based on pixel-wise differences but also on higher-level features and structures within the images. To capture these perceptual differences between images, LPIPS uses image features computed by deep layers in a pre-trained classification network. For two given images x_1 and x_2 and a pre-trained network F, LPIPS is computed as:

$$LPIPS = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} ||F^l_{hw}(x_1) - F^l_{hw}(x_2)||_2, \tag{3.3}$$

  where H_l, W_l correspond to the dimensions of features at layer l, while F^l_{hw} denotes the features computed at layer l at spatial location (h, w).

By comparing different pairs of generated images with LPIPS, one can estimate the degree of diversity present in the generated images. This measure, computed independently from FID, can provide a score that helps to detect training instabilities, mode collapse, and simply lack of diversity in generated images. Another method to use LPIPS is to compute it between generated and training images. This way, the average LPIPS from generated images to their "nearest" training samples would clearly indicate the degree of memorization in a generator. In this thesis, LPIPS is used to evaluate the diversity of our models in each of the Chapters 4-7.

3.2 Generative Adversarial Networks

In this thesis, we consider a special type of image synthesis models – generative adversarial networks (GANs). In the next sections, we explain their working principles and introduce the GAN architectures, losses, and training techniques that will be used in the next chapters.

3.2.1 GAN Working Principles

Like other generative models, GANs aim to train a generator G that transforms a prior distribution $\mathcal{Z}$ into the distribution of generated images $G(\mathcal{Z})$ which is close to $\mathcal{D}_{real}$. As also commonly used in other generative models, for the latent space $\mathcal{Z} : \mathbb{R}^k \rightarrow \mathbb{R}$, GANs select a multi-variate Gaussian distribution $z \sim \mathcal{N}(\mathbf{0}, I)$, where k is usually chosen as a power of 2 within the range of 64 to 512. In contrast to other approaches, instead of directly estimating the probability density $p(x)$ for generated images, GANs utilize an additional classifier network $D(x) \in [0, 1]$, estimating the probability that an input image x is real. The D network thus serves as a proxy for the measure $d(,)$, which evaluates the similarity between generated and real images.

The overview of the general GAN paradigm is presented in Fig. 3.1. A GAN consists of two networks: a generator G and a discriminator D. G takes a random noise vector z and generates an image $x_{gen} = G(z)$. D takes two inputs: the generated image x_{gen} and an image x_{real} from a provided dataset of real images; the output of D is a score $D(x)$ giving a probability of an image being real. The generator and discriminator are trained in alternate fashion, performing updates one after another. The goal of the generator is to produce images that the discriminator judges as real. In turn, the discriminator has to categorize all input images correctly. Therefore, GAN training is a game of two players with the following binary cross-entropy objective:

$$\min_G \max_D \mathbb{E}_{z\sim p_{\mathcal{Z}}}\big[\log(1 - D(G(z)))\big] + \mathbb{E}_{x\sim \mathcal{D}_{real}}\big[\log(D(x))\big]. \tag{3.4}$$

The solution to (3.4) is a saddle point, from which neither G nor D can improve. To train the networks, the ADAM (*Kingma and Ba*, 2015) optimizer is used with learning rates typically set in the range from 0.0001 to 0.001 for both G and D. It is worth noting that the gradients resulting from $log(1 - D(G(z)))$ can vanish if the discriminator produces confident predictions. Therefore, it is common to change the objective to a non-saturating variant of (3.4), referred to **NS-GAN**:

$$\begin{aligned} \mathcal{L}_G &= -\mathbb{E}_{z\sim p_{\mathcal{Z}}}\big[\log(D(G(z)))\big], \\ \mathcal{L}_D &= -\mathbb{E}_{z\sim p_{\mathcal{Z}}}\big[\log(1 - D(G(z)))\big] - \mathbb{E}_{x\sim \mathcal{D}_{real}}\big[\log(D(x))\big]. \end{aligned} \tag{3.5}$$

The described GAN training pipeline is summarized in Algorithm 1.

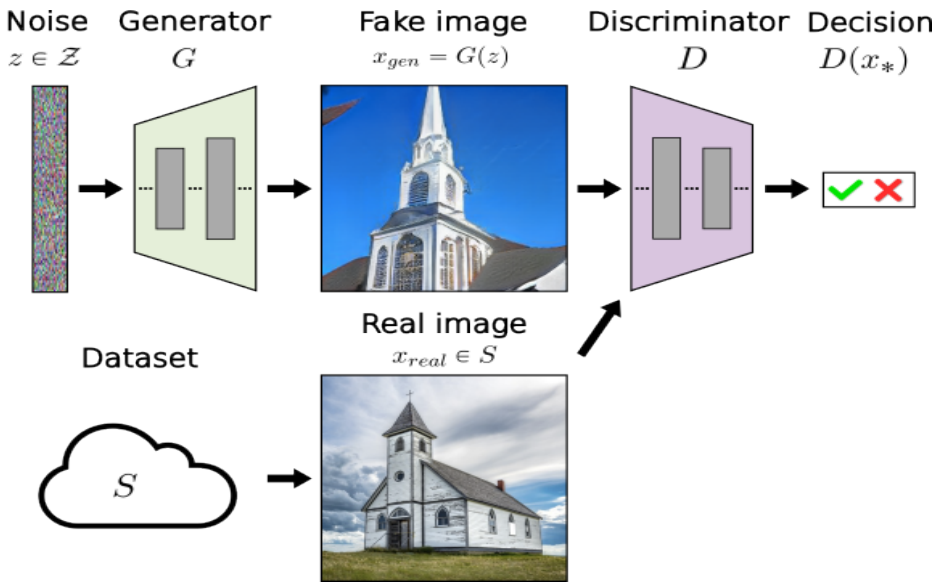

Figure 3.1: Overview of the GAN paradigm. A generator G produces an image x_{gen} from the input noise $z \in \mathcal{Z}$. A discriminator D tries to distinguish between generated images x_{gen} and real images from a provided dataset $x_{real} \in S$.

Algorithm 1 GAN training.

Input: G: Generator network, D: Discriminator network, S: Dataset, N: Batch size, $Opt.$: Parameters of the optimization algorithm

for *number of training iterations* **do**

· Sample a batch of noise vectors $\{z_1, ..., z_N\} \subset \mathcal{Z}$

· Sample a batch of images $\{x_1, ..., x_N\} \subset S$ (possibly, with data augmentation)

· Update D using $Opt.$ minimizing loss

$$-\frac{1}{N}\sum_{n=1}^{N}\Big[\log D(x_n) + \log(1 - D(G(z_n)))\Big]$$

· Update G using $Opt.$ minimizing loss

$$-\frac{1}{N}\sum_{n=1}^{N}\Big[\log D(G(z_n))\Big]$$

end for

3.2.2 Alternative Adversarial GAN Losses

The form of the GAN adversarial objective is mainly defined by the choice of the classification loss for the discriminator. While the NS-GAN loss, which is based on binary cross entropy, is used in the majority of GAN models, there exist several alternative formulations:

Least-square GAN (LSGAN) (*Mao et al.*, 2017)**:**

$$\begin{aligned}\mathcal{L}_G &= \frac{1}{2}\mathbb{E}_{z\sim p_{\mathcal{Z}}}\big[(D(G(z))-1)^2\big],\\ \mathcal{L}_D &= \frac{1}{2}\mathbb{E}_{z\sim p_{\mathcal{Z}}}\big[D(G(z))^2\big]+\frac{1}{2}\mathbb{E}_{x\sim \mathcal{D}_{real}}\big[(D(x)-1)^2\big].\end{aligned} \tag{3.6}$$

Hinge loss (*Lim and Ye*, 2017)**:**

$$\begin{aligned}\mathcal{L}_G &= -\mathbb{E}_{z\sim p_{\mathcal{Z}}}\big[D_{logits}(G(z))\big],\\ \mathcal{L}_D &= -\mathbb{E}_{z\sim p_{\mathcal{Z}}}\big[\min(0,-1+(D_{logits}(G(z)))\big]-\mathbb{E}_{x\sim \mathcal{D}_{real}}\big[\min(0,-1-(D_{logits}(x))\big].\end{aligned} \tag{3.7}$$

Note: the formulation of the Hinge loss uses the discriminator logits $D_{logits}(x) \in (-\inf, \inf)$ instead of the probabilities $D(x) \in [0,1]$ that are usually obtained via a non-linear activation layer.

Wasserstein GAN (WGAN) (*Arjovsky et al.*, 2017)**:**

$$\begin{aligned}\mathcal{L}_G &= -\mathbb{E}_{z\sim p_{\mathcal{Z}}}\big[D(G(z))\big],\\ \mathcal{L}_D &= +\mathbb{E}_{z\sim p_{\mathcal{Z}}}\big[D(G(z)))\big]-\mathbb{E}_{x\sim \mathcal{D}_{real}}\big[D(x)\big].\end{aligned} \tag{3.8}$$

In contrast to minimizing various statistical divergence measures (e.g., the JS-Divergence for NS-GAN and Pearson χ^2-Divergence for LSGAN), the WGAN loss uses the earth-mover Wasserstein distance between the real and generated data distributions. In Chapters 4-7, all our developed models use the NS-GAN objective, while some of the important comparison baselines employ the Hinge loss (such as SPADE, FastGAN, and BigGAN) or WGAN (used in SinGAN and ConSinGAN).

3.2.3 GAN Regularizations

In its basic form, the optimization of the adversarial losses from (3.5) can lead to training instabilities or mode collapse due to the non-convex minimax nature of the objective. For this reason, GANs often have additional regularization terms that are added either to $\mathcal{L}_G$ or $\mathcal{L}_D$. While the amount of GAN regularizations available in the literature is tremendous, below we present the four regularizations that are most relevant to our thesis:

Gradient penalty. The Gradient penalty (*Gulrajani et al.*, 2017) was introduced to stabilize GAN training by constraining the speed of the discriminator's gradient updates. This regularization penalizes the squared difference of the D's gradient for deviating from the unit sphere:

$$\mathcal{L}_{GP} = \lambda_{GP}\cdot\mathbb{E}_{\hat{x}\sim\mathcal{P}_{\hat{x}}}\big[(||\nabla D(\hat{x})||_2-1)^2\big]. \tag{3.9}$$

In Eq. (3.9), $P_{\hat{x}}$ is the distribution obtained by uniformly sampling along a straight line between the real and generated distributions. This way, $\hat{x}$ are obtained as linear interpolations between generated and real images. The Gradient penalty was proposed as a replacement for the earlier gradient clipping approach (*Arjovsky et al.*, 2017). This regularization significantly improved the performance and stability of previous GAN models, allowing them to model much more complex data distributions.

Spectral normalization. Another method to enforce the Lipschitz property on trained networks is to use the Spectral normalization (*Miyato et al.*, 2018). Spectral normalization tackles the issue of

drastic output changes by normalizing the spectral norm of the weight matrices. The normalization is typically applied to all learnable layers of both G and D, such as convolutions or fully-connected layers. For a given layer g, parameterized with a weight matrix A, the first step of spectral normalization is to compute the spectral norm of A:

$$\sigma(A) = \max_{\mathbf{h}\neq 0} \frac{||A\mathbf{h}||_2}{||\mathbf{h}||_2} = \max_{||\mathbf{h}||_2 \leqslant 1} ||A\mathbf{h}||_2, \tag{3.10}$$

which is equivalent to the largest singular value of A. The spectral normalization then normalizes the spectral norm of the weight matrix A so that it satisfies the Lipschitz constraint $\sigma(A) = 1$.

By using spectral normalization, GANs can achieve improved stability, better convergence properties, and higher-quality generated samples. It has been also shown to be effective in mitigating mode collapse, where the generator fails to capture the full diversity of the target distribution.

Diversity-sensitive loss. While Lipschitz regularizations like spectral normalization stabilize the training and mitigate mode collapse, they do not directly address the problem of GANs generating images with reduced diversity compared to real data. For this reason, *Yang et al.* (2019) proposed a diversity-sensitive regularization that explicitly encourages the generator to produce diverse outputs for different input noise vectors. This loss term is expressed as:

$$\mathcal{L}_{DS} = -\lambda_{DS} \cdot \mathbb{E}_{z_1, z_2 \sim p_{\mathcal{Z}}} \left[\min \left(\frac{||G(z_1) - G(z_2)||}{||z_1 - z_2||}, \varepsilon \right) \right], \tag{3.11}$$

where ε is an upper bound that assists numerical stability. $\mathcal{L}_{DS}$ helps to structure the generator in a way that the diversity of generated images is maximized without compromising image quality.

Path length regularization. *Odena et al.* (2017) and *Karras et al.* (2020b) observed a positive correlation between the quality of GAN-based generated images and the smoothness of the generator's mapping that was used to produce them. Mathematically, the generator's smoothness is expressed via a Jacobian matrix $J_G(z) = \partial G(z)/\partial z$, quantifying how G's output changes under noise shifts. In a smooth generator, random small steps in the latent space lead to minor but non-zero changes in output images. This property is usually visualized via latent space interpolations, which demonstrate that intermediate images $G(\alpha z_1 + (1-\alpha)z_2)$ are realistic for different pairs of z_1 and z_2. In prior work, it was therefore proposed to enforce smooth latent space interpolations explicitly using the generator's Jacobian matrix. This regularization is known as the path length regularization (PPL):

$$\mathcal{L}_{PPL} = \lambda_{PPL} \cdot \mathbb{E}_{z \sim p_{\mathcal{Z}}} (||J_G^T(z) \cdot y||_2 - a)^2. \tag{3.12}$$

Karras et al. (2020b) showed that minimizing Eq. (3.12) leads to a generator with an orthogonal $J_G(z)$ at any z. The fact that the Jacobian matrix is orthogonal indicates that every noise shift $\Delta z : ||\Delta z|| = \varepsilon$ leads to a fixed-size and non-zero step in the image space. This property was shown to induce several attractive properties to GANs, including the stability of training, higer quality and diversity of generated images, as well as the ability for GAN inversion.

In this thesis, the above regularizations are present in some of our models or in our most important comparison baselines. For example, the Gradient penalty is used by default in StyleGANv2, which is the baseline for our model in Chapter 7. Spectral normalization is used in our models OASIS (Chapter 4), SIV-GAN (Chapter 5), and OSMIS (Chapter 6). Lastly, the diversity-sensitive loss

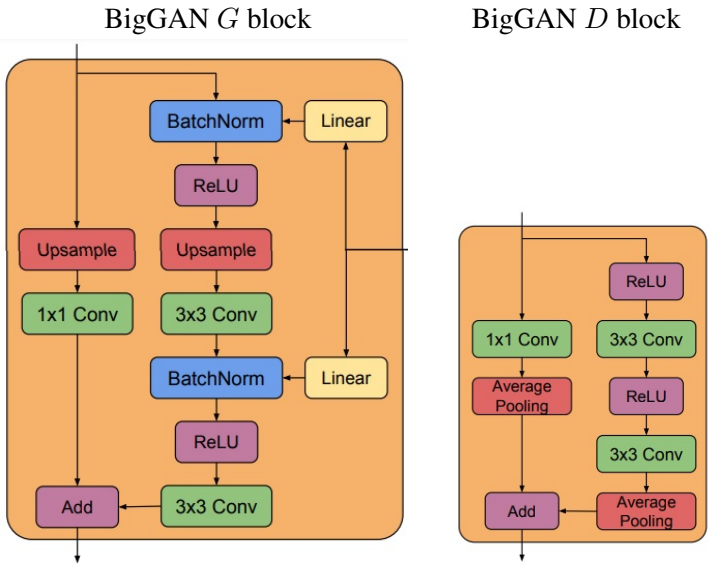

Figure 3.2: The architecture of the BigGAN (*Brock et al.*, 2019) G and D blocks.

and path length regularizer constitute important comparison methods for our proposed Diversity Regularization and Smoothness Similarity regularization introduced in Chapters 5 and 7.

3.2.4 GAN Architectures

Along with adversarial losses and regularizations, the performance of GAN models strongly depends on the exact architectures of G and D used for training. In what follows, we discuss the common neural architecture choices in the GAN literature.

Generator and discriminator blocks. GAN models borrow a lot of solutions from the recent advancements in general deep learning architectures. In state-of-the-art GAN models, both the generator and discriminator are structured as deep neural networks with multitude of layers, organized in repetitive blocks. A typical block usually consists of a skip-connection, convolutional layers with a kernel size of 3×3 and stride 1, non-linearities (e.g., ReLu), up/downsampling, and normalization layers. The exact order of operations and their parameters, however, depends on the model.

One of the most successful GAN architectures is BigGAN (*Brock et al.*, 2019). The structure of BigGAN's residual G and D blocks is shown in Fig. 3.2. In the generator, BigGAN uses conditional Batch normalization layers, ReLU activations, 2×2 bilinear upsampling, and 3×3 convolutions. The output features of the G block are therefore 2 times larger in spatial dimensions than the input. Along the channel dimensions, the output can remain the same or become 2 times smaller, depending on the number of output filters in the second convolution. In the skip-connection, the dimensions of the input features are matched to the output size via an upsampling layer and a 1×1 convolution. In the discriminator, BigGAN employs ReLU activations, 3×3 convolutions, and 2×2 average pooling for downsampling. The skip connection of BigGAN's D block consists of a 1×1 convolution and average pooling, which are used to match the shape of output features between different pathways. Overall, the architecture depicted in Fig. 3.2 influenced many subsequent models, and inspired the design of G and D blocks used in our SIV-GAN and OSMIS models (Chapters 5 and 6).

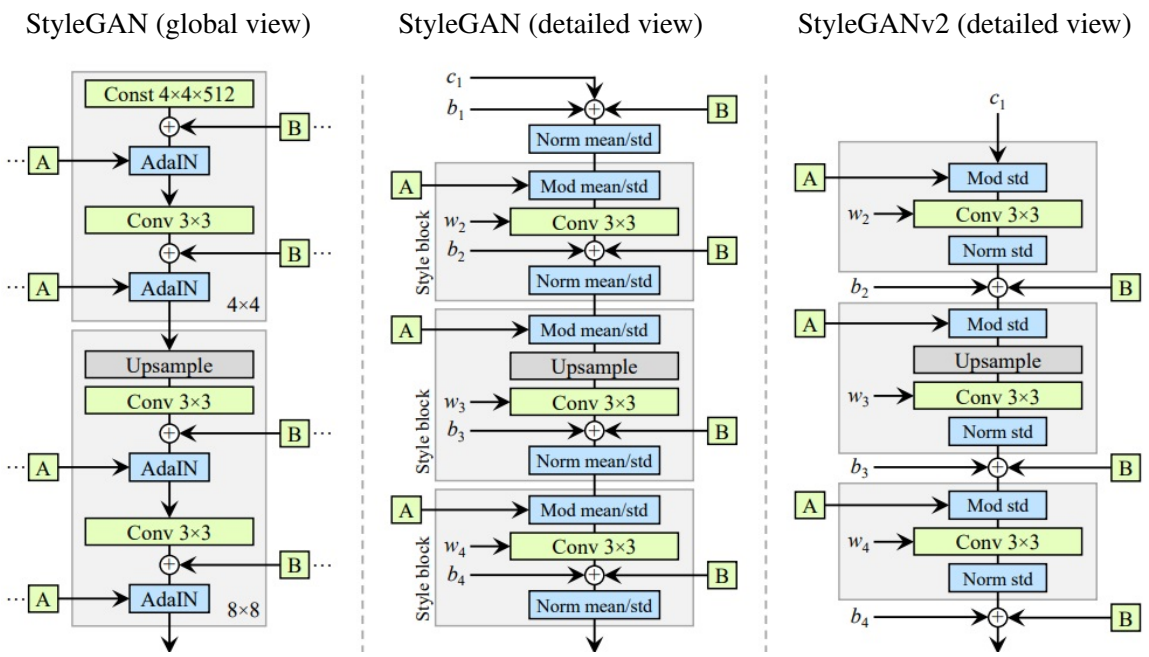

Figure 3.3: The architecture of StyleGAN (*Karras et al.*, 2019) and StyleGANv2 (*Karras et al.*, 2020b) G blocks. Convolutions in both models are followed by a LeakyReLu activation (not shown).

Another influential GAN architecture comes from a series of works on StyleGANs (*Karras et al.*, 2019, 2020b, 2021). The main novelty of StyleGANs lies on the generator's side (see Fig. 3.3). StyleGAN's generator contains Style blocks, consisting of a bilinear upsampling, 3×3 convolutions with LeakyReLu activations, and Adaptive Instance normalizations (*Huang and Belongie*, 2017) to adapt the style of generated images. In StyleGANv2, Adaptive Instance normalization was redesigned to adjust only the standard deviation of feature maps (see Fig. 3.3, right). In this thesis, StyleGANv2 architecture is used in Chapter 7 for the evaluation of the smoothness similarity regularization.

Connections between blocks and between G and D. While the low-level architecture of the G and D block is important, the performance of GANs also depends on how these blocks are connected to each other. Fig. 3.4 demonstrates several popular ways to organize GAN architectures at a higher level. The simplest among them is to use residual networks (Fig. 3.4, right), applying ResNet blocks sequentially one after another in both G and D. This solution is used, for example, in BigGAN. Another approach is to employ skip connections from all G's blocks to corresponding D's blocks (Fig. 3.4, left), as was proposed in MSG-GAN (*Karnewar and Wang*, 2020). This solution was shown to facilitate the flow of the gradient from D to the earliest blocks of G and thus to improve performance. This approach is used in our SIV-GAN and OSMIS models in Chapters 5 and 6. Lastly, *Karras et al.* (2020b) explored a mixed approach, in which a generated image is assembled as sum of images generated at different G blocks, and this image is processed at different resolutions at each D block (Fig. 3.4, center). Such skip-layer strategy is used in the StyleGANv2 generator.

Conditioning of G. Another degree of freedom in the GAN design is how to condition the generator on input noise and, optionally, on other data (e.g., class labels). There are several popular G conditioning mechanisms. The simplest way is to use noise simply as input to the first G block.

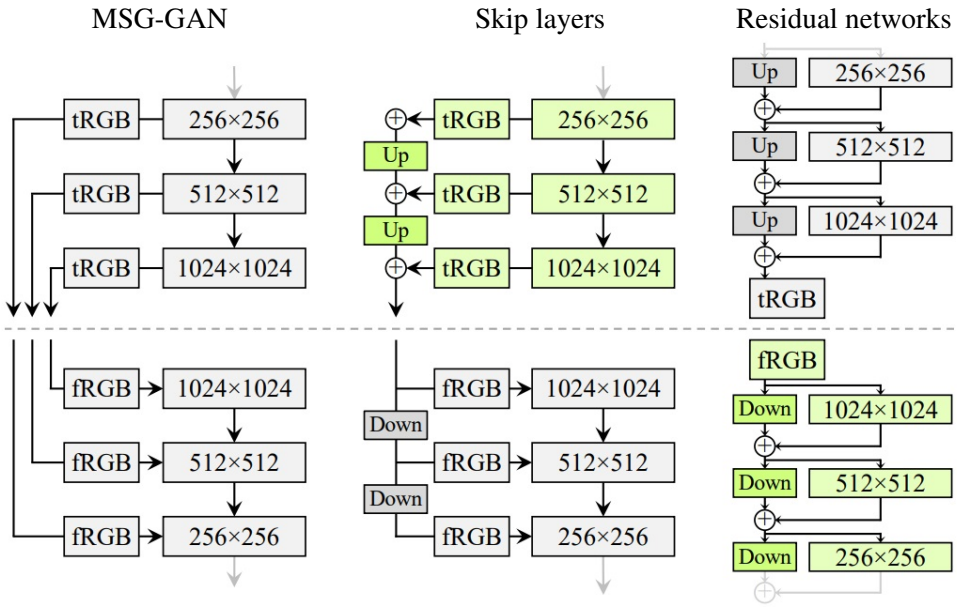

Figure 3.4: Different strategies for connecting the GAN generator and discriminator, as presented in (*Karras et al.*, 2020b). tRGB and fRGB denote convolutions that transform feature representations into images, and vice versa.

In this case, the class label is concatenated to noise as a one-hot vector or as output of a separate encoder. Alternatively, the combination of noise and class label can be projected into various G's blocks via self-modulation in conditional batch normalization (*Chen et al.*, 2019; *Brock et al.*, 2019). In StyleGANs, a separate MLP network maps the conditioning noise to a style space, after which a style vector is injected into each G block via adaptive instance normalization or by adjusting the standard deviation of intermediate features. Overall, the design of G conditioning plays an important role in its sensitivity to noise, diversity of generated images, and its ability to precisely follow the conditioning labels.

In semantic image synthesis, when generated images should adhere to given semantic label maps, G's sensitivity to input labels is especially important. In this field, most recent works inject the input label maps via the spatially-adaptive normalization (SPADE) (*Park et al.*, 2019b). The SPADE layer takes semantic label maps y and intermediate feature representations f of shape N×C×W×H as input (the dimensions N, C, W, H correspond to the batch size, channels, width and height of intermediate features). The output of SPADE is expressed as:

$$\hat{f}_{n,c,x,y} = \gamma_{c,y,x}(y)\frac{f_{n,c,x,y} - \mu_c}{\sigma_c} + \beta_{c,x,y}(y), \tag{3.13}$$

where γ and β are the learned scaling and bias parameters that depend on y, and the μ_c and σ_c are the mean and standard deviation of the activations f in channel c. Essentially, SPADE is a conditional batch normalization layer, in which the learned scaling and bias parameters are learned independently for pixels belonging to different semantic classes. In Chapter 4, we leverage the SPADE layer as the foundation for introducing a new 3D noise injection scheme in our OASIS model.

Discriminator architectures. The GAN discriminator is a neural network designed to classify input

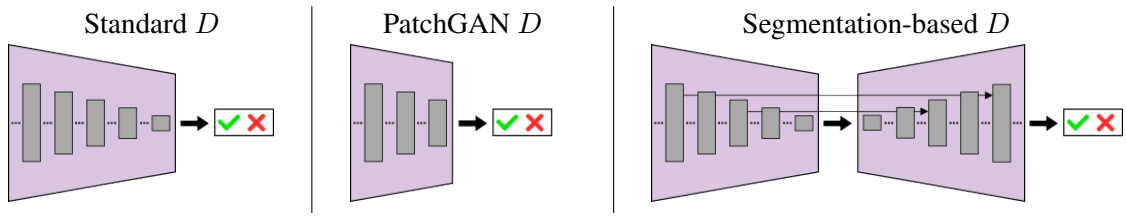

Figure 3.5: Different GAN discriminator architectures.

images as real or fake. It typically takes the form of an encoder, consisting of consecutive blocks $\{D^i\}_{i=1}^N$. When given an image x, the encoder computes a real/fake logit after the last block, denoted as $l = s^N \circ D^N(x)$, where s^N represents the final processing layer, such as convolution. This D architecture, depicted in Fig. 3.5 (left), has a receptive field that covers the entire image, enabling global-level assessment of image realism. This architecture is utilized in various state-of-the-art GAN models, including BigGAN and StyleGANs.

However, relying solely on global-level evaluation may not always yield optimal results. For example, when training data is limited, this approach can lead to easier memorization of the training images. To address this, prior work introduced an alternative solution to mitigate memorization of training data and to encourage more diverse synthesis. This solution involves disregarding the latest D blocks and making the real/fake decision at an intermediate layer: $l = s^k \circ D^k(x)$, where $k < N$. This approach, known as PatchGAN D (*Isola et al.*, 2017) (Fig. 3.5, center), has a limited receptive field, which allows to judge the input images at a more local level. PatchGAN D is used in several important comparison baselines in this thesis, including SPADE (Chapter 4) and CDC (Chapter 7).

The third discriminator architecture related to this thesis is a segmentation-based D. This discriminator utilizes an encoder-decoder architecture (see Fig.3.5, right), enabling it to generate individual real/fake predictions for each pixel of an image. Consequently, segmentation-based discriminators provide a rich training signal for the generator by precisely identifying the areas in generated images that require improvement. The effectiveness of segmentation-based discriminators has been demonstrated in applications such as semantic segmentation (*Souly et al.*, 2017) and unconditional image synthesis (*Schönfeld et al.*, 2020). In Chapter 4, we explore the potential of this architecture for semantic image synthesis and demonstrate its efficacy for the OASIS model.

3.2.5 Useful Techniques for GAN Training

Prior works on GANs also explored numerous training techniques that are complimentary to adversarial losses, regularizations, and network architectures. This section describes the techniques that are the most related to this thesis.

Differentiable image augmentation. In the field of machine learning, data augmentation is a standard technique to mitigate overfitting and improve the generalization of models. In the context of GANs, data augmentation of real data is also commonly employed to mitigate overfitting in discriminator. However, as argued in (*Zhao et al.*, 2020a), in this case the image transformations used for data augmentation can inadvertently affect the generated images. For instance, if vertical flipping is included as a data augmentation technique, the discriminator will encourage the generator to produce vertically flipped images.

Consequently, a challenge arises in striking a balance between preventing unrealistic generations

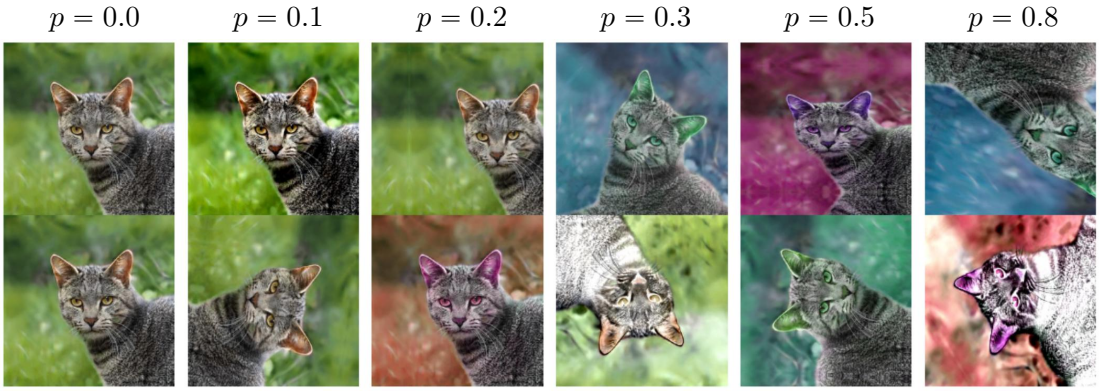

Figure 3.6: Differentiable image augmentation (DA), which is applied to both real and fake images before D takes them as input. DA includes a series of image transformations that are applied consequently, with probability p for each transformation. In the figure, the two rows illustrate possible outcomes of DA with different p, as presented in (*Karras et al.*, 2020a).

and leveraging the benefits of data augmentation. To tackle this issue, *Zhao et al.* (2020a) proposed a novel approach that applies data augmentation to both real and fake images. This strategy involves passing a real or fake image through a sequence of differentiable image transformations $T(x)$, each applied with a fixed probability p. Fig. 3.6 illustrates the outcomes of this image augmentation process, referred to as differentiable augmentation (DA), for various values of p. By increasing p, the discriminator is exposed to a significantly more diverse range of images, effectively preventing overfitting. As demonstrated in (*Karras et al.*, 2020a), if DA is applied during both the G and D training steps, the generator can learn to counteract the applied transformations, even when p is relatively large. In this thesis, DA is employed in all of our models that operate under limited data regimes (Chapters 5-7).

Perceptual loss. Perceptual loss is a widely used technique in paired image-to-image translation tasks, such as semantic image synthesis. A perceptual loss estimates deep features of real and fake images via a pre-trained classification network, aiming to make generated images perceptually closer to real data. In semantic image synthesis, a popular choice is the VGG-16 network (*Simonyan and Zisserman*, 2015), pre-trained on ImageNet (*Deng et al.*, 2009). For given batches of real and fake images x_{real} and x_{gen}, generated from the same label maps, the standard perceptual loss extracts ten feature representations $\{\Phi_i(x_{real})\}_{i=1}^{5}$, $\{\Phi_i(x_{gen})\}_{i=1}^{5}$, where Φ_i correspond to the VGG-16 features extracted at the *ReLu-1,2*, *ReLu-2,2*, *ReLu-3,3*, *ReLu-4,3*, *ReLu-5,3* layers. The perceptual loss is then formulated as:

$$\mathcal{L}_{VGG} = \lambda_{VGG} \cdot \sum_{i=1}^{5} w_i ||\Psi_i(x_{real}) - \Psi_i(x_{gen})||_1, \tag{3.14}$$

where w_i represents the weight assigned to the VGG-16 layer with index i. In prior works on semantic image synthesis, the standard choice for w is $[1/32, 1/16, 1/8, 1/4, 1]$. In Chapter 4, we demonstrate that using a perceptual loss is suboptimal for semantic image synthesis GANs, and propose a new model that achieves good performance without relying on this loss.

Feature matching Loss. Another technique employed in semantic image synthesis is the feature matching loss. This loss operates on a similar principle as the perceptual loss, but instead of using an external VGG-16 feature extractor, it leverages the discriminator network. For the real and fake batches x_{real} and x_{gen}, generated from the same label maps, the discriminator extracts feature representations $\{D^i(x_{real})\}_{i=1}^N$ and $\{D^i(x_{fake})\}_{i=1}^N$. Here, D^i corresponds to the output of the i-th discriminator block, and N represents the total number of blocks in D. The feature matching loss is defined as follows:

$$\mathcal{L}_{FM} = \lambda_{FM} \cdot \sum_{i=1}^{N} ||D^i(x_{real}) - D^i(x_{gen})||_1. \tag{3.15}$$

Similar to our findings regarding the perceptual loss, Chapter 4 demonstrates that the feature matching loss is unnecessary for achieving satisfactory performance in semantic image synthesis. As a result, we exclude this loss from our proposed OASIS model.

3.3 Applications of GAN-Based Image Synthesis

In previous sections, we explored the concept of image synthesis and the use of GANs for generating realistic images. However, the impact of GAN-based image synthesis goes beyond simply achieving impressive visual quality. In this thesis, we will study several downstream applications that can benefit from the high-quality images produced by GANs. The first application is semantic image editing, which enables controlled editing of distinct semantic regions within images. The second set of applications focuses on synthetic data augmentation. This involves using the generated images as a means of augmenting existing data for other computer vision applications. The next sections will provide an overview of the downstream applications that are considered in this thesis.

3.3.1 Semantic Image Editing

Image editing is the process of modifying an existing image according to user-defined specifications. In semantic image synthesis, this task takes a special form, referred to as semantic image editing. In this task, users have the ability to select a specific region of an image belonging to a particular object and generate new images where the chosen object is visually transformed, while the rest of the scene is preserved. More specifically, given an image x, its semantic label map $y \in N^{H \times W}$, and a specified class $c \in [1..N]$, semantic image editing aims to generate new content within the mask $m = \mathbf{1}_{y=c}$ while preserving the content outside of it. Figure 3.7 provides an example of semantic image editing, showcasing modifications made to mountains (upper row) or trees (bottom row) in the given scenes. In Chapter 4 of this thesis, we explore semantic image editing with our OASIS model

3.3.2 Synthetic Data Augmentation

In a well-trained GAN model, the generator produces new synthetic samples that closely resemble images from the original dataset. One notable advantage of well-trained models is their ability to avoid the memorization of training samples and produce new images that were not present in the original dataset. Therefore, the synthetic data can serve as a valuable resource for data augmentation, addressing limitations such as a small dataset size, class imbalance, or limited variation in the

Figure 3.7: Examples of semantic image editing. In the shown images, only the area belonging to one semantic class is modified: *mountain* in the 1st row, *tree* in the 2nd row.

original data. When added to the original data, synthetic samples thus have a potential to improve the robustness and overall performance of models in other computer vision applications. In this thesis, we study the application of synthetic data augmentation to several computer vision tasks (see Fig. 3.8):

Semantic image segmentation. In semantic image segmentation, the task is to train a neural network F that predicts a semantic label map y of a given image x (Fig. 3.8, upper left corner). In the simplest scenario, the network F is trained on a dataset consisting of paired images and label maps $S = (\mathbf{x}_i, y_i)_{i=1}^m$. Notably, this type of data is also employed for training semantic image synthesis models, where a generator G produces an image x based on a provided label map y. Therefore, in Chapter 4, our model OASIS can generate new images $\hat{x} = G(y)$ for each label map y from S, thereby serving as data augmentation $(\hat{x}, y)$ for training F. In Chapter 4, we demonstrate that synthetic data augmentation produced by OASIS helps to improve the performance of semantic segmentation on two datasets, providing larger gains compared to the baseline SPADE.

One-shot semantic image segmentation. The task of one-shot semantic image segmentation aims to adapt a pre-trained semantic segmentation network F to new object classes based on a single labeled example (Fig. 3.8, lower left corner). This is a highly challenging task as the network must learn the appearance of new classes and achieve generalization from just a single image-mask pair (x, y) during the test phase. In the context of GANs, this type of data is used in our newly introduced task OSMIS in Chapter 6. Therefore, we apply our OSMIS model to this task, generating new diverse image-mask pairs $(\hat{x}, \hat{y}) = G(z)$ that contain the same objects as in the original pair (x, y). The high quality and diversity of the OSMIS generations in this one-shot scenario allow us to effectively augment the original data, resulting in significant mIoU gains in the one-shot image segmentation benchmark COCO-20^i (*Lin et al.*, 2014).

One-shot video object segmentation. One-shot video object segmentation is the task which segments objects in videos, provided a mask of objects only in the first video frame (Fig. 3.8, right). At test time, this task provides a network F with a set of video frames $(x_1, ..., x_n)$ and the semantic mask y_1. To generate data augmentation for this task, we train OSMIS on the image-mask pair

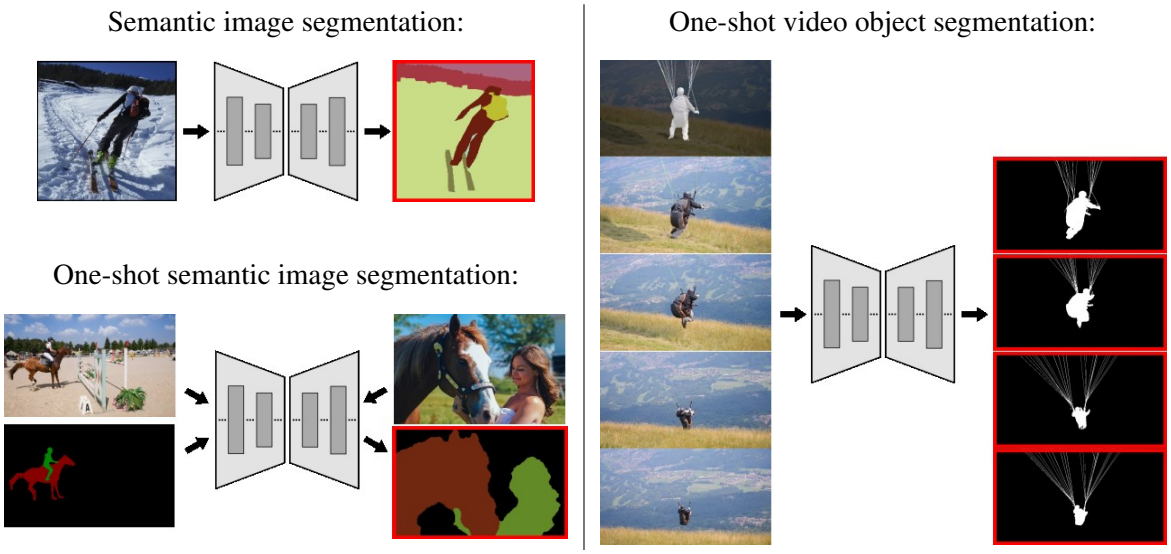

Figure 3.8: Downstream applications for GAN-based synthetic data augmentation studied in this work. For one-shot semantic image segmentation and one-shot video object segmentation tasks, only the test stage is illustrated. For all the tasks, the desired output is highlighted in red frames.

(x_1, y_1). As demonstrated in Chapter 6, the generated image augmentation helps the model to avoid overfitting and achieve better generalization to other video frames.

CHAPTER 4

You Only Need Adversarial Supervision for Semantic Image Synthesis

In this chapter, we focus on semantic image synthesis. This task can be considered as an extension of class-conditional image generation, where the class index is provided for each pixel in an image. Semantic label maps are of great help to the generation process, as they provide the generator with pixel-wise guidance on how to realistically arrange objects in the scene at a pixel level. Thus, ideally, semantic image synthesis GANs should outperform unconditional and class-conditional GANs in terms of synthesis quality and diversity. However, in reality, these GANs often struggle to produce images with correct colors and textures and fail to achieve diversity. In this chapter, we identify that these problems are caused by the overreliance of previous models on additional training techniques like perceptual losses. To this end, in this chapter we introduce a new GAN model that achieves impressive performance using only the adversarial loss. Our model, called OASIS, is simpler than prior models, while offering better performance and introducing several new capabilities.

Individual Contribution

This chapter is based on the following journal and conference publications (*Sushko et al.*, 2022; *Schönfeld et al.*, 2021):

OASIS: Only Adversarial Supervision for Semantic Image Synthesis
Vadim Sushko*, Edgar Schönfeld*, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva
International Journal of Computer Vision (IJCV), 2022. DOI: 10.1007/s11263-022-01673-x

You Only Need Adversarial Supervision for Semantic Image Synthesis
Edgar Schönfeld*, Vadim Sushko*, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva
International Conference on Learning Representations (ICLR), 2021.

This publication was a result of a highly collaborative effort involving Vadim Sushko, Edgar Schönfeld, Dan Zhang, and Anna Khoreva. Bernt Schiele and Juergen Gall provided scientific guidance and generously supported this work with valuable feedback and suggestions. Anna Khoreva initially proposed to explore a segmentation-based discriminator in the context of semantic image synthesis. Subsequently, through extensive joint discussions, all co-authors contributed to the development of this idea, leading to additional proposals. In this publication, Vadim Sushko and Edgar Schönfeld are joint first authors who contributed equally to all aspects of the paper, including discussions, codebase development, ablations, final experiments, evaluations, and paper writing.

Contents

Figure 4.1: Existing semantic image synthesis models heavily rely on the VGG-based perceptual loss to improve the quality of generated images. In contrast, our model (OASIS) can synthesize diverse and high-quality images while only using an adversarial loss, without any external supervision.

4.1 Introduction

Conditional generative adversarial networks (GANs) (*Mirza and Osindero*, 2014) synthesize images conditioned on class labels (*Brock et al.*, 2019; *Casanova et al.*, 2021), text (*Reed et al.*, 2016; *Zhang et al.*, 2018a, 2021a), other images (*Isola et al.*, 2017; *Huang et al.*, 2018; *Park et al.*, 2020b), or semantic label maps (*Park et al.*, 2019b; *Liu et al.*, 2019; *Wang et al.*, 2021b). In this work, we focus on the latter, addressing semantic image synthesis. Taking pixel-level annotated semantic maps as input, semantic image synthesis enables the rendering of realistic images from user-specified layouts, without the use of an intricate graphics engine. Therefore, its applications range widely from content creation and image editing to producing training data for downstream applications that adhere to specific semantic requirements (*Park et al.*, 2019a; *Ntavelis et al.*, 2020).

Despite the recent progress on stabilizing GANs (*Miyato et al.*, 2018; *Zhang and Khoreva*, 2019; *Karras et al.*, 2020a; *Sauer et al.*, 2021) and developing their architectures (*Karras et al.*, 2021, 2019, 2020b; *Brock et al.*, 2019; *Liu et al.*, 2021), state-of-the-art GAN-based semantic image synthesis models (*Park et al.*, 2019b; *Liu et al.*, 2019; *Wang et al.*, 2021b) still greatly suffer from training instabilities and poor image quality when the generator is only trained to fool the discriminator in an adversarial fashion (see Fig. 4.1). An established practice to overcome this issue is to employ a perceptual loss (*Wang et al.*, 2018a) to train the generator, in addition to the discriminator loss. The perceptual loss aims to match intermediate features of synthetic and real images, that are estimated via an external perception network. A popular choice for such a network is VGG (*Simonyan and Zisserman*, 2015), pre-trained on ImageNet (*Deng et al.*, 2009). Although the perceptual loss substantially improves the performance of previous methods, it comes with the computational overhead introduced by utilizing an extra network for training. Moreover, as we show in our experiments, it dominates over the adversarial loss during training, as the generator starts to learn mostly through minimizing the VGG loss, which has a negative impact on the diversity and quality of generated images. Therefore, in this work we propose a novel, simplified model that establishes new state-of-the-art results without requiring a perceptual loss.

To achieve semantic image synthesis of high quality, the training signal to the GAN generator should contain feedback on whether the generated images are well aligned to the input label maps. Thus, a fundamental question for GAN-based semantic image synthesis models is how to design the discriminator that would efficiently utilize information from given semantic label maps, in addition to judging the realism of given images. Conventional methods (*Park et al.*, 2019b; *Wang et al.*, 2018a; *Liu et al.*, 2019; *Isola et al.*, 2017; *Wang et al.*, 2021b; *Ntavelis et al.*, 2020) adopt a multi-scale classification network, taking the label map as input along with the image, and making a global image-level real/fake decision. This discriminator has limited representation power, as it is not incentivized to learn high-fidelity pixel-level details of the images and their precise alignment with the input semantic label maps. For example, such a classification-based discriminator can base its decision solely on image realism, without the need of examining the alignment between the image and label map. To mitigate this issue, we propose an alternative architecture for the discriminator, re-designing it as an encoder-decoder semantic segmentation network (*Ronneberger et al.*, 2015), and directly exploiting the given semantic label maps as ground truth via an (N+1)-class cross-entropy loss. This new discriminator provides semantically-aware pixel-level feedback to the generator, partitioning the image into segments belonging to one of the N real semantic classes or the fake class. With this design, the network cannot ignore the provided label maps, as it has to predict a correct class label for each pixel of an image. Enabled by the discriminator per-pixel response, we further introduce a LabelMix regularization, which fosters the discriminator to focus more on the semantic and structural differences of real and synthetic images. The proposed changes lead to a much stronger discriminator, that maintains a powerful semantic representation of objects, giving more meaningful feedback to the generator, and thus making the perceptual loss supervision superfluous (see Fig. 4.1).

Semantic image synthesis is naturally a one-to-many mapping, where one label map can correspond to many possible real images. Thus, a desirable property of a generator is to generate a diverse set of images from a single label map, only by sampling noise. This property is known as multi-modality. Previously, only using a noise vector as input was not sufficient to achieve multi-modality, because the generator tended to mostly ignore the noise or synthesized images of poor quality (*Isola et al.*, 2017; *Wang et al.*, 2018a). Thus, prior work (*Wang et al.*, 2018a; *Park et al.*, 2019b) resorted to using an image encoder to produce multi-modal outputs. In this work, we enable multi-modal synthesis of the generator via a newly-introduced 3D noise sampling method, without requiring an image encoder and not relying on availability of a reference image to produce new image styles. Empowered by our stronger discriminator, the generator can now effectively synthesize different images by simply resampling a 3D noise tensor, which is used not only as the input, but is also combined with intermediate features via conditional normalization at every layer. This procedure makes the generator spatially sensitive to noise, so we can re-sample it both globally (channel-wise) and locally (pixel-wise), allowing to change not only the appearance of the whole scene, but also of specific semantic classes or any chosen area (see Fig. 4.2). As shown in our experiments, the proposed 3D noise injection

Figure 4.2: OASIS multi-modal synthesis results. 3D noise can be sampled globally (first 2 rows), changing the whole scene, or locally (last 2 rows), partially changing the image. For the latter, we sample different noise per region, like the bed segment (in red) or arbitrary areas defined by shapes.

scheme enables a significantly higher diversity of synthesis compared to previous methods.

With the proposed modifications in the discriminator and generator design, we outperform the prior state of the art in synthesis quality across the commonly used ADE20K (*Zhou et al.*, 2017b), COCO-Stuff (*Caesar et al.*, 2018) and Cityscapes (*Cordts et al.*, 2016) datasets. Omitting the necessity of the VGG perceptual loss, our model generates samples of higher quality and diversity, and follows the color and texture distributions of real images more closely.

A well known challenge for semantic segmentation applications is the problem of class imbalance. In practice, a dataset can contain underrepresented classes (representing a very small fraction of the dataset pixels), which can lead to suboptimal performance of models (*Sudre et al.*, 2017). However, to the best of our knowledge, this problem has not been studied in the context of semantic image synthesis. For this reason, we propose to extend the evaluation setup used in previous works by using the highly imbalanced LVIS dataset (*Gupta et al.*, 2019). Originally introduced as a dataset for long-tailed object recognition, LVIS contains a large set of 1203 classes, the majority of which appear only in a few images. Moreover, to simplify dataset curation, label maps in LVIS were annotated sparsely, with large image areas being occupied with a generic background label. The above properties make LVIS a very challenging evaluation setting for previous semantic image synthesis models, as we demonstrate by the example of the state-of-the-art SPADE model (*Park et al.*, 2019b). As the classification-based discriminator of SPADE makes a global real/fake decision for each image-label pair, the loss contribution originating from underrepresented classes can be dominated by the loss contribution of well represented classes. In contrast, our proposed discriminator mitigates this issue: with the $(N{+}1)$-class cross-entropy loss computed for each image pixel, it becomes possible to assign higher weights for the pixels belonging to underrepresented classes. As shown in our experiments, our model successfully deals with both the extreme class imbalance and sparsity in label maps, outperforming SPADE on the LVIS dataset by a large margin.

To extend the evaluation of our model further, we test the efficacy of generated images when applied as synthetic data augmentation for the training of semantic segmentation networks. This way, the performance

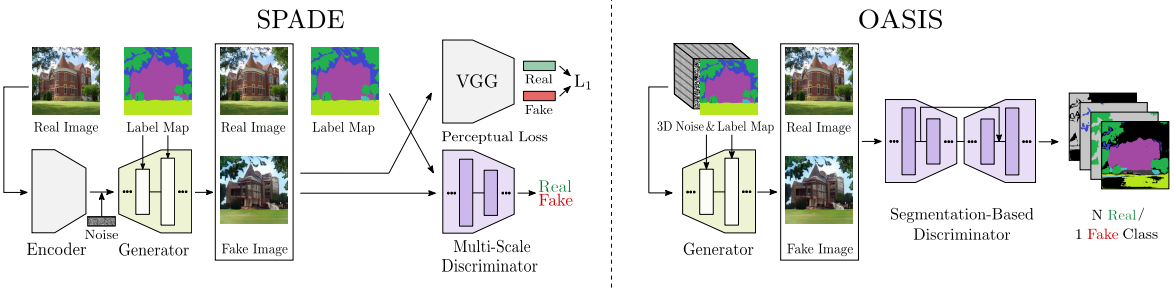

Figure 4.3: SPADE (left) vs. OASIS (right). OASIS outperforms SPADE, while being simpler and lighter: it uses only an adversarial loss as supervision and a single segmentation-based discriminator, without relying on heavy external networks. Furthermore, OASIS learns to synthesize multi-modal outputs by directly re-sampling the 3D noise tensor, instead of using an image encoder as in SPADE.

of semantic image synthesis is assessed through a task that holistically requires high image quality, diversity, and precise image alignment to the label maps. We demonstrate that the synthetic data produced by our model achieves high performance on this test, eliciting a notable increase in downstream segmentation performance. In doing so, our model outperforms a strong baseline SPADE (*Park et al.*, 2019b), indicating its high potential to be applied in segmentation applications. In addition, we also demonstrate how our model for the first time enables the application of a GAN-based semantic image synthesis model to unlabelled images, without requiring external segmentation networks. Thanks to a good segmentation performance of our trained discriminator, we can infer the label map of an image and generate many alternative versions of the same scene by varying the 3D noise. We find these results promising for future utilization of our model in applications.

We call our model OASIS, as it needs **o**nly **a**dversarial supervision for **s**emantic **i**mage **s**ynthesis. In summary, our main contributions include: (i) We propose a novel segmentation-based discriminator architecture, that gives more powerful feedback to the generator and eliminates the necessity of the perceptual loss supervision. (ii) We present a simple 3D noise sampling scheme, notably increasing the diversity of multi-modal synthesis and enabling both complete or partial resampling of a generated image. (iii) With the OASIS model, we achieve high-quality results on the ADE20K, Cityscapes and COCO-Stuff datasets, outperforming previous state-of-the-art models while relying only on adversarial supervision. (iv) We show that images synthesized by OASIS exhibit much higher diversity and more closely follow the color and texture distributions of real images. (v) We propose to use the LVIS dataset (*Gupta et al.*, 2019) to assess image generation in the regime with many underrepresented semantic classes, leading to a severe class imbalance. (vi) We show how the OASIS design directly addresses these issues and thereby outperforms the strong baseline SPADE (*Park et al.*, 2019b) by a large margin. (vii) We test the efficacy of generated images for synthetic data augmentation, as a unified measure that simultaneously depends on image quality, diversity, and label map alignment. The images generated by OASIS elicit a stronger increase in downstream segmentation performance compared to SPADE, suggesting a higher potential of our model for future utilization in applications.

4.2 Method

In this section, we present our OASIS model, which, in contrast to other semantic image synthesis methods, needs only adversarial supervision for training. Using SPADE as a starting point (Sec. 4.2.1), we first propose to re-design the discriminator as a semantic segmentation network, directly using the given semantic label maps as ground truth (Sec. 4.2.2). Empowered by spatially- and semantically-aware feedback of the new discriminator, we next re-design the SPADE generator, enabling its effective multi-modal synthesis via 3D noise sampling (Sec. 4.2.3).

4.2.1 The SPADE Baseline

We choose SPADE as our baseline as it is a state-of-the-art model and a relatively simple representative of conventional semantic image synthesis models. As depicted in Fig. 4.3, the discriminator of SPADE largely follows the PatchGAN multi-scale discriminator (*Isola et al.*, 2017), adopting two image classification networks operating at different resolutions. Both of them take the channel-wise concatenation of the semantic label map and the real/fake image as input, and produce real/fake classification scores. On the generator side, SPADE adopts spatially-adaptive normalization layers to effectively integrate the semantic label map into the synthesis process from low to high scales. Additionally, the image encoder is used to extract the style vector from the reference image, which is then combined with a 1D noise vector for multi-modal synthesis. The training loss of SPADE consists of three terms, namely, an adversarial loss, a feature matching loss and the VGG-based perceptual loss:

$$\mathcal{L} = \max_{G} \min_{D} \mathcal{L}_{\text{adv}} + \lambda_{\text{FM}} \mathcal{L}_{\text{FM}} + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}}. \tag{4.1}$$

Overall, SPADE is a resource-demanding model at both training and test time, i.e., with two PatchGAN discriminators, an image encoder in addition to the generator, and the VGG loss. In the following, we revisit its architecture and introduce a simpler and more efficient solution that offers better performance and reduces the model complexity.

4.2.2 The OASIS Discriminator

To train the generator to synthesize high-quality images that are well aligned with the input semantic label maps, we need a powerful discriminator that coherently captures discriminative semantic features at different image scales. While classification-based discriminators, such as PatchGAN, take label maps as input concatenated to images, they can afford to ignore them and make the decision solely on image patch realism. Thus, we propose to cast the discriminator task as a multi-class semantic segmentation problem to directly utilize label maps for supervision, and accordingly alter its architecture to an encoder-decoder segmentation network (see Fig. 4.3). Encoder-decoder networks have proven to be effective for semantic segmentation (*Badrinarayanan et al.*, 2016; *Chen et al.*, 2018). Thus, we build our discriminator architecture upon U-Net (*Ronneberger et al.*, 2015), which consists of the encoder and decoder connected by skip connections. This discriminator architecture is multi-scale through its design, integrating information over up- and down-sampling pathways as well as through the encoder-decoder skip connections. The segmentation task of the discriminator is formulated to predict the per-pixel class label of the real images, using the given semantic label maps as ground truth. In addition to the N semantic classes from the label maps, all pixels of fake images are categorized as one extra class. As the formulated semantic segmentation problem has $N+1$ classes, we propose to use an (N+1)-class cross-entropy loss for training.

In practice, the N semantic classes are often imbalanced, as some of the classes represent significantly less pixels of the dataset compared to others. The loss contribution for such underrepresented classes can be dominated by well represented classes, which can lead to suboptimal performance. To mitigate this issue, empowered by the pixel-level loss computation of our discriminator, we propose to weight each class by its inverse pixel-wise frequency in a batch, thus giving underrepresented semantic classes more weight. In doing so, the loss contributions of each class are equally balanced, and, thus, the generator is also encouraged to pay more attention to underrepresented classes. Mathematically, the new discriminator loss is expressed as:

$$\mathcal{L}_D = -\mathbb{E}_{(x,t)}\left[\sum_{c=1}^{N} \alpha_c \sum_{i,j}^{H\times W} t_{i,j,c} \log D(x)_{i,j,c}\right] - \mathbb{E}_{(z,t)}\left[\sum_{i,j}^{H\times W} \log D(G(z,t))_{i,j,c=N+1}\right], \tag{4.2}$$

where x denotes the real image; (z,t) is the noise-label map pair used by the generator G to synthesize a fake image; and the discriminator D maps the real or fake image into a per-pixel (N+1)-class prediction

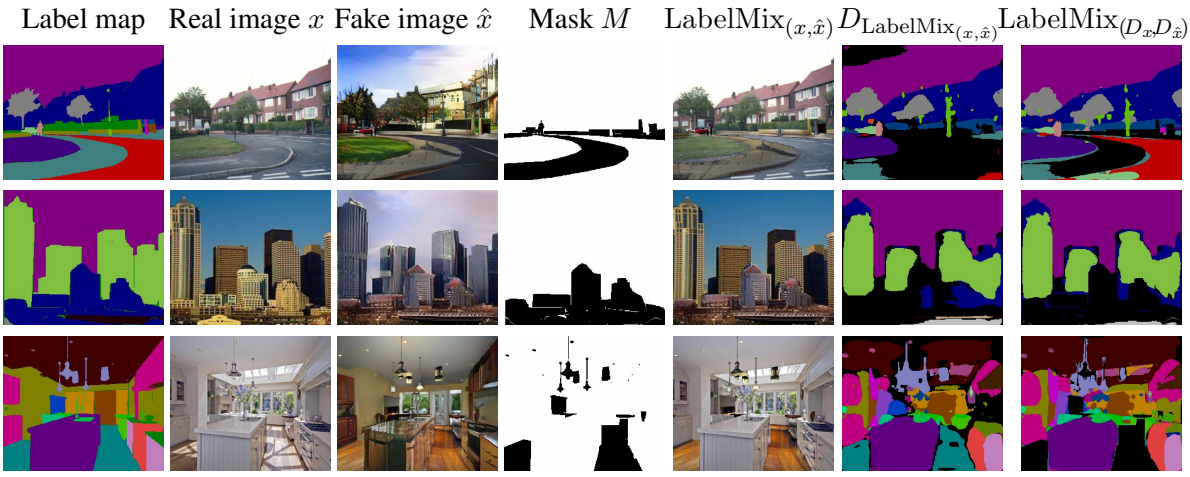

Figure 4.4: LabelMix regularization. Real x and fake $\hat{x}$ images are mixed using a binary mask M, sampled based on the label map, resulting in $\mathrm{LabelMix}_{(x,\hat{x})}$. The consistency regularization minimizes the L2 distance between the logits of $D_{\mathrm{LabelMix}_{(x,\hat{x})}}$ and $\mathrm{LabelMix}_{(D_x,D_{\hat{x}})}$. In this visualization, **black** corresponds to the fake class in the N+1 segmentation output.

probability. The ground truth label map t has three dimensions, where the first two correspond to the spatial position $(i,j) \in H \times W$, and the third one is a one-hot vector encoding the class $c \in \{1,..,N+1\}$. The class balancing weight α_c is the inverse pixel-wise frequency of a class c per batch:

$$\alpha_c = \frac{H \times W}{\sum_{i,j}^{H \times W} E_t\left[\mathbb{1}[t_{i,j,c}=1]\right]}. \tag{4.3}$$

In effect, improving the synthesis of underrepresented and well represented classes is equally necessary to minimize the loss. As we show in Sec. 4.3.3, this step helps to improve the synthesis quality of underrepresented classes.

LabelMix regularization. In order to encourage our discriminator to focus on differences in content and structure between the fake and real classes, we propose a LabelMix regularization. Based on the semantic layout, we generate a binary mask M to mix a pair $(x,\hat{x})$ of real and fake images conditioned on the same label map: $\mathrm{LabelMix}(x,\hat{x},M) = M \odot x + (1-M) \odot \hat{x}$, as visualized in Fig. 4.4. Given the mixed image, we further train the discriminator to be equivariant under the $\mathrm{LabelMix}$ operation. This is achieved by adding a consistency loss term $\mathcal{L}_{cons}$ to Eq. 4.2:

$$\mathcal{L}_{cons} = \Big\| D_{\mathrm{logits}}\Big(\mathrm{LabelMix}(x,\hat{x},M)\Big) - \mathrm{LabelMix}\Big(D_{\mathrm{logits}}(x), D_{\mathrm{logits}}(\hat{x}), M\Big)\Big\|^2, \tag{4.4}$$

where D_{logits} are the logits attained before the last $\mathrm{softmax}$ activation layer, and $\|\cdot\|$ is the L_2 norm. This consistency loss compares the output of the discriminator on the LabelMix image with the LabelMix of its outputs, penalizing the discriminator for inconsistent predictions. LabelMix is different to CutMix (*Yun et al.*, 2019), which randomly samples the binary mask M. A random mask will introduce inconsistency between the pixel-level labels and the scene layout provided by the label map. For an object with the class label c, it will contain pixels from both real and fake images, resulting in two labels, i.e. c and $N+1$. To avoid such inconsistency, the mask of LabelMix is generated according to the label map, providing natural borders between semantic regions, see Mask M in Fig. 4.4. Under LabelMix regularization, the generator is encouraged to respect the natural semantic boundaries, improving pixel-level realism while also considering the class segment shapes.

Alternative ways to encode label maps. Besides the proposed (N+1)-class cross entropy loss, there are other

ways to incorporate a label map into the training of a segmentation-based discriminator. One can concatenate the label map to the input image, analogous to SPADE. Another option is to use projection, by taking the inner product between the last linear layer output and the embedded label map, analogous to class-label conditional GANs (*Miyato and Koyama*, 2018). For both alternatives, the training loss is the pixel-level real/fake binary cross-entropy (*Schönfeld et al.*, 2020). As in these two variants the label maps are used as input to the discriminator (concatenated to the input image or fed to the last linear layer), they are propagated *forward* through the network. In contrast, the (N+1)-setting uses label maps only as targets for the loss computation, so they are propagated *backward* through the network via the gradients updates. Backward propagation ensures that the discriminator learns semantic-aware features, in contrast to forward propagation, where the alignment of a generated image to the input label map can be ignored. The comparison between the above label map encodings is shown in Table 4.10.

4.2.3 The OASIS Generator

To stay in line with the OASIS discriminator design, the training loss for the generator is changed to

$$\mathcal{L}_G = -\mathbb{E}_{(z,t)}\left[\sum_{c=1}^{N}\alpha_c \sum_{i,j}^{H\times W} t_{i,j,c}\log D(G(z,t))_{i,j,c}\right], \tag{4.5}$$

which is a direct outcome of the non-saturation trick (*Goodfellow et al.*, 2014) to Eq. 4.2. We next re-design the generator to enable multi-modal synthesis through noise sampling. SPADE is deterministic in its default setup, but can be trained with an extra image encoder to generate multi-modal outputs. We introduce a simpler version, that enables synthesis of diverse outputs directly from input noise. For this, we construct a noise tensor of size $M\times H\times W$, matching the spatial dimensions of the label map of size $N\times H\times W$, where N is the number of semantic labels and $H \times W$ corresponds to the height and width of the image. Note that for simplicity during training we sample the 3D noise tensor globally, i.e. per-channel, replicating each channel value spatially along the height and width of the tensor. In other words, a M-dimensional latent vector is sampled and then broadcasted to each pixel of an image. After sampling, the noise and the label map are concatenated along the channel dimensions to form a combined noise-label 3D tensor of size $(M+N)\times H\times W$. This combined tensor serves as input to the first generator layer, but also as input to the spatially-adaptive normalization layers in every generator block. This way, all intermediate feature maps are conditioned on both the semantic labels and the noise (see Fig. 4.3), making the noise hard to ignore. As the 3D noise is channel- and pixel-wise sensitive, at test time, one can sample the noise globally, per-channel, and locally, per-segment or per-pixel, for controlled synthesis of the whole scene or of specific semantic objects. For example, when generating a scene of a bedroom, one can re-sample the noise locally and change the appearance of the bed alone (see Fig. 4.2).

Note that using image styles via an encoder, as in SPADE, is also possible in our setting, as the 3D noise can be simply concatenated to the encoder style features. Lastly, to further reduce the complexity, we remove the first residual block in the generator, reducing the number of parameters from 96M to 72M without a noticeable performance loss (see Table 4.8).

4.3 Experiments

We provide an extensive experimental evaluation of our contributions, using the official implementation of SPADE[1] as our baseline. The setup of our experiments is described in detail in Sec. 4.3.1. Firstly, we compare OASIS with other methods on common semantic image synthesis benchmark datasets, comparing their performance in terms of both image quality and diversity (Sec. 4.3.2). To further highlight the advantages of OASIS over the SPADE baseline, we provide additional discussions on different aspects of the semantic image

[1] `github.com/NVlabs/SPADE`

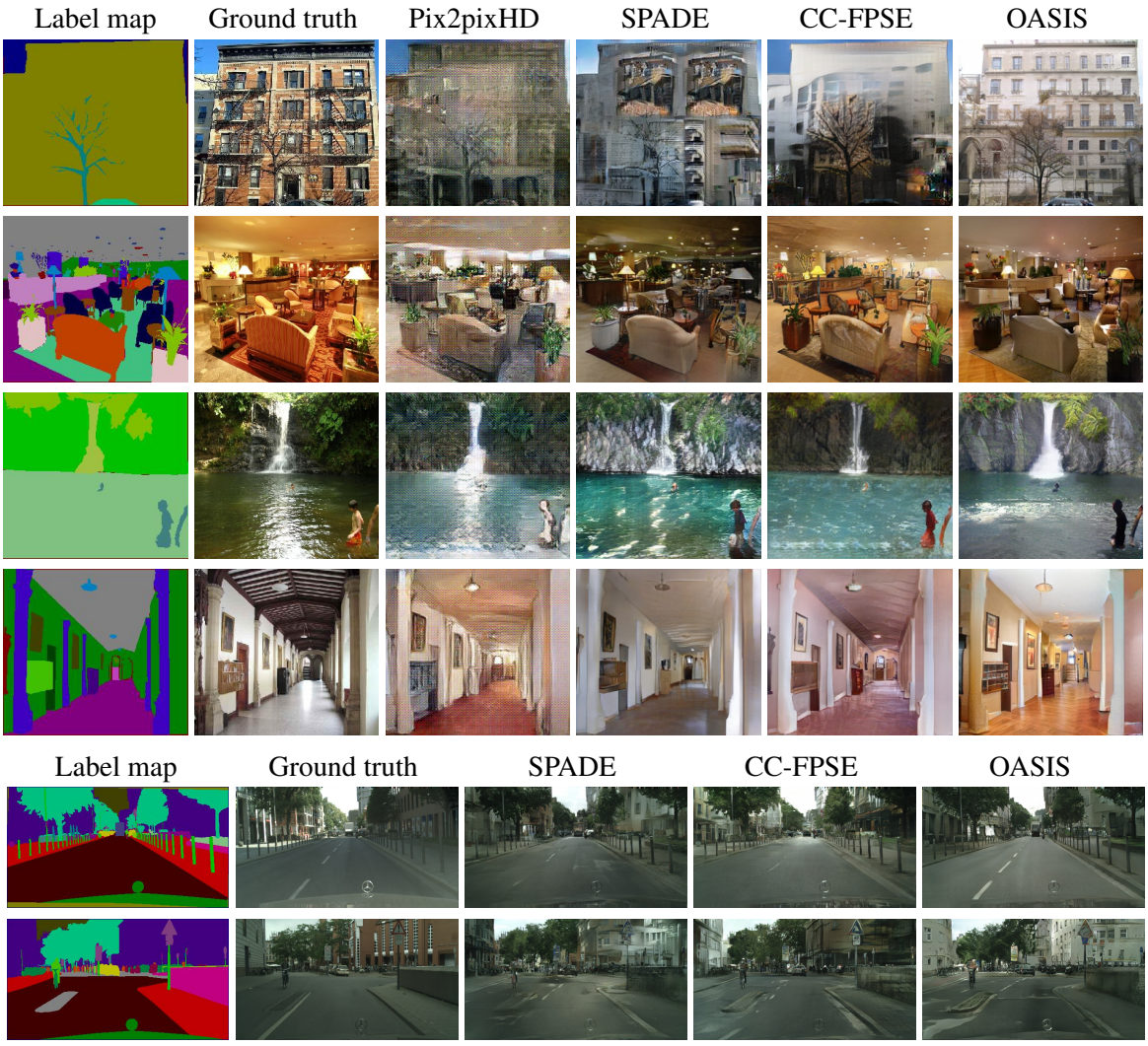

Figure 4.5: Qualitative comparison of OASIS with other methods on ADE20K and Cityscapes. Trained with only adversarial loss, OASIS generates images with better visual quality and structure.

synthesis. In particular, Sec. 4.3.3 is devoted to the performance analysis on the underrepresented classes, extending the comparison of the models to the LVIS dataset (*Gupta et al.*, 2019). Sec. 4.3.4 demonstrates new semantic image editing techniques enabled by OASIS. Sec. 4.3.5 explores the application of generated images as synthetic data augmentation for the training of semantic segmentation networks. Lastly, we provide an extensive ablation study to verify the effectiveness of the proposed contributions (Sec 4.3.6).

4.3.1 Experimental Setup

Datasets. We conduct experiments on several challenging datasets. Firstly, to compare OASIS with other models, we use the ADE20K (*Zhou et al.*, 2017b), COCO-Stuff (*Caesar et al.*, 2018) and Cityscapes (*Cordts et al.*, 2016), which are the three benchmark datasets commonly used in the semantic image synthesis literature (see Sec. 4.3.2). The image resolution is set to 256x256 for ADE20K and COCO-Stuff, and 256x512 for experiments on Cityscapes. Following *Qi et al.* (2018), we also evaluate OASIS on ADE20K-outdoors, the subset of ADE20K containing only outdoor scenes.

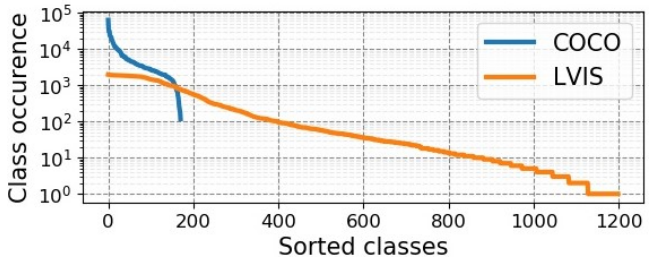

Figure 4.6: Comparison of class distributions of the COCO and LVIS datasets. LVIS has a much larger vocabulary of 1203 classes with a long tail of underrepresented classes.

Secondly, to test the capability of models to learn underrepresented classes, we conduct additional evaluations on the ADE20K and LVIS dataset (*Gupta et al.*, 2019) (see Sec. 4.3.3). We select ADE20K among conventional datasets for its notable class imbalance, as among its 150 classes, more than 86% of the image pixels belong only to the 30 best represented ones (see Table 4.4). In addition, to test the networks under more extreme class imbalance, we propose to use LVIS, the dataset that has been originally introduced for the task of long-tailed instance segmentation. LVIS employs the same set of training images as COCO-Stuff, but its annotations are different in two important ways. Firstly, LVIS provides a significantly larger set of 1203 annotated classes, following a long-tailed distribution in which some classes are present only in one or a few training samples (see Fig. 4.6). Secondly, due to a fixed labelling budget, different background types were not considered for annotation in LVIS. Consequently, the images in LVIS dataset contain large areas belonging to the background class, which sometimes covers more than 90% of the pixels in an image (see grey areas in Fig. 4.9). For the above two reasons, the structure of LVIS poses a new challenge for semantic image synthesis, as models need to account for a much more extreme class imbalance. We conduct experiments on LVIS at the image resolution of 128x128.

Training. We follow the experimental setting of *Park et al.* (2019b). The Adam (*Kingma and Ba*, 2015) optimizer was used with momenta $\beta = (0, 0.999)$ and constant learning rates $(0.0001, 0.0004)$ for G and D. We did not use the GAN feature matching loss for OASIS, as we did not observe any improvement with it, and used the VGG loss only for ablations with $\lambda_{\text{VGG}} = 10$. The parameter for LabelMix λ_{LM} was set to 5 for ADE20k and Cityscapes, and to 10 for COCO-Stuff and LVIS. The latent dimension M was set to 64. We did not experience any training instabilities and, thus, did not employ any extra stabilization techniques. All our models use an exponential moving average (EMA) of the generator weights with 0.9999 decay. All the experiments were run on 4 Tesla V100 GPUs, with a batch size of 20 for Cityscapes and 32 for the other datasets. The training epochs are 200 on ADE20K and Cityscapes, and 100 for the larger COCO-Stuff and LVIS datasets. On average, a complete forward-backward pass with batch size 32 on ADE20k takes around 0.95ms per image.

Evaluation metrics. Following prior work (*Park et al.*, 2019b; *Liu et al.*, 2019), we evaluate the *quality* of semantic image synthesis by computing the FID (*Heusel et al.*, 2017) and evaluate the *alignment* of the generated images with their semantic label maps via mIoU (mean intersection-over-union) or mAP (mean average precision) on the test set (see Sec. 4.3.2). mIoU evaluates the alignment of generated images with their ground truth label maps, as measured by an external pre-trained semantic segmentation network. We use UperNet101 (*Xiao et al.*, 2018) for ADE20K, multi-scale DRN-D-105 (*Yu et al.*, 2017) for Cityscapes, and DeepLabV2 (*Chen et al.*, 2015) for COCO-Stuff. Differently, for the LVIS dataset, the alignment of generated images to ground truth label maps is measured using mAP instead of mIoU, following the official guidelines for evaluating instance segmentation models on this dataset (see Sec. 4.3.3). We compute mAP using a state-of-the-art instance segmentation model from *Wang et al.* (2021a), pre-trained on LVIS.

In addition, to better understand how the perceptual loss influences the synthesis performance, we propose to compare the *color and texture statistics* of generated and real images. For this, we compute color

Method	# param	VGG	ADE20K		ADE-outd.		Cityscapes		COCO-stuff	
			FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
CRN	84M	✓	73.3	22.4	99.0	16.5	104.7	52.4	70.4	23.7
SIMS	56M	✓	n/a	n/a	67.7	13.1	49.7	47.2	n/a	n/a
Pix2pixHD	183M	✓	81.8	20.3	97.8	17.4	95.0	58.3	111.5	14.6
LGGAN	n/a	✓	31.6	41.6	n/a	n/a	57.7	68.4	n/a	n/a
CC-FPSE	131M	✓	31.7	43.7	n/a	n/a	54.3	65.5	19.2	41.6
SC-GAN	66M	✓	29.3	45.2	n/a	n/a	49.5	66.9	18.1	42.0
SESAME	104M	✓	31.9	**49.0**	n/a	n/a	54.2	66.0	n/a	n/a
SPADE	102M	✓	33.9	38.5	63.3	30.8	71.8	62.3	22.6	37.4
SPADE+	102M	✓	32.9	42.5	51.1	32.1	47.8	64.0	21.7	38.8
		✗	60.7	21.0	65.4	22.7	61.4	47.6	99.1	16.1
OASIS	94M	✗	**28.3**	48.8	**48.6**	**40.4**	**47.7**	**69.3**	**17.0**	**44.1**

Table 4.1: Comparison with other GAN methods across datasets. Bold denotes the best performance.

histograms in the LAB space and measure the earth mover's distance between the real and generated image sets (*Rubner et al.*, 2000). We also measure the texture similarity to the real data as the χ^2-distance between Local Binary Patterns histograms (*Ojala et al.*, 1996). As different semantic classes have different color and texture distributions, we aggregate the histogram distances separately per class and compute their average.

To measure the *diversity* among synthesized samples in the multi-modal generation regime, we evaluate MS-SSIM (*Wang et al.*, 2003) and LPIPS (*Zhang et al.*, 2018b) between the images generated from the same label map. For each label map in the test set, we generate 20 images and compute the mean pairwise scores. For the final numbers, the scores are averaged over all label maps.

Lastly, we propose to test the efficacy of generated images when applied as *synthetic data augmentation* for the task of semantic segmentation (see Sec. 4.3.5). For this, we take a DeepLab-V3 segmentation network with a ResNeSt-50 backbone (*Zhang et al.*, 2020b) and train it on ADE20K and Cityscapes. At each training step of DeepLab-V3, we add for each training image its synthetic counterpart to the batch, generated from the same label map. The efficacy of synthetic images is therefore measured by its effect on the downstream mIoU performance of DeepLab-V3.

4.3.2 Evaluation of the Synthesis Quality and Diversity

In this section, we compare OASIS to other state-of-the-art methods, focusing on GANs. For a fair comparison to the baseline SPADE, we additionally train this model without the feature matching loss and using EMA (*Yazici et al.*, 2018) at the test phase. We refer to this improved baseline as SPADE+. In addition, the end of this section provides a quantitative comparison between OASIS and most recent diffusion models.

Synthesis quality. Table 4.1 compares the image synthesis quality achieved by OASIS and other GAN methods. In this table, we report the results of our evaluation for OASIS and SPADE+, and the officially reported numbers for all the other models. As seen from Table 4.1, OASIS outperforms prior state-of-the-art GAN models in FID on all benchmark datasets. Our model also achieves the highest mIoU scores on three out of four datasets, being almost on par with the highest score on ADE20K achieved by SESAME (*Ntavelis et al.*, 2020) Importantly, OASIS achieves the improvement using only adversarial supervision from its segmentation-based discriminator. On the contrary, in the absence of the VGG loss, the baseline SPADE+ does not produce images of high visual quality (see Fig. 4.1), with two-digit drops in FID scores observed for all the datasets in Table 4.1. The strong adversarial supervision also allows OASIS to produce images with color and texture distributions closer to the real data. Such improvement over SPADE+ on the ADE20K dataset is shown in Fig. 4.7,

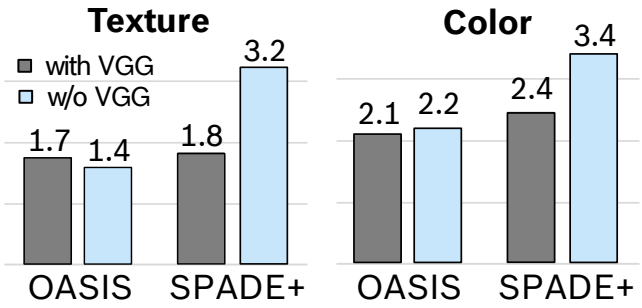

Figure 4.7: Histogram distances to real data on the ADE20K validation set. While SPADE+ relies on the VGG loss to learn colors and textures, OASIS achieves low scores without it.

Method	Multi-mod.	VGG	MS-SSIM↓	LPIPS↑	FID↓	mIoU↑
SPADE+	Encoder	✓	0.85	0.16	33.4	40.2
SPADE+	3D noise	✗	**0.35**	**0.50**	58.4	18.7
		✓	0.53	0.36	34.4	36.2
OASIS	3D noise	✗	0.65	0.35	**28.3**	48.8
		✓	0.88	0.15	31.6	**50.8**

Table 4.2: Multi-modal synthesis evaluation on ADE20K. Bold and red numbers show the best and the worst performance.

where OASIS achieves the lowest color and texture distances to the target distribution. In contrast, SPADE+ needs to compensate a weaker discriminator signal with the VGG loss, struggling to learn the color and texture distribution of real images without it (see Fig. 4.7).

Fig. 4.5 shows a qualitative comparison of our results to previous models. Our approach noticeably improves image quality, synthesizing finer textures and more natural colors. While the previous methods occasionally produce areas with unnatural checkerboard artifacts, OASIS generates large objects and surfaces with higher photorealism. Notably, the improvement over previous models is especially remarkable for the semantic classes that occupy large areas, e.g, wall (rows 1,4 in Fig. 4.5), road (rows 5,6) or water (row 3).

Synthesis diversity. By resampling the input 3D noise, OASIS can produce diverse images given the same label map (see Fig. 4.2). To measure the diversity of such multi-modal synthesis, we evaluate MS-SSIM (*Wang et al.*, 2003) and LPIPS (*Zhang et al.*, 2018b). The lower the MS-SSIM and the higher the LPIPS scores, the more diverse the generated images are. As seen from Table 4.2, OASIS outperforms SPADE+ in both diversity metrics, improving the MS-SSIM scores from 0.85 to 0.65 and LPIPS from 0.16 to 0.35. To assess the effect of the perceptual loss and the noise sampling on diversity, we train SPADE+ with 3D noise or the image encoder, and with or without the perceptual loss. Table 4.2 shows that OASIS, without the perceptual VGG loss, improves over SPADE+ with the image encoder, both in terms of image diversity (MS-SSIM, LPIPS) and quality (mean FID, mIoU across 20 realizations). Using 3D noise further increases diversity for SPADE+. However, a strong quality-diversity trade-off exists for SPADE+: 3D noise improves diversity at the cost of quality, and the perceptual loss improves quality at the cost of diversity. We conclude that our 3D noise injection strongly improves the synthesis diversity, while the VGG loss decreases it.

While the increased diversity is a big advantage, it can also lead to failures in rare cases: for some samples the colors and textures of objects may lie further from the real distribution and seem unnatural to the human eye (see Fig. 4.8).

Comparison to diffusion models. In addition to the previous comparisons with GAN models, Table 4.3 provides a qualitative comparison of OASIS to the diffusion models SDM (*Wang et al.*, 2022b), PITI (*Wang*

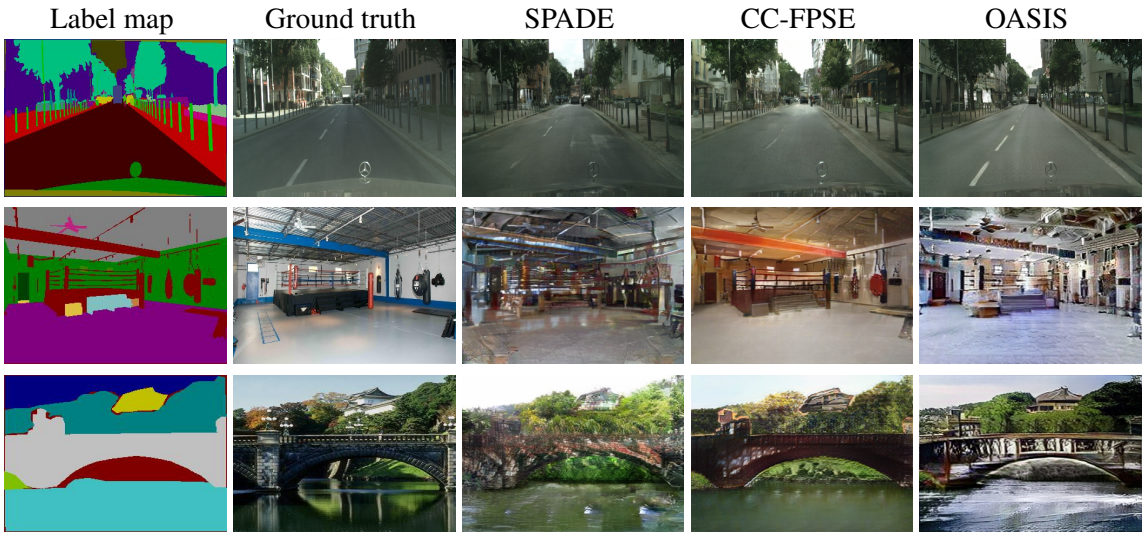

Figure 4.8: Failure mode of OASIS. Without the VGG loss, OASIS has less constraints on the diversity in colors and textures. This helps to achieve higher diversity among the generated samples, but sometimes leads to synthesis of objects with outlier colors and textures which may look less realistic compared to (*Park et al.*, 2019b) and (*Liu et al.*, 2019).

Method	Type	ADE20K			COCO-stuff		
		FID↓	mIoU↑	LPIPS↑	FID↓	mIoU↑	LPIPS↑
OASIS	GAN	28.3	**48.8**	0.35	17.0	**44.1**	0.42
SDM	DM	27.3	39.2	0.52	15.8	40.2	0.52
PITI*	DM	27.9	29.4	0.48	16.1	34.1	0.52
FreestyleNet*	DM	**25.0**	41.9	**0.59**	**14.4**	40.7	**0.59**

Table 4.3: Comparison with diffusion models. Bold denotes the best performance. * indicates that a model was pre-trained on a very large text-image data corpus and not trained from scratch.

et al., 2022a), and FreestyleNet (*Xue et al.*, 2023). It is worth noting that diffusion models are a relatively recent technology compared to GANs, and the aforementioned models were published significantly later than OASIS. Hence, the diffusion models outperform OASIS in terms of overall image quality and diversity, as seen in FID and LPIPS scores in Table 4.3. It is notable that their improvement in LPIPS is particularly significant, indicating the potential of diffusion models to address common diversity issues in GANs, such as reduced recall and mode collapse. On the other hand, we note that diffusion models lag behind OASIS in the mIoU measure. This suggests that better mechanisms for the injection of label maps into diffusion models are required. Other disadvantages of diffusion models include their very slow sampling time (tens of seconds per image) and requirements for very large pre-training datasets (e.g., as in PITI and FreestyleNet), which can complicate their adaptation to datasets of rare structures.

As this thesis focuses on GANs, in the next sections we concentrate on comparing OASIS with other GAN models, primarily with our baseline SPADE+.

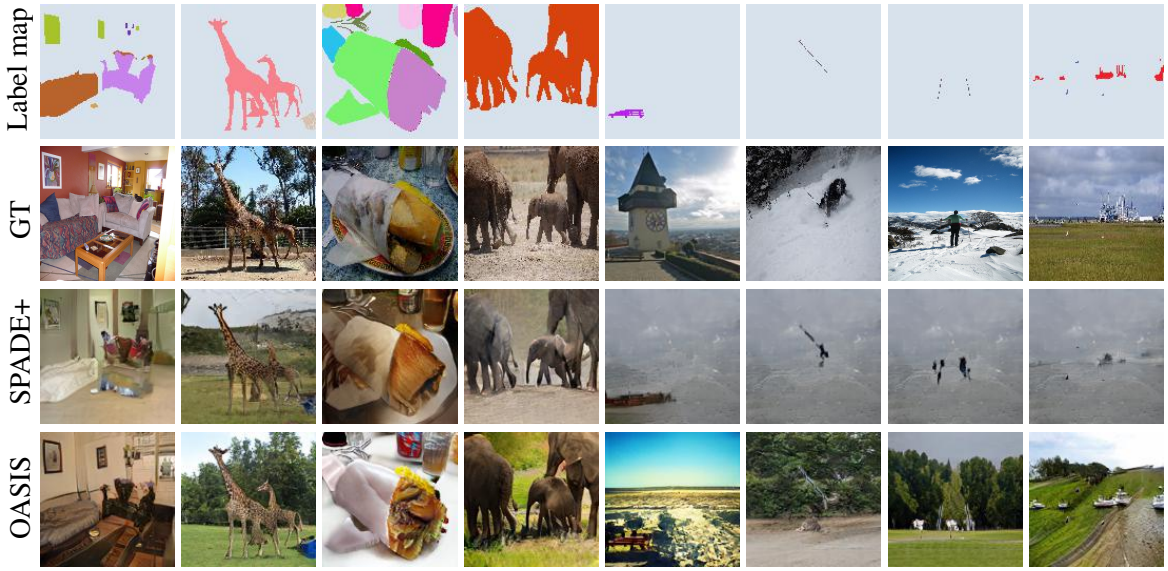

Figure 4.9: Qualitative comparison between OASIS and SPADE+ on the long-tailed LVIS dataset with 1203 classes. OASIS generates higher-quality images with more natural colors and textures. For label maps covered mostly by the background class (four right columns), OASIS hallucinates plausible and diverse images, while SPADE+ suffers from mode collapse.

4.3.3 Synthesis Performance on Underrepresented Classes

Class imbalance is a well-known challenge in semantic segmentation applications (*Sudre et al.*, 2017). Similarly to semantic segmentation, to ensure good performance in real-life test scenarios, semantic image synthesis models should account for a possible dataset class imbalance, especially considering that GANs are notorious for dropping modes of training data (*Arjovsky and Bottou*, 2017). However, to the best of our knowledge, this issue was not addressed in prior works. Thus, in what follows, we evaluate the performance of OASIS and SPADE+ on the ADE20K and LVIS datasets, considering their class imbalances. While the class imbalance in ADE20K is notable (e.g., 86.4% of all image pixels belongs to the 30 best represented classes), this issue is much more amplified in LVIS, which has a long tail of underrepresented classes (see Fig. 4.6).

Evaluation on ADE20K. OASIS significantly outperforms the SPADE+ baseline in the alignment between generated images and label maps, as measured by mIoU (see Table 4.1). As shown in Table 4.4, the improvement in mIoU on ADE20K comes mainly from the better IoU scores achieved for underrepresented semantic classes. To illustrate this, the semantic classes are sorted by their pixel-wise frequency in the training images, obtained by dividing the number of pixels a class occupies in the dataset by the total number of pixels of all images (2nd column in Table 4.4). Table 4.4 highlights that the relative gain in mIoU is especially high for the groups of underrepresented semantic classes, that cover less than 3% of all pixels in the dataset. For these classes, the relative gain over the SPADE+ baseline exceeds 40%. Remarkably, the gain for this group mainly comes from the per-class balancing applied in the OASIS loss function (columns "w/o α_c" and "w. α_c"), which draws the attention of the discriminator to underrepresented semantic classes, thus allowing a higher quality of their generation. This class balancing computes a weight α_c for the losses of each class c on a per-batch basis, for which the total number of pixels in a given batch is divided by the number of pixels belonging to the class (see Eq. 4.2 and 4.3). We note that the possibility to introduce the pixel-wise frequency based balancing requires the loss to be computed separately for each image pixel. This is a unique property of the OASIS discriminator, in contrast to conventional classification-based discriminators, which have to evaluate realism with a single score for images containing both well- and underrepresented classes together.

Evaluation on LVIS. A quantitative comparison between the models on the LVIS dataset is shown in Table

Classes IDs	Pixel-wise frequency	mIoU		
		SPADE+	OASIS (w/o α_c)	OASIS (w. α_c)
0 - 29	86.4%	63.7	**69.1**	68.8
30 - 59	7.2%	47.4	52.4	**56.6**
60 - 89	3.5%	45.3	47.0	**51.5**
90 - 119	1.8%	29.3	36.2	**41.5**
120 - 149	1.0%	26.2	31.2	**39.7**
0-149 (all classes)	100%	42.4	47.2	**51.6**

Table 4.4: Per-class IoU scores on ADE20k, grouped by pixel-wise frequency (the fraction of all pixels in the datasets belonging to one class). Bold denotes the best performance. Training with per-class loss balancing is denoted by α_c.

Method	FID ↓	mAP, % ↑	classes with AP > 0 ↑
SPADE+	26.8	4.56	439
OASIS	**15.3**	**5.38**	**510**
real data	0	6.70	624

Table 4.5: Comparison of SPADE+ and OASIS on the LVIS dataset with 1203 classes and a long tail of underrepresented classes. Bold denotes the best performance. Last row shows the scores for the LVIS validation set.

4.5. In this more extremely imbalanced data regime, the gain of our model is pronounced: OASIS outperforms SPADE+ by a large margin, lowering the FID by 43% (from 26.8 to 15.3). Fig. 4.9 shows a qualitative comparison between the models. OASIS produces images of higher visual quality with more natural colors and textures. In Table 4.5 we report the mean Average Precision (mAP) of the instance segmentation network evaluated on the set of generated images. OASIS outperforms SPADE+ in mAP by a notable margin (5.38 vs 4.56), thus producing objects with a more realistic appearance and largely reducing the gap to real data (mAP of 6.70). To evaluate the ability of the models to generate underrepresented classes at the tail of the LVIS data distribution, we count the number of classes for which a non-zero AP score is achieved. Table 4.5 shows that OASIS can model more semantic classes: OASIS achieves a positive AP for 510 semantic classes compared to 439 for SPADE+, thus exhibiting a better capability to synthesize underrepresented classes.

In addition to better handling the class imbalance, OASIS also visually outperforms SPADE+ on the LVIS label maps with a very large proportion of the background class. As seen in Fig. 4.9 (four rightmost columns), from such label maps, SPADE+ fails to produce plausible images and suffers from mode collapse. In contrast, OASIS successfully deals with such kinds of inputs, producing diverse and visually plausible images even for the least annotated label maps, with the highest proportion of the background class.

In conclusion, we consider long-tailed datasets, such as LVIS, an interesting direction for future work, as the improved synthesis of multiple tail classes under severe imbalance can significantly boost the applicability of semantic image synthesis to real-world applications.

Figure 4.10: Images generated by OASIS on ADE20K with 256×256 resolution using different 3D noise inputs. For both input label maps, the noise is re-sampled globally (first row) or locally in the areas marked in red (second row).

4.3.4 Image Editing with OASIS

OASIS can generate many different-looking images for a single label map by directly resampling input 3D noise. In the following, we present qualitative multi-modal results and discuss two unique semantic image editing techniques enabled by our model: local resampling of selected semantic classes and diverse resampling of unlabelled images.

Global and local resampling of the 3D noise. The 3D noise of OASIS modulates the activations directly at every generator layer, matching the spatial resolution of features at different generation scales. Therefore, such modulation affects both global and local characteristics of a generated image. At test time, this allows different strategies for noise sampling. For example, the noise can be sampled globally for all pixels, varying the whole image (see Fig. 4.10, first and third rows). Alternatively, a noise vector can be re-sampled only for specified image regions, resulting in local image editing while preserving the rest of the scene. For example, the local strategy allows to re-sample only the sky area in a landscape scenery, or only the window in a scene of a bedroom (see Fig. 4.10, second and fourth rows). Spatial sensitivity of OASIS to 3D noise is further demonstrated in Fig. 4.11, showing interpolations in the latent space. The learned latent space captures well the semantic meaning of objects and allows smooth interpolations not only globally, but also locally for selected objects (see Fig. 4.11, two last rows).

Creating diverse images from unlabelled data. In contrast to previous semantic image synthesis methods, the OASIS discriminator can be reused as a stand-alone image segmenter. To obtain a segmentation prediction for a given image, a user just needs to feed it to our pre-trained discriminator and select the highest activation among real classes in its $(N{+}1)$-channel output for each pixel. When tested as an image segmenter on the validation set of ADE20K, the OASIS discriminator reaches a mIoU of 40.0. For comparison, the state-of-the-art model DeepLab-V3 with a ResNeST backbone (*Zhang et al.*, 2020b) achieves an mIoU of 46.91. The good

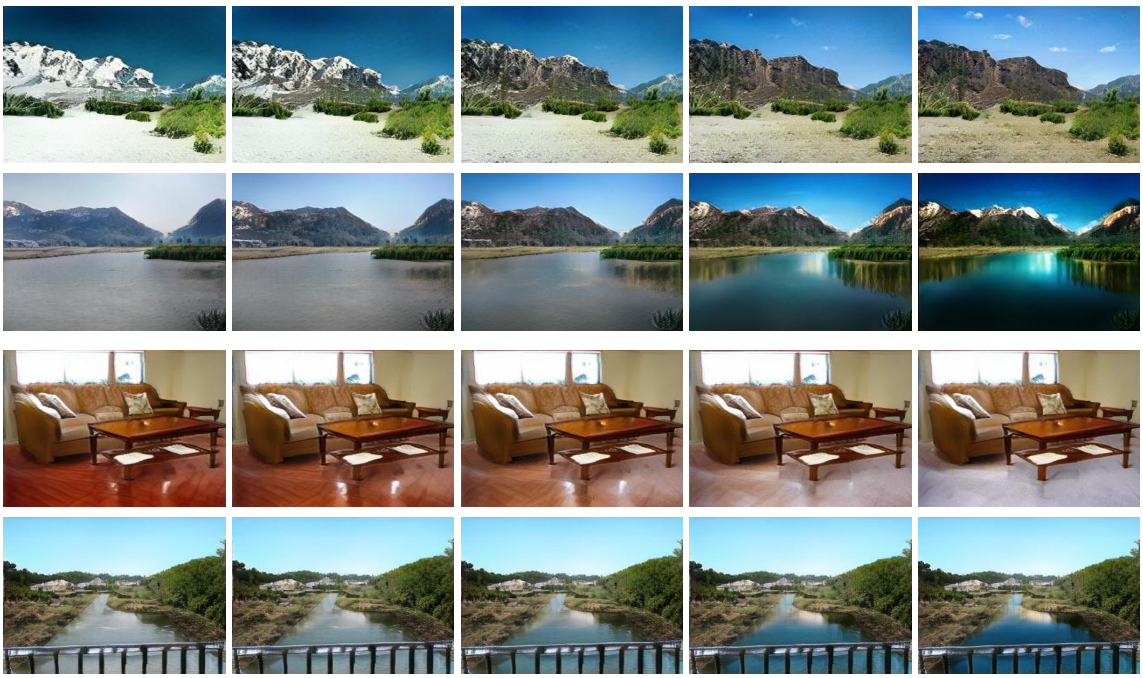

Figure 4.11: Latent space interpolations between images generated by OASIS for the ADE20K dataset at resolution 256×256. The first two rows display *global* interpolations. The second two rows show *local* interpolations of the floor or water only.

segmentation performance allows OASIS to be applied to unlabelled images: given an unseen image without the ground truth annotation, OASIS can predict a label map via the discriminator. Subsequently feeding this prediction to the generator allows to synthesize a scene with the same layout but different style (see Fig. 4.12). The recreated scenes closely follow the ground truth label map of the original image and vary considerably, due to the high sensitivity of OASIS to the 3D noise. We note that OASIS uniquely reaches this ability using only adversarial training, without the need for an external segmentation network or additional loss functions. We believe that the ability to create multiple versions of one image while retaining the layout, but not requiring the ground truth label map, may provide useful data augmentation for various applications in future research.

4.3.5 Synthetic Data Augmentation

As an additional evaluation method, we test the efficacy of generated images when applied as synthetic data augmentation for the task of semantic segmentation. Synthetic data augmentation is a task that benefits from both image quality and diversity, as well as the ability to generate semantic classes that are underrepresented in the original data (see Table 4.4). Therefore, the effect of synthetic data augmentation on downstream performance can constitute a more holistic evaluation of semantic image synthesis models. To test the efficiency of OASIS, we train a DeepLab-V3 segmentation network on ADE20K and Cityscapes, at each step augmenting each training image with its synthetic augmentation, produced by OASIS from the same label map.

We compare OASIS against the strong baseline SPADE in Table 4.6. Between the two methods, OASIS elicits a stronger increase in segmentation performance with an improvement of 2.0 mIoU on Cityscapes and 0.8 mIoU on ADE20K, compared to DeepLab-V3 trained without synthetic augmentation. The higher performance improvement of OASIS compared to SPADE is explained by all the previously observed gains in image quality, diversity, and the alignment to input label maps (see Fig. 4.7, Tables 4.1 and 4.2). In addition to that, the segmentation performance is also improved due to the fact that OASIS tends to synthesize underrepresented classes better than SPADE, which is evident from Table 4.7. This table compares the IoU

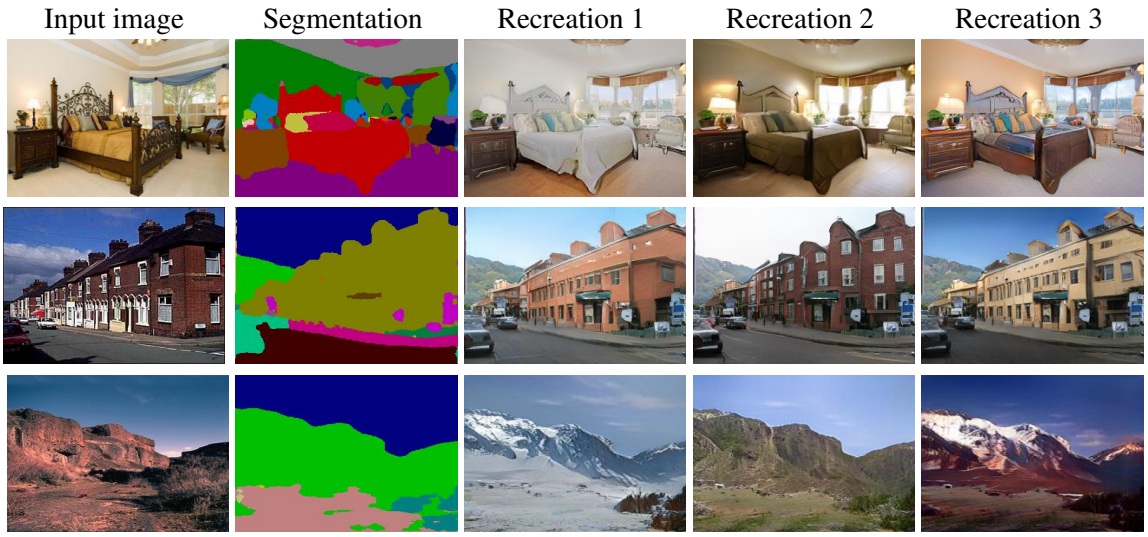

Figure 4.12: After training, the OASIS discriminator can be used to segment images. The first two columns show the real image and the segmentation of the discriminator. Using the predicted label map, the generator can produce multiple versions of the original image by resampling noise (Recreations 1-3). Note that no ground truth maps are required.

Data augmentation	Cityscapes mIoU↑	ADE20K mIoU↑
no synthetic DA	62.7	41.0
with SPADE	62.6	41.6
with OASIS	**64.7**	**41.8**

Table 4.6: Semantic segmentation performance of ResNeSt-50 with and without synthetic data augmentation (DA). Bold denotes the best performance.

performance of DeepLab-V3 on the well represented and underrepresented classes of Cityscapes, as measured by the pixel-wise frequency of the semantic class in the dataset. Examples of well represented classes are road and building (see the 1st row of Table 4.7), while classes like bicycle or traffic light are the least represented in the dataset (see 4th row in Table 4.7). Note that the IoU comparison in Table 4.7 is different from Table 4.4, where the IoU was measured directly on synthetic data using a pretrained segmenter. It can be seen that the improvement in IoU through OASIS can be mostly attributed to better performance on underrepresented classes, as the gap in performance between OASIS and SPADE becomes larger for the classes which are less represented. Lastly, since the OASIS generator was trained to fool an image segmenter (the OASIS discriminator), it may synthesize harder examples for semantic segmentation than SPADE, thus having higher potential to improve the generalization of segmentation networks to challenging corner cases. We find the above results promising for future utilization of OASIS in various downstream applications. Moreover, for future research, we find it interesting to explore synthetic data augmentation in combination with other data augmentation techniques, e.g., RandAugment (*Cubuk et al.*, 2020), which has the potential to provide further performance gains for downstream applications.

Sorted classes	Pixel-wise frequency	None	SPADE		OASIS	
			abs	rel	abs	rel
0 - 4	82.7%	90.6	90.6	+0.0	**90.9**	**+0.3**
5 - 8	12.5%	66.2	66.2	+0.0	**67.4**	**+1.2**
9 - 12	3.3%	50.2	49.1	-1.1	**52.2**	**+2.0**
13 - 18	1.6%	51.9	52.3	+0.4	**55.4**	**+3.5**
all classes	100%	62.7	62.6	-0.1	**64.7**	**+2.0**

Table 4.7: Per-class IoU scores on Cityscapes, obtained without (None) and with synthetic data augmentation using SPADE or OASIS. The classes are sorted and grouped by class pixel-wise frequency, as measured by the total fraction of pixels in the dataset belonging to one class. Bold denotes the best performance. The absolute (abs) and relative (rel) mIoU gain via data augmentation is shown.

G	D	VGG	LabelMix	FID↓	mIoU↑
SPADE+	SPADE+	✗	✗	60.7	21.0
SPADE+	OASIS	✗	✗	29.0	**52.1**
OASIS	OASIS	✗	✗	29.3	51.6
		✗	✓	28.4	50.6
OASIS	OASIS	✗	✓	**28.3**	48.8
+3D noise		✓	✓	31.6	50.8

Table 4.8: Main ablation on ADE20K. The OASIS generator is a lighter version of the SPADE+ generator (72M vs 96M parameters). Bold denotes the best performance.

4.3.6 Ablations

We conduct all our ablations on the ADE20K dataset. We choose this dataset as it more challenging (with 150 classes) than Cityscapes (35 classes) and ADE20K-Outdoors (110 classes), and has more reasonable training time (5 days) compared to COCO-Stuff and LVIS (4 weeks). Our main ablation shows the impact of the main technical components of OASIS, including the new discriminator, lighter generator, LabelMix and the 3D noise. Further ablations are concerned with the architecture changes in the discriminator, the label map encoding in the discriminator, different noise sampling strategies, LabelMix and the GAN feature matching loss.

Main ablation. Table 4.8 shows that SPADE+ achieves low performance on the image quality metrics without the perceptual loss. Replacing the SPADE+ discriminator with the OASIS discriminator, while keeping the generator fixed, improves FID and mIoU by more than 30 points. Changing the SPADE+ generator to the lighter OASIS generator leads to a negligible degradation of 0.3 in FID and 0.5 in mIoU, but reduces the number of parameters from 96M to 72M. With LabelMix FID improves further by about 1 point. Adding 3D noise improves FID but degrades mIoU, as diversity complicates the task of the pre-trained semantic segmentation network used to compute the mIoU score. For OASIS the perceptual loss deteriorates FID by more than 2 points, but improves mIoU. Overall, without the VGG loss the new discriminator is the key to the performance boost over SPADE+.

Ablation on the discriminator architecture. We train the OASIS generator with three alternative discriminators: the original multi-scale PatchGAN consisting of two networks, a single-scale PatchGAN, and a ResNet-based discriminator, corresponding to the encoder of the U-Net shaped OASIS discriminator. Table 4.9 shows

D architecture	w/o VGG		with VGG	
	FID↓	mIoU↑	FID↓	mIoU↑
MS-PatchGAN (2x)	60.7	21.0	32.9	42.5
PatchGAN	197	0.62	34.2	42.2
ResNet-PatchGAN	147	0.42	32.4	45.1
OASIS	**29.3**	**51.6**	**29.2**	**51.1**

Table 4.9: Ablation on D architecture. Bold denotes the best performance, red shows collapsed runs.

Label encoding	w/o VGG		with VGG	
	FID↓	mIoU↑	FID↓	mIoU↑
Input concatenation	280	0.02	30.0	43.9
Projection	32.4	44.9	**28.0**	46.9
N+1 loss	**28.3**	47.2	28.6	49.8
Balanced N+1 loss	29.3	**51.6**	29.2	**51.1**

Table 4.10: Ablation on the label map encoding. Bold denotes the best performance, red shows collapsed runs.

that the alternative discriminators only perform well with perceptual supervision, while the OASIS discriminator achieves superior performance independent of it. The single-scale discriminators even collapse without the perceptual loss (red colors in Table 4.9).

Ablation on the discriminator label map encoding. We study four different ways to use label maps in the discriminator: the first encoding is input concatenation, as in SPADE. The second option is a pixel-wise projection-based GAN loss (*Miyato and Koyama*, 2018). Unlike *Miyato and Koyama* (2018), we condition the GAN loss on the label map instead of a single label. The third and fourth option is to employ the label maps as ground truth for the N+1 segmentation loss, or for the class-balanced N+1 loss (see Sec. 4.2.2). For a fair comparison we use neither 3D noise nor LabelMix. As shown in Table 4.10, input concatenation is not sufficient without additional perceptual loss supervision, leading to training collapse. Without the perceptual loss, the N+1 loss outperforms the input concatenation and the projection in both the FID and mIoU metrics. Finally, the class balancing enables enhanced supervision for underrepresented semantic classes, which noticeably improves mIoU scores. On the other hand, we observed that the FID metric is more sensitive to the synthesis of well represented classes and not underrepresented classes, which explains the negative effect of the class balancing on FID.

Ablation on LabelMix. Consistency regularization for the segmentation output of the discriminator requires a method of generating binary masks. Therefore, we compare the effectiveness of CutMix (*Yun et al.*, 2019) and our proposed LabelMix. Both methods produce binary masks, but only LabelMix respects the boundaries between semantic classes in the label map. Table 4.11 compares the FID and mIoU scores of OASIS trained with both methods on the Cityscapes dataset. As seen from the table, LabelMix improves both FID (51.5 vs. 47.7) and mIoU (66.3 vs. 69.3), in comparison to OASIS without consistency regularization. CutMix-based consistency regularization only improves the mIoU (66.3 vs. 67.4), but not as much as LabelMix (69.3). We suspect that since the images are already partitioned through the label map, an additional partition through CutMix results in a dense patchwork of areas that differ by semantic class and real/fake class identity. This may introduce additional label noise during training for the discriminator. To avoid such inconsistency between semantic classes and real/fake identity, the mask of LabelMix is generated according to the label map, providing natural borders between semantic regions, so that the real and fake objects are placed side-by-side

Transformation	FID↓	mIoU ↑
No CR	51.5	66.3
CutMix	52.1	67.4
LabelMix	**47.7**	**69.3**

Table 4.11: Ablation study on the impact of LabelMix and CutMix for consistency regularization (CR) in OASIS on Cityscapes. Bold denotes the best performance.

without interfering with each other. Under LabelMix regularization, the generator is encouraged to respect the natural semantic class boundaries, improving pixel-level realism while also considering the class segment shapes.

4.4 Conclusion

In this chapter we studied semantic image synthesis, the task of generating diverse and photorealistic images from semantic label maps. Conventionally, semantic image synthesis GAN models employed a perceptual VGG loss to overcome training instabilities and improve the synthesis quality. In our experiments we demonstrated that the VGG-based perceptual loss imposes unnecessary constraints on the feature space of the generator, significantly limiting its ability to produce diverse samples from input noise, as well as the ability to produce images with colors and textures closely matching the distribution of real images. Therefore, in this work we propose OASIS, a GAN model for semantic image synthesis that needs only adversarial supervision to achieve high-quality results.

The improvement over the prior work in image synthesis quality is achieved via the detailed spatial and semantic-aware supervision from our novel segmentation-based discriminator, which uses semantic label maps as ground truth for training. With this powerful discriminator, OASIS can easily generate diverse outputs from the same semantic label map by resampling 3D noise, eliminating the need for additional image encoders to achieve multi-modality. The proposed 3D noise injection scheme can work both in a global and local regime, allowing to change the appearance of the whole scene and of individual objects. With the proposed modifications, OASIS significantly improves over previous state-of-the-art GAN models in terms of image synthesis quality.

Furthermore, we proposed to use the LVIS dataset to evaluate semantic image synthesis under severe class imbalance and sparse label annotations. Thanks to the class balancing mechanism enabled by its segmentation-based discriminator, OASIS achieves more realistic synthesis of underrepresented classes, achieving pronounced gains on the extremely unbalanced LVIS dataset. Lastly, the design of OASIS can be better suited for image editing applications compared to the SPADE baseline, enabling diverse resampling of scenes from unlabelled images, as well as for synthetic data augmentation, improving the performance of a segmentation network by a larger margin.

While semantic image synthesis is an interesting task that has a lot of applications, it comes with some practical restrictions. Firstly, achieving good performance in this task necessitates the availability of large training datasets. For example, the size of datasets used for the training of OASIS in Table 4.1 ranges from 3000 images (Cityscapes) to over 100000 images (COCO-stuff, LVIS). This requirement naturally limits the utilization of OASIS in many restricted image domains, where finding such datasets is challenging. Secondly, semantic image synthesis datasets require semantic label maps for each image. Annotating pixel-wise segmentation masks of objects is widely recognized as an expensive and time-consuming process, thereby making the curation of large datasets costly. Therefore, in the upcoming chapters, our focus will shift towards exploring unconditional GANs in scenarios with limited data, with the aim of expanding the application of GANs to new practical domains and applications.

CHAPTER 5

Generating Novel Scene Compositions from Single Images and Videos

In Chapter 4, we showed that GANs can achieve impressive performance with sufficient training data and rich conditioning. However, in many scenarios it is interesting to train GANs on much smaller image datasets and without annotations. Therefore, in this chapter, we explore GAN training on extremely limited unconditional datasets. Previous research in this direction focused on two types of models: training GANs on a single image and on few-shot datasets (at least 100 diverse images). We find a significant gap between these models: the former struggles to learn even from two images, while the latter suffers from overfitting and memorization when the number and diversity of training images are reduced from standard few-shot datasets. To address these limitations, we introduce a new GAN model, called SIV-GAN. Our model generates new diverse images given a small set of very similar images, such as frames from a single video. SIV-GAN consists of two main components: a two-branch discriminator to mitigate overfitting and a diversity regularization technique for diverse synthesis in challenging data scenarios. Compared to previous single-image and few-shot GAN methods, as well as existing image manipulation techniques, our model demonstrates higher efficiency in producing diverse and realistic scene compositions from extremely limited data.

Individual Contribution

This chapter is based on the following conference workshop publication (*Sushko et al.*, 2021a):

One-Shot GAN: Learning to Generate Samples from Single Images and Videos
Vadim Sushko, Juergen Gall, Anna Khoreva
IEEE Computer Vision and Pattern Recognition Conference (CVPR) workshops, 2021.
DOI: 10.1109/CVPRW53098.2021.00293

This work was done in very close collaboration between Vadim Sushko and Anna Khoreva. The initial idea to explore the GAN training on frames of a single video was proposed by Anna Khoreva. Subsequently, the joint discussions shaped the architecture of the SIV-GAN model, and the model was later found to generalize in the single image setting. Throughout the entire development process, Juergen Gall provided invaluable scientific guidance, feedback, and suggestions. Additionally, Dan Zhang joined the project after the initial submission and provided guidance and suggestions for the further exploration of the experimental settings and paper writing. In this paper, Vadim Sushko is the first author who made contributions to all stages of the project, including discussions, implementation, evaluations, and paper writing.

Contents

5.1 Introduction

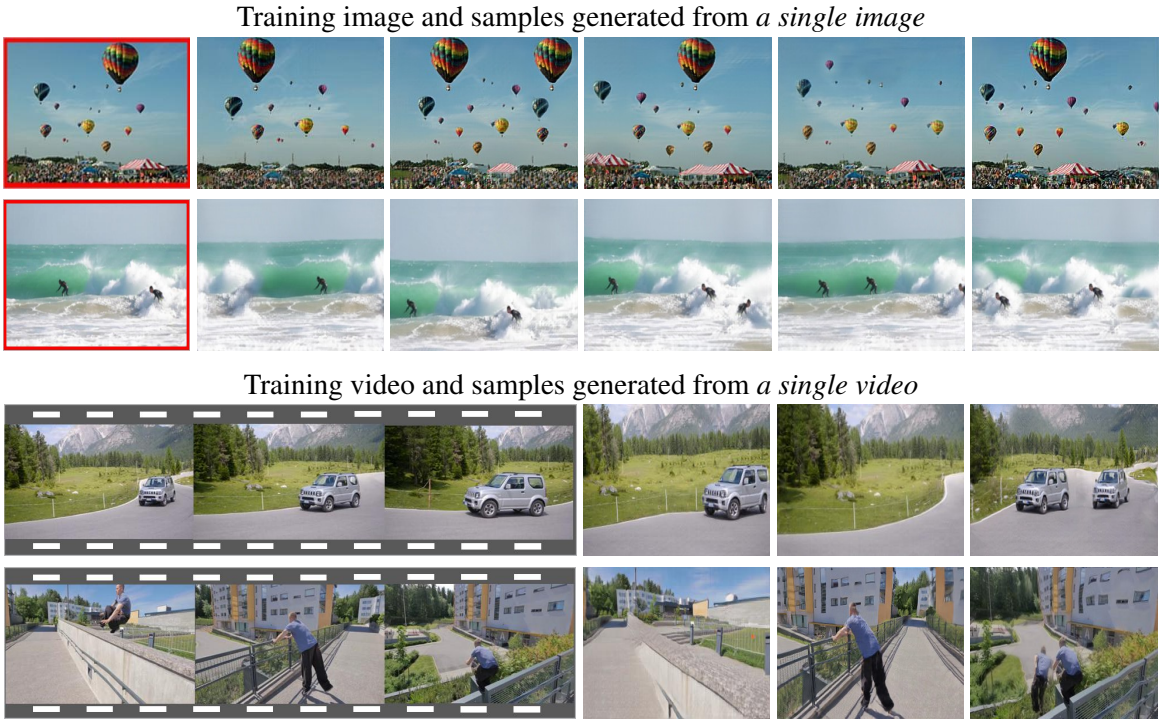

Figure 5.1: Images generated by SIV-GAN. Our model successfully operates in extremely low data regimes, generating new scene compositions with varying content and layout from a single image (first two rows) or a single video (last two rows). For example, from the single training surfing image, it can synthesize layouts with a different position and configuration of waves and change the number of surfers; and from a single video with a car on the road, SIV-GAN generates images without a car or with two cars. Original training samples are shown in red or grey frames.

The quality of synthetic images produced by generative adversarial networks (GANs) has greatly improved in recent years (*Brock et al.*, 2019; *Schönfeld et al.*, 2020; *Karras et al.*, 2021, 2020b). These impressive results are in large part enabled by the availability of large, diverse datasets, typically consisting of tens of thousands of images. This dependency on the availability of training data limits the applicability of GANs in domains where collecting a large dataset is not feasible. In some real-world applications, collection of even a small dataset remains challenging due to specific constraints related to privacy, copyright status, subject type, geographical

location, time, and dangerous or hazardous environments. It may happen that rare objects or events are present only in one image or video, and it is difficult to obtain a second one. For example, this includes images of rare animal species or videos of traffic accidents recorded in extreme conditions. Thus, enabling GANs to generate diverse and high-quality images in extremely low data regimes can improve their utilization in practice.

In this context, prior work mainly focused on developing GAN models for two low data regimes: learning from a single image (*Shocher et al.*, 2019; *Shaham et al.*, 2019; *Hinz et al.*, 2021) (Fig. 5.1, first two rows), or from few-shot datasets, which usually consist of 100 diverse images or more (*Liu et al.*, 2021). As our experiments revealed, the latter methods are still strictly limited by the number of images that are available for training, or their overall diversity. For example, we observed a severe performance degradation of the few-shot model FastGAN (*Liu et al.*, 2021) when trained on few-shot datasets consisting of *very similar* images. To explore the limitations of existing models further, in this work we introduce a new task of generating images from a *single video* (Fig. 5.1, last two rows), where the training data is a collection of 60-100 frames taken from a short (2-10 seconds) video clip. Similarly to *Liu et al.* (2021), in this task we aim to generate diverse images, but not temporally-coherent videos. The introduced single video setting is interesting for training unconditional GANs for two reasons. Firstly, capturing a short video is easy in practice, which can facilitate the data collection needed for successful GAN training in restricted image domains. Secondly, compared to a few-shot dataset containing 60-100 images (e.g., as used in (*Liu et al.*, 2021)), the overall diversity of frames taken from the same video is much smaller due to a high correlation between adjacent video frames. As shown in our experiments, both single-image and few-shot GAN models cannot successfully deal with such a data regime, which makes it an interesting evaluation benchmark that can boost the applications of GANs to new image domains.

The challenge of training GANs from only a single image, as well as from only a few highly-correlated video frames, is the problem of overfitting (*Karras et al.*, 2020a). For example, applying few-shot image synthesis models, such as FastGAN (*Liu et al.*, 2021), to learn from a single image or video leads to severe memorization problems (see Sec. 5.3). As one approach to mitigate the memorization issues, single image GAN models (*Shaham et al.*, 2019; *Hinz et al.*, 2021) proposed to learn an internal patch-based distribution of an image, employing a cascade of multi-scale patch-GANs (*Isola et al.*, 2017) trained in multiple stages. Though these models overcome memorization, producing different versions of a training image, they cannot learn high-level semantic properties of the scene, e.g., to dissect objects from the background. Consequently, they often suffer from incoherent shuffling of image patches, distorting objects and producing unrealistic layouts (see Fig. 5.4). Furthermore, the success of these models is inherently limited to learning from a single image. As we demonstrate in our experiments (see Sec. 5.3), these models struggle to generate realistic scenes when trained on multiple images, e.g., on frames collected from a single video.

In this work, we aim to go beyond patch-based learning and to achieve image synthesis of high diversity and quality at the same time. Given only a single image or video, we aim to learn a model which is able to compose new layouts, re-arrange objects in the scene, remove or duplicate instances, and change their shape and size. The generated scene compositions should be visually plausible, with objects preserving their appearance and natural shape, and scene layouts looking realistic to the human eye.

Although there exists an alternative way to compose new images by using image manipulation methods, such as image inpainting (*Dong et al.*, 2022) or blending of objects (*Zhang et al.*, 2020c), training an unconditional GAN model on limited data has several advantages. In particular, image manipulation methods need to be pre-trained, and they typically require a lot of data to achieve high performance. Moreover, for each new image, they require pixel-level user input to indicate which objects should be edited, which poses a limitation on a number of images these methods can produce. The above properties limit the utilization of image manipulation methods in practical applications dealing with restricted data domains, for example one-shot or few-shot image classification (*Tian et al.*, 2020a). As in such scenarios classification models are required to learn to recognize a new class (e.g., previously unseen dog breed) by using only one or a few examples, they are severely prone to overfitting. Therefore, augmenting the limited available training data with diverse novel scene compositions, for example by changing the shape or location of objects in the images, has the potential to reduce overfitting of classification models and thereby improve their performance.

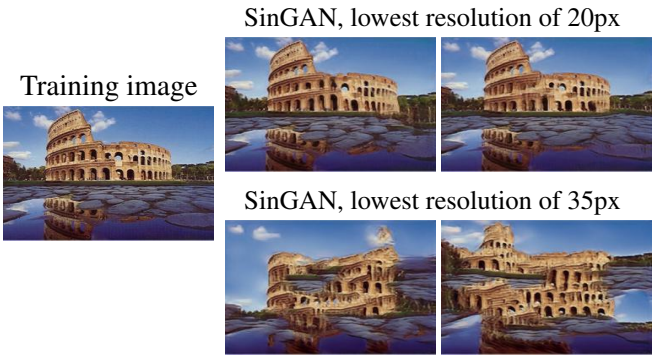

Figure 5.2: Limitation of the multi-stage training of the single image SinGAN model (*Shaham et al.*, 2019). As the finer generation stages cannot correct the layout decision made by the coarser scale generators, without a careful tuning of the lowest resolution size the model produces images of very low diversity (first row) or lacking global coherency (last row).

To this end, we introduce SIV-GAN (**S**ingle **I**mage and **V**ideo GAN), an unconditional, one-stage GAN model, capable of learning from the single data instances to generate images that are substantially different from the original training sample, while still preserving its context. This is achieved by two key ingredients: the novel discriminator design and the proposed diversity regularization for the generator. Our discriminator has two branches, separately judging image content and scene layout realism. The content branch evaluates objects' fidelity irrespective of their spatial arrangement, while the layout branch looks only at the global scene coherency. Disentangling the discriminator's decision about content and layout helps to prevent overfitting and provides a more informative signal to the generator. To achieve a high diversity among generated samples, we further extend the regularization technique of *Yang et al.* (2019) to unconditional image synthesis in single data instance regimes. The prior work of *Yang et al.* (2019) encourages the generation of different images depending on their input latent codes, thus the difference between images is proportional to the distance between their codes in the latent space. Assuming that in case of a single image or a single video all generated images should belong to one semantic domain (i.e. preserve the original training sample context), and thus should be more or less equally different from each other, we apply diversity regularization uniformly, independent of the latent space distance. Moreover, we use the regularization in the feature space, inducing both high- and low-level diversity.

We demonstrate the effectiveness of our model for the single image setting, as well as for the novel single video setting in Sec. 5.3. SIV-GAN is the first model that successfully learns in both of these extremely low-data settings, improving over prior work (*Shaham et al.*, 2019; *Hinz et al.*, 2021; *Liu et al.*, 2021) in both image quality and diversity. As shown in our experiments, in contrast to FastGAN (*Liu et al.*, 2021), our model does not suffer from memorization, successfully dealing with a high similarity between training images, while compared to single image GANs (*Shaham et al.*, 2019; *Hinz et al.*, 2021), our model does not distort objects and preserves their appearance. In summary, our main contributions are: i) We propose a new task of learning generative models from frames of a single video, which introduces a new challenge for few-shot image synthesis models due to a high similarity between training images. ii) We present a novel two-branch discriminator, encouraging the generation of new scene compositions with layouts and content substantially different from training samples. Our proposed diversity regularization ensures a high variability among generated samples in the challenging single data instance regimes. iii) With SIV-GAN, we achieve high quality and diversity when learning from both single images and videos, outperforming prior GAN models and image manipulation methods.

5.2 Method

In this section, we present SIV-GAN, an unconditional GAN model that learns from a single image or a single video to generate new plausible compositions of a given scene with varying content and layout. The key ingredients of SIV-GAN are a novel design of a two-branch discriminator (Sec. 5.2.1) and a diversity regularization introduced for synthesis in single data instance regimes (Sec. 5.2.2).

5.2.1 Content-Layout Discriminator

One challenge of training GANs in single data instance regimes is the problem of overfitting to original samples. In many cases the model can simply memorize the original training images and their augmented versions (if used during the training). To avoid this memorization effect, *Shaham et al.* (2019) and *Hinz et al.* (2021) proposed to learn an internal patch-based distribution of a single image by using a hierarchy of patch-GANs (*Isola et al.*, 2017) at different image scales. As the employed patch-GANs have small receptive fields and limited capacity, they are prevented from memorizing the full image. However, the downside of training each scale of the patch-GANs in a separate stage is that any layout decisions made by the coarser scale generators cannot be corrected at later, finer generation stages. Thus, both the quality and diversity of generated images are highly dependent on the chosen lowest resolution size. This parameter needs careful tuning for specific images at hand, otherwise image layouts may lack diversity or loose global coherency (see Fig. 5.2). Moreover, this approach does not generalize to learning from multiple images, as in the single video case (see Fig. 5.5).

We therefore introduce an alternative solution to overcome the memorization effect but still to produce high-quality images. We note that in order to produce realistic and diverse images, the generator should learn the appearance of objects and combine them in the image in a globally-coherent way. To this end, we propose a discriminator that judges the *content* distribution of a given image separately from its *layout* realism. To achieve the disentanglement, we design a two-branch discriminator architecture, with separate content and layout branches. Note that the branching of the discriminator happens after intermediate layers; this is done in order to learn relevant representations for building the branches. As seen from Fig. 5.3, our discriminator consists of the low-level feature extractor $D_{low\text{-}level}$, the content branch $D_{content}$, and the layout branch D_{layout}. For a given image x, the purpose of $D_{low\text{-}level}$ is to learn low-level features and to produce an image representation $F(x) = D_{low\text{-}level}(x)$ for the branches. Next, $D_{content}$ will judge the content of $F(x)$, irrespective from its spatial layout, while on the other hand D_{layout} will inspect only the spatial information extracted from $F(x)$. Inspired by the attention modules of *Park et al.* (2018); *Woo et al.* (2018), we implement the content-layout disentanglement by squeezing channels or spatial dimensions of the intermediate features $F(x)$. Note that afterwards the branches $D_{content}$ and D_{layout} receive only limited information about the image from $F(x)$, preventing them from overfitting to the whole image, and thus mitigating the negative effect of memorizing the original image.

Content branch. The content branch decision should be based upon the image content, i.e. the fidelity of objects composing the image, independent of their spatial location in the scene. Let the feature map $F(x)$ have dimensions $H(\text{height}) \times W(\text{width}) \times C(\text{channel})$. Note that the spatial dimensions $H \times W$ capture spatial information, while the channels C encode the semantic representation. As we want the content branch to ignore the spatial location of objects, we apply global average pooling to aggregate the spatial information $H \times W$ across the channels C. The resulting feature map $F_{content}(x)$ has size $1 \times 1 \times C$, which is then processed by several layers for further real/fake decision making. By removing the spatial information, $D_{content}(x)$ is induced to respond to content features encoded in different channels regardless of their spatial location (see Fig. 5.7).

Layout branch. The layout branch, in contrast, should assess the spatial location of objects in the scene, but not their specific appearance. Thus, the layout branch is designed to judge only the spatial information of $F(x)$, filtering out the content details. Since the layout information is encoded only in spatial dimensions $H \times W$, and not in channels C, we aggregate the channel information from $F(x)$ via a (1×1) convolution

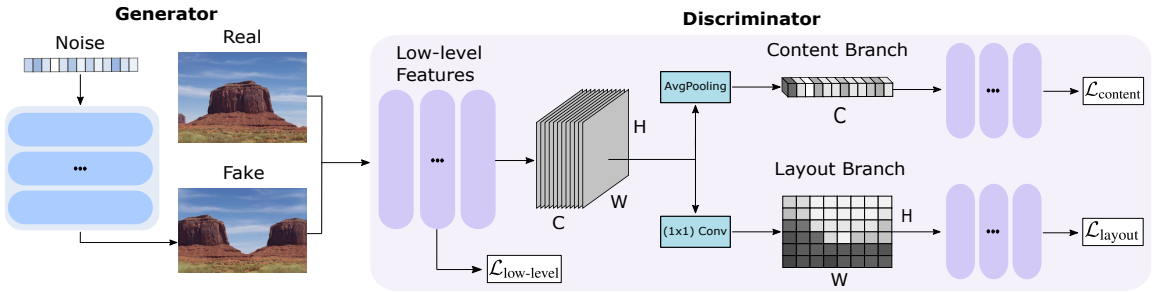

Figure 5.3: SIV-GAN architecture. Two separate discriminator branches judge the image content separately from the scene layout realism and thus enable the generator to produce images with varying content and global layouts. Before the branching, a low-level feature extractor is trained via a separate low-level loss, enabling to learn the low-level image realism and to build relevant representations for the content and layout branches.

with only one output channel, which forms a feature map $F_{layout}(x)$ with size $H \times W \times 1$. This channel aggregation weakens the content representation but does not affect the spatial information. The $F_{layout}(x)$ features are further processed by several layers before a real/fake decision is made. As $D_{layout}(x)$ is designed to be sensitive only to the spatial representation of the input image, it learns to judge the realism of scene layouts (see Fig. 5.7).

Feature augmentation. The proposed two-branch discriminator prevents the memorization of training samples, enabling the generation of images with content and layouts different from the original sample. To further improve the diversity of generated images, we propose to augment the content $F_{content}(x)$ and layout $F_{layout}(x)$ features of real images. For the single image setting this is done by mixing the features of two different augmentations of the original image, and for the single video setting by mixing the features of augmentations of two different video frames. For two real samples x_1 and x_2 we apply a mixing transformation $F_*(x_1) = T_{mix}(F_*(x_1), F_*(x_2))$. We use two types of mixing: 1) For the layout branch, we sample a rectangular crop of $F_{layout}(x_2)$ and paste it on to $F_{layout}(x_1)$ at the same spatial location, similarly to CutMix (*Yun et al.*, 2019). In contrast to CutMix, our approach augments features, not input images, and mixes only features of real images. 2) For the content branch, we sample a set of channels from $F_{content}(x_2)$ and copy their values to the corresponding channels of $F_{content}(x_1)$. As the channels encode semantic features of images, we expect the resulting augmented tensor to represent objects seen in two different images. We also found it useful to remove channels from $F_{content}(x)$, thus removing some object representations. For this, we sample a set of channels and drop out their values (*Srivastava et al.*, 2014; *Zhengsu et al.*, 2018). With the above augmentations, $D_{content}$ and D_{layout} see significantly more variance in both the content and layout representations of real images, which prevents overfitting and improves the diversity of generated samples. The effect of feature augmentation (FA) is shown in Table 5.3.

Adversarial loss. To evaluate images at different scales, we design our discriminator to make a binary true/fake decision at each intermediate resolution. For each discriminator part D_*: $D_{low\text{-}level}$, $D_{content}$, and D_{layout}, the loss is computed by aggregating the contributions across all layers constituting the corresponding discriminator part:

$$\mathcal{L}_{D_*} = \frac{1}{N_*} \sum_{l=1}^{N_*} \mathcal{L}_{D_*^l}, \tag{5.1}$$

where D_*^l is the l-th ResNet block of D_*, N_* is the number of ResNet blocks used in D_*, and the loss $\mathcal{L}_{D_*^l}$ is the binary cross-entropy: $\mathcal{L}_{D_*^l} = -\mathbb{E}_x[\log D_*^l(x)] - \mathbb{E}_z[\log(1 - D_*^l(G(z)))]$. D_*^l aims to distinguish between real x and generated $G(z)$ images based on their corresponding features at block l, which captures either their low-level details, content, or layout at a certain resolution. The overall adversarial loss for SIV-

GAN is then computed by taking the decisions from the content branch $D_{content}$, the layout branch D_{layout}, and the low-level features of $D_{low\text{-}level}$:

$$\mathcal{L}_{adv}(G, D) = \mathcal{L}_{D_{content}} + \mathcal{L}_{D_{layout}} + 2\mathcal{L}_{D_{low\text{-}level}}, \tag{5.2}$$

where D aims to distinguish between real and generated images based on their low-level, content, and layout realism. As two D branches operate on high-level image features, contrary to only one $D_{low\text{-}level}$ operating on low-level features, we double the weighting for $\mathcal{L}_{D_{low\text{-}level}}$. This is done in order to properly balance the contributions of different feature scales and encourage the generation of images with good low-level details, plausible content, and coherent scene layouts. The effect of $\mathcal{L}_{D_{low\text{-}level}}$ is discussed in Sec. 5.3.3.

5.2.2 Diversity Regularization

To improve the variability among the generated images, we propose to add a diversity regularization (DR) loss term $\mathcal{L}_{DR}$ to the SIV-GAN objective. Prior work (*Yang et al.*, 2019; *Zhao et al.*, 2021; *Choi et al.*, 2020) also proposed to use diversity regularization for GAN training, but mainly to avoid mode collapse, and assuming the availability of a large training set. The regularization of *Yang et al.* (2019) aimed to encourage the generator to produce different outputs depending on the input latent code, in such a way that the generated samples with closer latent codes should look more similar to each other, and vice versa. In contrast, our diversity regularization is tuned for synthesis from single data instance regimes. Assuming that in case of a single image or a single video we are operating in one semantic domain, the generator should produce images that are in-domain but more or less equally different from each other, and substantially different from the original training sample. Thus, in such regimes the difference of generated images should not be dependent on the distance between their latent codes, so we propose to encourage the generator to produce perceptually different image samples independent of their distance in the latent space. Mathematically, the new diversity regularization is expressed as:

$$\mathcal{L}_{DR}(G) = \mathbb{E}_{z_1, z_2}\left[\frac{1}{L}\sum_{l=1}^{L}\|G^l(z_1) - G^l(z_2)\|\right], \tag{5.3}$$

where $\|\cdot\|$ denotes the $L1$ norm, $G^l(z)$ indicates a feature extracted after l-th resolution block of the generator G given input z, and z_1, z_2 are randomly sampled latent codes in the batch, i.e. $z_1, z_2 \sim N(0, 1)$. By regularizing the generator to maximize Eq. (5.3), we force it to produce diverse outputs for different latent codes z. Note that, in contrast to *Yang et al.* (2019); *Zhao et al.* (2021); *Choi et al.* (2020), we compute the distance between samples in the feature space of the generator. Computing the distance in the feature space results in a more meaningful diversity among the generated images, as different layers of the generator capture different image semantics, inducing both high- and low-level diversity. Computing the distance in the image space, i.e. $\mathcal{L}_{DR}(G) = \|G(z_1) - G(z_2)\|$ as in (*Choi et al.*, 2020), leads to reduced image diversity in our experiments (see Table 5.4).

The overall SIV-GAN objective can be written as:

$$\max_{G}\min_{D} \; \mathcal{L}_{adv}(G, D) + \lambda\mathcal{L}_{DR}(G), \tag{5.4}$$

where λ controls the strength of the diversity regularization and $\mathcal{L}_{adv}$ is the adversarial loss in Eq. (5.2). The proposed diversity regularization is shown to be highly-effective for SIV-GAN, while prior regularizations (*Yang et al.*, 2019; *Zhao et al.*, 2021; *Choi et al.*, 2020) underperform in our experiments (see Table 5.4).

5.2.3 Implementation and Training

The overall architecture of SIV-GAN is shown in Fig. 5.3. In our implementation, the SIV-GAN generator employs ResNet blocks which are similar to BigGAN (*Brock et al.*, 2019). However, we do not use BatchNorm

or self-attention. As in MSG-GAN (*Karnewar and Wang*, 2020), we generate images at intermediate ResNet blocks of G, passing them to $D_{low\text{-}level}$ to facilitate the gradient flow from the discriminator. The latent vector z, of length 64, is sampled from $N(0, 1)$. It is by default broadcasted to the spatial dimensions of 3x5, which can be adjusted to closer fit the shape of a training sample. For diversity regularization, we use the $\tanh$ activation on the features from the final convolutions of the G blocks.

The SIV-GAN discriminator also uses ResNet blocks. We set $N_{low-level} = 3$, $N_{layout} = N_{content} = 4$, thus using 3 ResNet blocks before branching and 4 ResNet blocks for the content and layout branches (the ablation is presented in Sec. 5.3.3). To enable multi-scale gradients, we incorporate images at different scales using the ϕ_{lin_cat} strategy from *Karnewar and Wang* (2020). The proposed feature augmentation (FA) is applied with probability 0.4 at every discriminator forward pass. We also use differentiable image augmentation (DA) (*Karras et al.*, 2020a; *Zhao et al.*, 2020c,a), applying translation, cropping, rotation, and horizontal flipping for real and fake images with a probability of 0.7 at each forward pass. As in *Karras et al.* (2020a), we observe no signs of leaking augmentations in the generated samples.

In contrast to previous single image GANs (*Shaham et al.*, 2019; *Hinz et al.*, 2021), which employ a multi-stage training scheme, SIV-GAN is trained end-to-end in one stage, with the losses from Eq. (5.4), with $\lambda = 0.15$ for $\mathcal{L}_{DR}$ (see the ablation on λ in Sec. 5.3.3). We use spectral normalization (*Miyato et al.*, 2018) for both G and D, and do not use a reconstruction loss as in *Shaham et al.* (2019); *Hinz et al.* (2021), or any other stabilization techniques. SIV-GAN is trained using the ADAM optimizer with $(\beta_1, \beta_2) = (0.5, 0.999)$, a learning rate of 0.0002 for both G and D, and a batch size of 5 (using different augmentations of a single image or of video frames).

5.3 Experiments

5.3.1 Experimental Setup

We evaluate SIV-GAN by conducting experiments in two different settings: learning from a *single image* and from a *single video*. For both of them, we use the same model configuration as described in Sec. 5.2.3. We train our model for 100k iterations in the Single Image setting and for 300k iterations in the Single Video setting.

Datasets. Following SinGAN (*Shaham et al.*, 2019), we evaluate the Single Image setting on 50 images extracted from the Places dataset (*Zhou et al.*, 2017a). In addition to their protocol, we also select 15 videos from the DAVIS (*Perazzi et al.*, 2016) and YFCC100M (*Thomee et al.*, 2016) datasets. In the Single Video setting, we use all the frames as training images, while for the Single Image setup we use only one frame from the middle of each sequence. The chosen videos last for 2-10 seconds and consist of 60-100 frames.

Metrics. To assess the quality of generated images, we measure the mean single FID (SIFID) (*Shaham et al.*, 2019). Following the evaluation from *Shaham et al.* (2019); *Hinz et al.* (2021), in the Single Image setting we also report the *best* SIFID among the generated samples. The original SIFID formulation uses InceptionV3 features before the first pooling layer at $\frac{H \times W}{4}$ resolution. We observed that such metric captures only low-level image details, such as colors and textures, and not high-level semantic image properties, such as appearance of objects or global layouts. Therefore, to evaluate higher-level realism, we additionally use later features, obtained before the final classification layer (at $\frac{H \times W}{16}$ resolution). To evaluate the diversity of samples, we adopt the pixel diversity metric from *Shaham et al.* (2019). To measure perceptual diversity, we also report the average LPIPS (*Dosovitskiy and Brox*, 2016) across pairs of generated images. To verify that the models do not simply reproduce the training set, we report average LPIPS to the nearest image in the training set, augmented in the same way as during training (Dist. to train). We note that SIFID tends to penalize diversity, favouring overfitting (*Robb et al.*, 2021). Thus, to account for this quality-diversity trade-off, a fair analysis should assess both diversity and quality.

Comparison models. We compare our model with single image methods, SinGAN (*Shaham et al.*, 2019) and

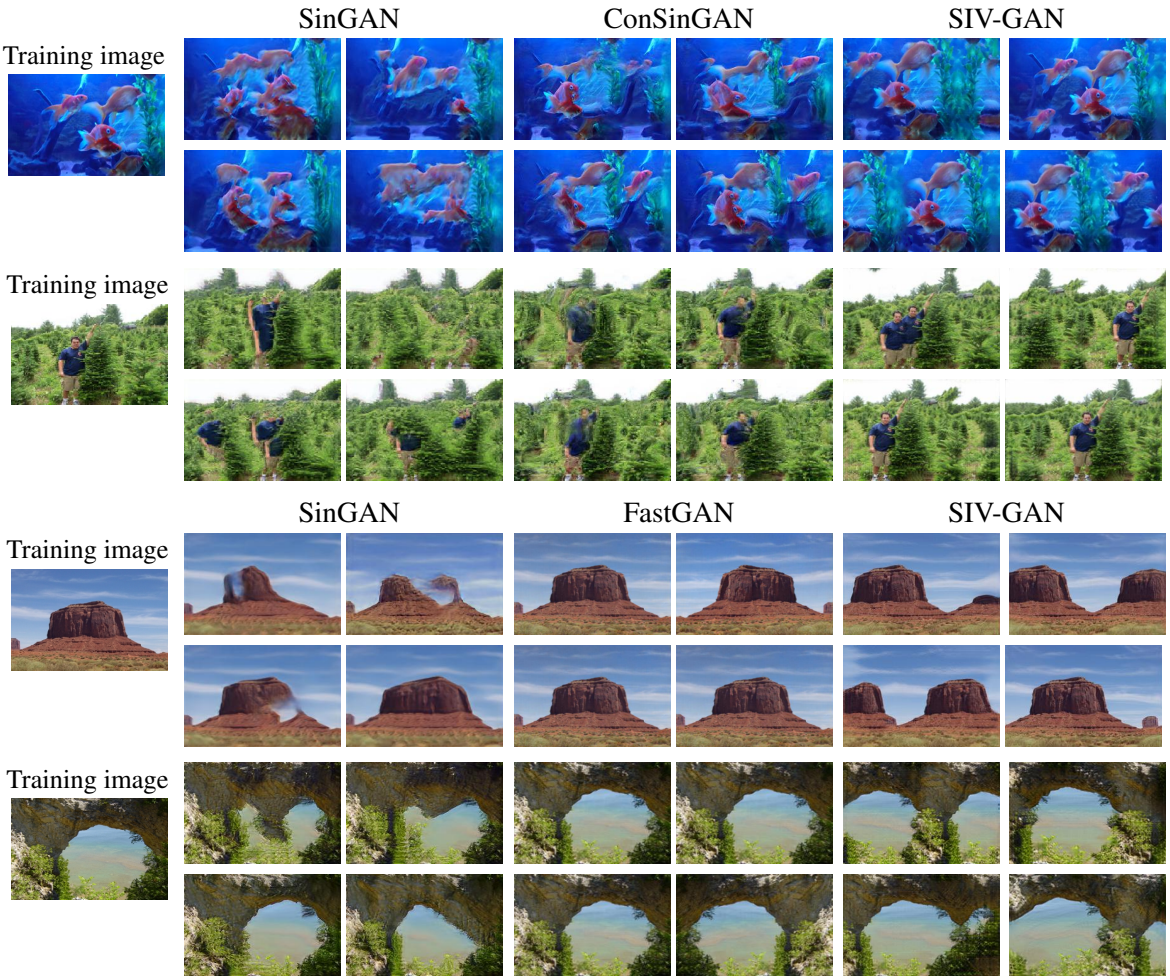

Figure 5.4: Visual comparison between models in the Single Image setting. Single image GANs (*Shaham et al.*, 2019; *Hinz et al.*, 2021) are prone to shuffle patches incoherently (e.g. sky textures below horizon, perturbed fish contours), while the few-shot FastGAN (*Liu et al.*, 2021) suffers from memorization, reproducing only the original image or its flipped version. In contrast, SIV-GAN achieves both high quality and diversity, preserving the realism of image content and layout.

ConSinGAN (*Hinz et al.*, 2021), and with a recent model for few-shot image synthesis, FastGAN (*Liu et al.*, 2021). We use the original implementation codes provided by the authors. While training single image GANs (*Shaham et al.*, 2019; *Hinz et al.*, 2021) on a single video, we applied the reconstruction loss on all frames, as we found this helpful in stabilizing the training.

5.3.2 Comparison to Previous GAN Models

Tables 5.1 and 5.2 present a quantitative comparison between the models in the Single Image and Single Video settings, while the respective visual results are shown in Fig. 5.4 and 5.5. As seen from the tables, SIV-GAN notably outperforms other models in both studied settings. Despite a potential trade-off between quality and diversity, our model achieves better performance in both, reaching lower SIFID values and higher diversity scores. Importantly, only SIV-GAN successfully learns from both single images and videos, generating globally-coherent images of high diversity. Next, we analyse results in these settings separately.

Method	Places					DAVIS-YFCC100M				
	SIFID ↓ $\frac{H\times W}{4}$	SIFID ↓ $\frac{H\times W}{16}$	LPIPS ↑	Pixel ↑ Diversity	Dist. to train	SIFID ↓ $\frac{H\times W}{4}$	SIFID ↓ $\frac{H\times W}{16}$	LPIPS ↑	Pixel ↑ Diversity	Dist. to train
SinGAN	0.15	25.33	0.22	0.52	0.24	0.13	34.52	0.26	0.54	0.30
ConSinGAN	0.08	23.45	0.24	0.50	0.25	0.09	27.33	0.29	0.59	0.31
FastGAN	0.14	16.52	0.15	0.48	0.08	0.13	19.48	0.18	0.49	0.11
SIV-GAN	**0.06**	**12.12**	**0.28**	**0.57**	0.31	**0.08**	**16.30**	**0.33**	**0.66**	0.37

Table 5.1: Comparison with other methods in the Single Image setting on Places (*Zhou et al.*, 2017a) and DAVIS-YFCC100M (*Perazzi et al.*, 2016; *Thomee et al.*, 2016) datasets.

Method	SIFID ↓ $\frac{H\times W}{4}$	SIFID ↓ $\frac{H\times W}{16}$	LPIPS ↑	Dist. to train
SinGAN	2.47	96.35	0.32	0.51
ConSinGAN	2.74	74.50	0.34	0.53
FastGAN	0.79	9.24	**0.43**	0.13
SIV-GAN	**0.55**	**5.14**	**0.43**	0.34

Table 5.2: Comparison in the Single Video setting on DAVIS-YFCC100M.

Single Image. As seen from Fig. 5.1 and Fig. 5.4, given a single image for training, SIV-GAN produces diverse samples of high visual quality. For example, in Fig. 5.4 our model can change the number and placement of foreground objects (e.g., fish and people), or edit the contour and position of rocks in landscape images. Note that such changes preserve the original scene context, retaining the appearance of objects and maintaining the scene layout realism. In contrast, the prior single image GAN models, SinGAN and ConSinGAN, tend to disturb the appearance of objects (e.g., by washing away the contours of fish and people) and disrespect layouts (e.g., sky textures can appear below the horizon), while exhibiting lower diversity in content and layouts. This is reflected in their higher SIFID and lower diversity scores in Table 5.1. On the other hand, the few-shot FastGAN model suffers from memorization issues, only reproducing the training image or its flipped version. In Table 5.1 this is reflected in lowest diversity and Dist. to train (in red) metrics on both datasets. Despite having the lowest diversity, we observe that FastGAN does not reach a low SIFID due to leaking augmentations (horizontal flipping).

Single Video. The Single Video setting provides multiple video frames for training. Consequently, generative models have the potential to combine the knowledge observed in different video frames, and thereby to synthesize more interesting combinations of objects and scenes. Fig. 5.1 and Fig. 5.5 show the images generated by SIV-GAN in this setting. Our model generates high-quality images that are substantially different from the training frames, adding/removing objects and changing the scene geometry. For example, having seen a car following a road (Fig. 5.1), SIV-GAN generates the scene without a car or with two cars. In Fig. 5.5 our model varies the length of a bus and placement of trees, or removes a horse from the scene and changes the jumping obstacle configuration. In contrast, SinGAN, tuned to learn from a single image, does not generalize to the Single Video setting, distorting objects and producing unrealistic layouts (note a low diversity and an extremely high SIFID in Table 5.2). The few-shot FastGAN, on the other hand, generates images with reasonable fidelity, but is still unable to produce samples with non-trivial layout changes, achieving only a very low Dist. to train score (0.13 in Table 5.2). We conclude that only SIV-GAN deals with the challenging Single Video setting successfully, producing globally-coherent images and avoiding memorization of training data.

In Fig. 5.6 we provide an extended analysis, exploring the performance of SIV-GAN and FastGAN while

Figure 5.5: Visual comparison in the Single Video setting. While other models fall into reproducing the training frames or fail to correctly generate objects, SIV-GAN produces high-quality images substantially different from the original training frames.

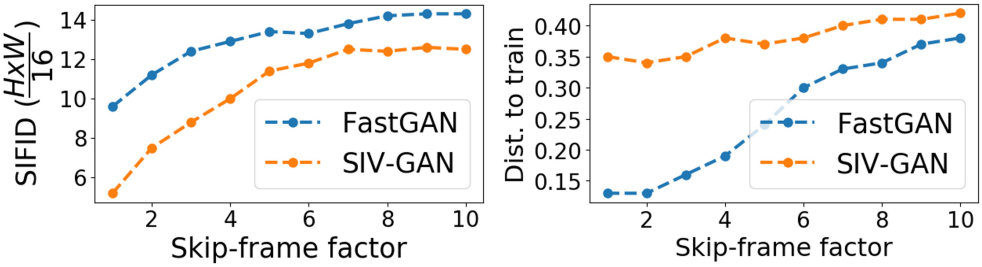

Figure 5.6: Comparison between SIV-GAN and FastGAN in the single video setting. A skip-factor factor of N indicates that each N-th consecutive video frame was used as training data (a smaller factor means higher average similarity between the training frames).

varying the average similarity between training frames of a given video. For this, we take 5 long frame sequences from YFCC100M (*Thomee et al.*, 2016) and for each of them construct 10 different subsets: a subset with index i includes each i^{th} video frame (the total number of frames is always kept to 100). This way, a lower skip-frame factor indicates that the chosen frames are closer in time (and thus more similar to each other), and vice versa. As seen from Fig. 5.6, FastGAN suffers from memorization the most when learning from a set of very similar images: its Dist. to train scores fall dramatically when the skip-frame factor is decreased. This indicates that learning from a few-shot dataset consisting of very similar images can be challenging for prior few-shot GAN models. In contrast, SIV-GAN preserves high Dist. to train scores even for lowest skip-frame factors. Notably, our model also outperforms FastGAN in SIFID for all skip-frame factors.

5.3.3 Ablations

Ablations on the main model components. In Table 5.3 we demonstrate the importance of model's components. In each line we remove only one component, starting from the full SIV-GAN model.

Method	Single Image					Single Video			
	SIFID ↓ $\frac{H\times W}{4}$	SIFID ↓ $\frac{H\times W}{16}$	LPIPS ↑	Pixel ↑ diversity	Dist. to train	SIFID ↓ $\frac{H\times W}{4}$	SIFID ↓ $\frac{H\times W}{16}$	LPIPS ↑	Dist. to train
Full model	0.08	16.30	0.33	0.66	0.37	0.55	5.14	**0.43**	0.34
No Layout br.	0.14	20.29	**0.35**	**0.67**	0.40	0.71	11.70	0.42	0.38
No Content br.	0.08	23.25	0.34	0.64	0.36	0.73	10.43	0.41	0.33
No branches	**0.03**	**7.73**	0.13	0.43	0.12	0.42	**3.73**	0.37	0.18
No DR	0.05	11.99	0.04	0.33	0.06	**0.40**	9.81	0.30	0.32
No FA	0.08	14.81	0.27	0.58	0.33	0.51	4.85	0.41	0.32
No $\mathcal{L}_{D_{low-level}}$	0.08	15.92	0.27	0.56	0.29	0.58	5.32	0.40	0.31

Table 5.3: Ablation study in the Single Image and Video settings on DAVIS-YFCC100M. Indicators of collapsed diversity (low LPIPS, Pixel Diversity) or poor quality (high SIFID) are marked in red.

Regularization	SIFID ↓	LPIPS ↑	Pixel ↑ diversity	Dist. to train
None	**0.05**	0.04	0.33	0.06
zCR	0.05	0.06	0.37	0.09
DS	0.06	0.14	0.45	0.14
DR (im. space)	0.07	0.21	0.52	0.25
DR	0.08	**0.33**	**0.66**	0.37

Table 5.4: Comparison of diversity regularization techniques in the Single Image setting on DAVIS-YFCC100M.

Firstly, we ablate our discriminator architecture, testing it without any branches (No branches), corresponding to a standard GAN discriminator, and without the layout branch or the content branch. The model without branches is trained together with our proposed diversity regularization (DR) and feature augmentation (FA), as well as differentiable augmentations (DA) as in *Karras et al.* (2020a); *Zhao et al.* (2020a). However, as seen from Table 5.3, it memorizes the training images and reproduces them with poor diversity. Using only one of the branches shows good diversity, but the model fails to generate globally-coherent images, having a high $\frac{H\times W}{16}$ SIFID. The qualitative results for these ablation models in the Single Image setting are presented in Fig. 5.7. We observe that the visual results correspond well to the conclusions from Table 5.3. For example, employing none of the branches visually leads to reproducing only the training image. The model without the layout branch generates different objects in various combinations, but the model often fails to reproduce correct positioning of objects or globally-coherent layouts. In particular, there might be a horizon discontinuity, or air balloons may follow unrealistically structured positions in a grid. On the other hand, the model trained without the content branch generates images with more realistic layouts, but does not preserve the content distribution of the original training image, distorting the appearance of objects or perturbing their shapes.

To illustrate further our intuition on the content and layout learning, in Fig. 5.8 we analyse the feature distances between real images in the content and layout embeddings of a trained SIV-GAN discriminator. We take the discriminators trained on the "bus" and "parkour" videos from DAVIS. The plots show the content and layout distances between the middle frame and other frames of the same video. We observe that the distances correlate well with our intuition. Firstly, as nearby frames have very similar content and layouts, the lowest embedding distances are always between adjacent frames, while higher temporal distances from the middle

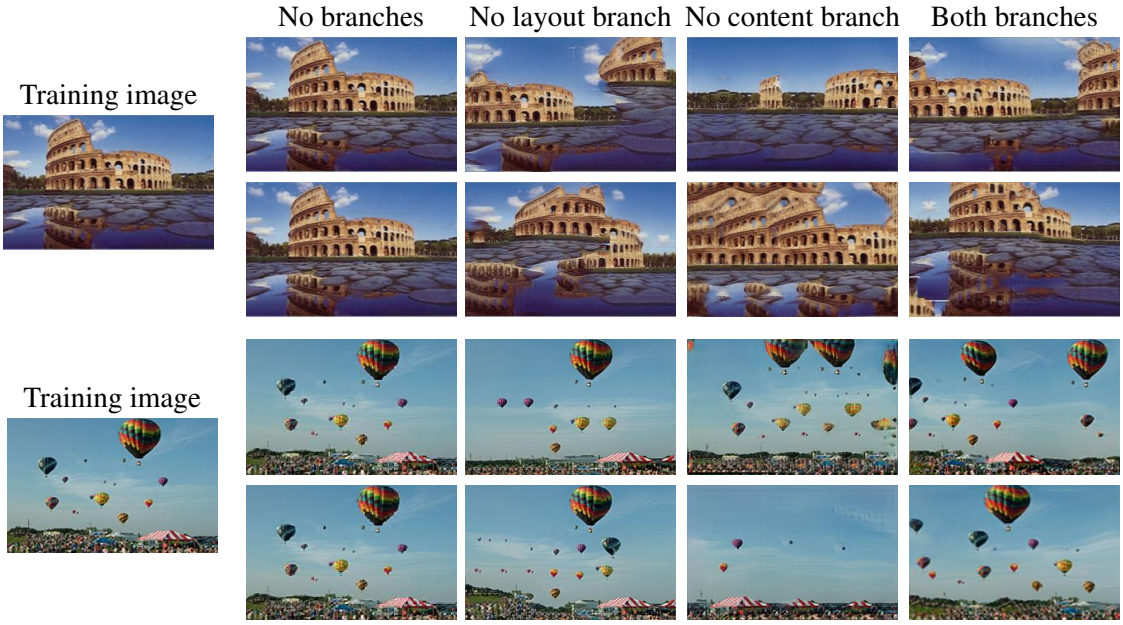

Figure 5.7: Visual results for the ablation on the two-branch discriminator in the Single Image setting. The model with a standard GAN discriminator (No branches) memorizes the training image. The model without the layout branch fails to produce images with realistic layouts or positioning of objects. Absence of the content branch leads to a model that does not preserve well the appearance of objects. Finally, the model with both branches generates diverse images with realistic content and layouts. The qualitative comparison of these models is presented in Table 5.3.

frame lead to higher feature distances. Secondly, the embeddings correlate with the perceptual judgement: the layout distances (solid red) between "bus" frames are lower than for the "parkour" video. Finally, as frames of a short video depict similar content (e.g. same objects), the content distances between frames of the same sequence (solid blue) are much lower than between the middle frames of two different videos (dashed blue).

Next, we observe the effect of the proposed DR and FA. Without DR, the model does not achieve multi-modality, scoring low in all the diversity metrics. The absence of FA notably decreases diversity, resulting in the diversity scores dropping by 0.02-0.08 points. The qualitative results for these models are shown in Fig. 5.9. SIV-GAN without DR does not mitigate overfitting, suffering from mode collapse. The model with DR but without FA manages to achieve diverse image synthesis. However, such a model produces only modest diversity in content and layouts, e.g, it only slightly translates a rock to new locations in the image.

Finally, removing the low-level loss $\mathcal{L}_{D_{low\text{-}level}}$ also results in decreased diversity. According to Eq. (5.1) and (5.2), this term shifts the attention of the loss function from the latest discriminator layers towards earlier layers with smaller receptive field, which complicates the memorization of the whole image. This way, $\mathcal{L}_{D_{low\text{-}level}}$ not only helps to learn low-level image statistics, but also to regularize the discriminator and thus allow a synthesis of higher diversity.

Comparison of DR to alternative techniques. In Table 5.4 we compare our proposed DR to the latent consistency regularization (zCR) (*Zhao et al.*, 2021), diversity-sensitive loss (DS) (*Yang et al.*, 2019), and using no regularization (None). We apply zCR only to the generator loss, leaving the discriminator objective intact. As both zCR and DS operate in the image space, we also test our proposed DR in the image space instead of the G feature space, as in *Choi et al.* (2020). As seen from Table 5.4, our DR noticeably improves over zCR and DS in all diversity metrics. Moreover, we find it beneficial to use DR in the feature space, which leads to more variation in the generated samples. Interestingly, Table 5.4 illustrates a quality-diversity trade-off

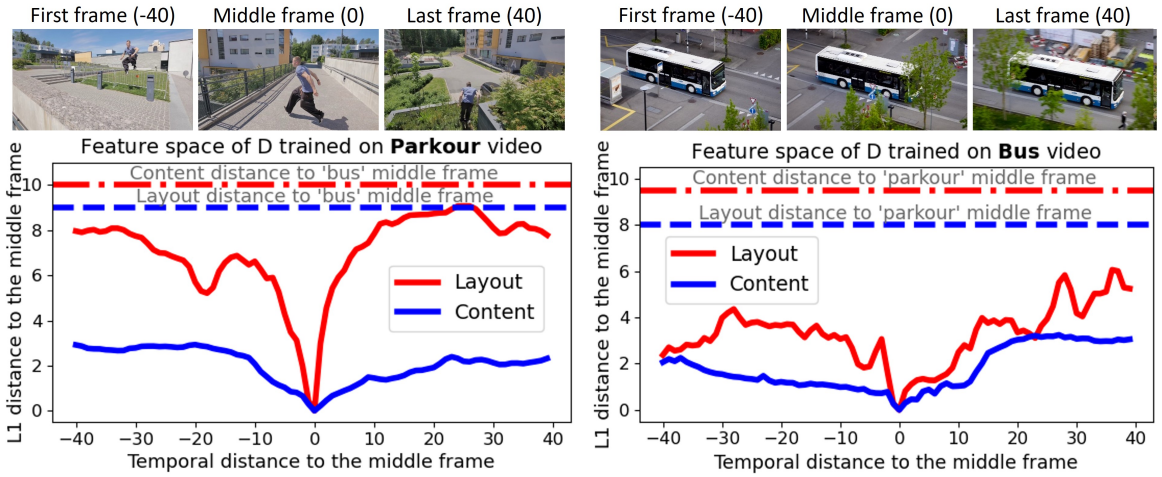

Figure 5.8: Feature distances from video frames to middle frame. The layout embedding distances (solid red) between "bus" frames are lower than for the "parkour" video. The content embedding distances between frames of the same sequence (solid blue) are significantly lower than distance between middle frames of different videos (dashed blue).

λ	SIFID ↓	LPIPS ↑	Pixel ↑ diversity	Dist. to train
0.00	**0.05**	0.04	0.33	0.06
0.05	0.07	0.19	0.50	0.23
0.15	0.08	0.33	0.66	0.37
0.50	0.13	**0.39**	**0.69**	0.46

Table 5.5: Effect of the diversity regularization (DR) strength in the Single Image setting on DAVIS-YFCC100M.

in the Single Image setting, where improvements in diversity lead to deterioration in SIFID.

Ablation on the DR strength. In all our experiments, we used DR with $\lambda = 0.15$. In Table 5.5 we show the effect of setting different λ for DR in the Single Image setting, changing the strength of the diversity regularization. Table 5.5 illustrates that setting the multiplier too high ($\lambda = 0.50$) leads to good diversity, but harms image quality, while setting small values ($\lambda = [0.00, 0.05]$) is beneficial for quality, but deteriorates diversity. We observed that using $\lambda = 0.15$ leads to a good trade-off, resulting in a high diversity among generated samples, while not corrupting the quality of textures and the global layout coherency, so we picked this value for the final version of the model.

Ablation on the number of low-level ResNet blocks. The SIV-GAN discriminator has 3 ResNet blocks before the branching for $D_{low\text{-}level}$. In Table 5.6 and Fig. 5.10 we analyse the effect of applying branching at an earlier or a later discrimination stage, keeping the overall depth of the network equal to 7 ResNet blocks. The results indicate that the branching should be applied neither too early nor too late. Using too few ResNet blocks (1-2) before the branching leads to a reduced capacity of the low-level feature extractor $D_{low\text{-}level}$, so this network becomes unable to learn meaningful content and layout features. Such a model learns the color distribution of an image, but cannot produce a globally-coherent scene and generate textures of good quality (see Fig. 5.10). In Table 5.6 this effect is indicated by a very high SIFID. On the other hand, using too many

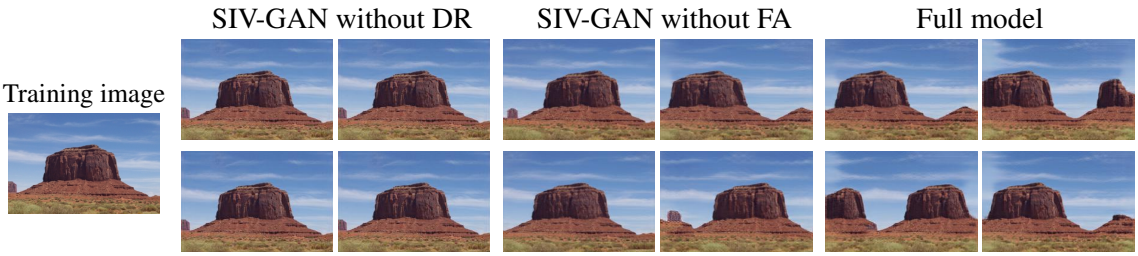

Figure 5.9: Qualitative ablation on the proposed diversity regularization (DR) and feature augmentation (FA). Without DR, the model does not mitigate memorization, producing only images that are perceptually indistinguishable from the original sample. Without FA, SIV-GAN achieves only modest diversity in content and layouts. Finally, using both DR and FA enables generating more interesting scene compositions, varying global scene layouts, or duplicating and removing objects.

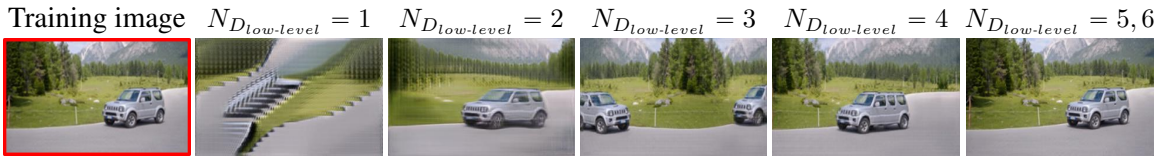

Figure 5.10: Effect of the number of discriminator blocks used before branching. Using too few blocks (1-2) leads to reduced synthesis quality, as $D_{low\text{-}level}$ is unable to extract the features necessary to build the content and layout representations. Increasing the number of blocks (3-4) results in improved quality while maintaining good diversity. Using too many blocks (5-6) leads to the memorization of the training image.

blocks before the branching (5-6) increased the capacity of $D_{low\text{-}level}$, so it becomes easier to memorize the whole image (see Fig. 5.10), which corresponds to low diversity metrics in Table 5.6. We found that using $N_{low-level} = 3$ leads to an optimal quality-diversity trade-off in both the Single Image and Single Video settings.

5.3.4 Comparison to Image Manipulation Methods

In Fig. 5.11 and Table 5.7 we compare the ability of SIV-GAN to compose new scenes with image manipulations methods. For this, for each single image from DAVIS-YFCC100M, using a provided segmentation mask,

$N_{D_{low\text{-}level}}$	Single Image		Single Video	
	SIFID ↓	LPIPS ↑	SIFID ↓	LPIPS ↑
1	0.59	**0.42**	2.75	**0.46**
2	0.13	0.40	1.12	0.45
3	0.08	0.33	0.55	0.43
4	0.06	0.24	0.36	0.38
5	**0.03**	0.15	0.40	0.37
6	**0.03**	0.13	**0.35**	0.36

Table 5.6: Ablation on the number of blocks $N_{D_{low\text{-}level}}$ used before the content-layout branching on the DAVIS-YFCC100M dataset.

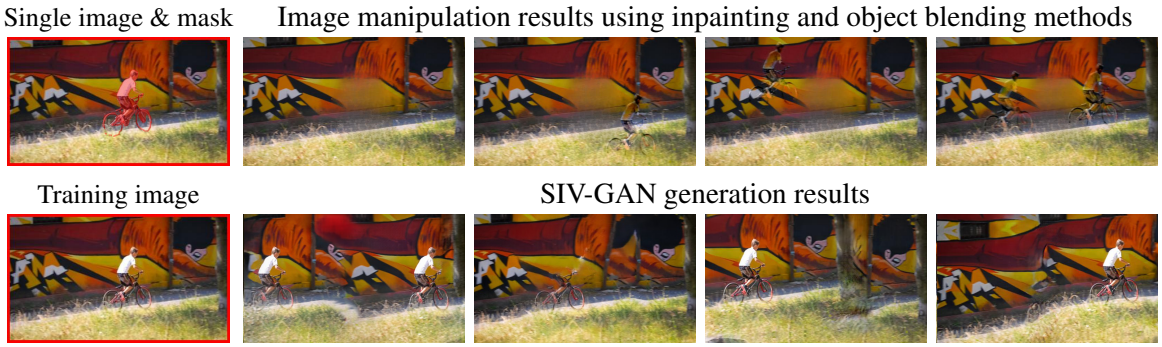

Figure 5.11: Comparison of SIV-GAN synthesis to image manipulation methods in the Single Image setting. The ability of SIV-GAN to remove objects from scenes can be performed by image inpaiting (*Dong et al.*, 2022), while moving or duplicating objects can be done with an object blending method ()zhang2020deep applied at random locations. Nonetheless, unlike SIV-GAN these methods require object annotations, which can be restrictive in practice.

Approach	SIFID ↓	LPIPS ↑	Dist. to train
Image manipulation	**0.07**	0.19	0.18
SIV-GAN	0.08	**0.33**	0.37

Table 5.7: Comparison to image manipulation methods (*Dong et al.*, 2022; *Zhang et al.*, 2020c) on DAVIS-YFCC100M in the Single Image setting.

we first remove a foreground object using a state-of-the-art image inpainting method (*Dong et al.*, 2022) (see Fig. 5.11, second image). Next, we paste the removed object at random locations within the scene by using deep image blending (*Zhang et al.*, 2020c) (see Fig. 5.11, images 3-5). We observe that the images manipulated this way sometimes exhibit inpainting distortions, while the blended objects can be modified too strongly and appear unrealistic. As seen in Table 5.7, SIV-GAN allows a synthesis of significantly higher diversity than image manipulation alternatives. We note that SIV-GAN achieves this without any mask annotations, which eliminates the need for manual annotations.

5.4 Conclusion

In this chapter, we proposed SIV-GAN, a new unconditional generative model which successfully learns from a single image or a single video. In such extremely low-data regimes, our model prevents memorization and generates diverse images that are significantly different from the training set. Inherently, the synthesis of our model is constrained by the appearance of objects present in the original sample. Nevertheless, SIV-GAN can synthesize novel scene compositions by blending objects in different combinations, changing their shape or position, while preserving the original context and plausibility of the scene. Astonishingly, such compositionality is enabled by the model's ability to distinguish objects and backgrounds learnt just from a single image or video.

The fact that one image is enough to train a GAN to generate new diverse images with realistic-looking objects raises an interesting question: can such synthesis be utilized for data augmentation for other applications? Consider various one-shot applications, where neural networks are trained or fine-tuned based on a single provided training example. In such cases, models are naturally susceptible to overfitting, making it reasonable to assume that augmentations produced by GANs can enhance their generalization. This question is studied

in Chapter 6, where we explore the application of one-shot GAN synthesis to synthetic data augmentation in one-shot segmentation applications.

CHAPTER 6

One-Shot Synthesis of Images and Segmentation Masks

In the previous chapter, we introduced a GAN model capable of generating highly realistic scene compositions, even with extremely limited data such as a single image. Notably, this model can rearrange objects within the scene, alter their positions, remove or duplicate them, while preserving their realistic appearance. Building on this, we hypothesize that the model's ability to distinguish between objects and backgrounds can be exploited for segmentation applications. To this end, in this chapter we present a novel GAN training setup where models are trained using a single image along with its corresponding segmentation mask. Using SIV-GAN as the baseline, we introduce a mask synthesis branch in the generator and a masked content attention mechanism in the discriminator. These additions enable us not only to avoid memorization, in contrast to previous image-mask GANs, but also to surpass previous single-image GAN models in terms of both image quality and diversity. Moreover, we demonstrate that our model, called OSMIS, provides valuable data augmentation for one-shot segmentation applications.

Individual Contribution

This chapter is based on the following publication (*Sushko et al.*, 2023b):

One-Shot Synthesis of Images and Segmentation Masks
Vadim Sushko, Dan Zhang, Juergen Gall, Anna Khoreva
IEEE Winter Conference on Applications of Computer Vision (WACV), 2023.
DOI: 10.1109/WACV56688.2023.00622

This publication is the outcome of a collaborative effort involving Vadim Sushko, Dan Zhang, and Anna Khoreva. Juergen Gall supported the project with scientific guidance and valuable suggestions. The initial idea to explore the joint synthesis of images and segmentation masks in low data regimes was proposed by Anna Khoreva. Anna Khoreva also proposed the idea to incorporate segmentation masks into GAN training through a masked attention mechanism, which was further refined through collaborative discussions. The final stage of the project, including final evaluations and paper writing, was largely shaped by Vadim Sushko. Overall, as the first author, Vadim Sushko made significant contributions to discussions, implementation, evaluations, and the writing of the paper.

Contents

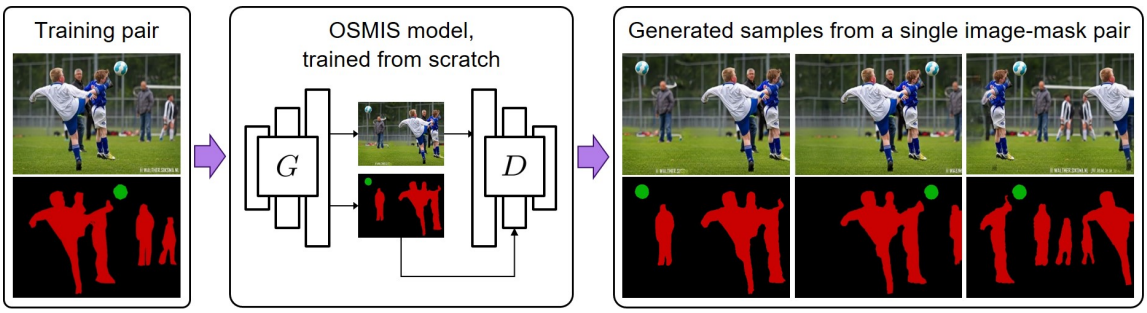

Figure 6.1: We introduce a new task of generating new images and their segmentation masks from a single training pair, without access to any pre-training data. Under this challenging regime, our proposed model achieves a synthesis of a high structural diversity, preserving the photorealism of original images and a precise alignment of produced segmentation masks to the generated content.

6.1 Introduction

Deep neural networks have been shown powerful at solving various segmentation problems in computer vision (*Chen et al.*, 2018; *He et al.*, 2017; *Kirillov et al.*, 2019; *Perazzi et al.*, 2016; *Nilsson and Sminchisescu*, 2018; *Wang et al.*, 2019). The success of these segmentation models strongly relies on the availability of a large-scale collection of labelled data for training. Nevertheless, annotation of a large dataset is not always feasible in practice due to a very high cost of manual labelling of segmentation masks (*Caesar et al.*, 2018). For example, accurately labelling a single image with many objects can take more than 30 minutes (*Zhang et al.*, 2021b). Therefore, diminishing the human effort required for obtaining diverse and precisely aligned image-mask data is an important problem for many practical applications.

Recently, several works (*Tritrong et al.*, 2021; *Zhang et al.*, 2021b; *Li et al.*, 2021a; *Saha et al.*, 2021) proposed to tackle this issue by jointly generating images and segmentation masks with generative adversarial networks (GANs). Utilizing a few provided pixel-level annotations in addition to an image dataset for training, such GAN models become a source of labelled data that can be used to train neural networks in various practical applications. Despite achieving impressive synthesis of segmentation masks based on limited annotated examples, existing image-mask GAN models still require large pre-training image datasets to learn high-fidelity image synthesis. This naturally restricts their application only to the data domains where such datasets are available (e.g., images of faces or cars). However, in some practical scenarios such a dataset can be difficult to find, for example in one-shot segmentation applications (*Amirreza et al.*, 2017), where the object types can be rare. Therefore, in this work we aim to learn a high-fidelity joint mask and image synthesis having as little limitations on the data domain as possible. To this end, we propose a novel GAN training setup, in which we assume availability only of a single training image and its segmentation mask, not relying on any image dataset for pre-training (see Fig. 6.1). After training, we aim to generate diverse new image samples and supplement them with accurate segmentation masks. To the best of our knowledge, we are the first to consider such a training scenario for GANs.

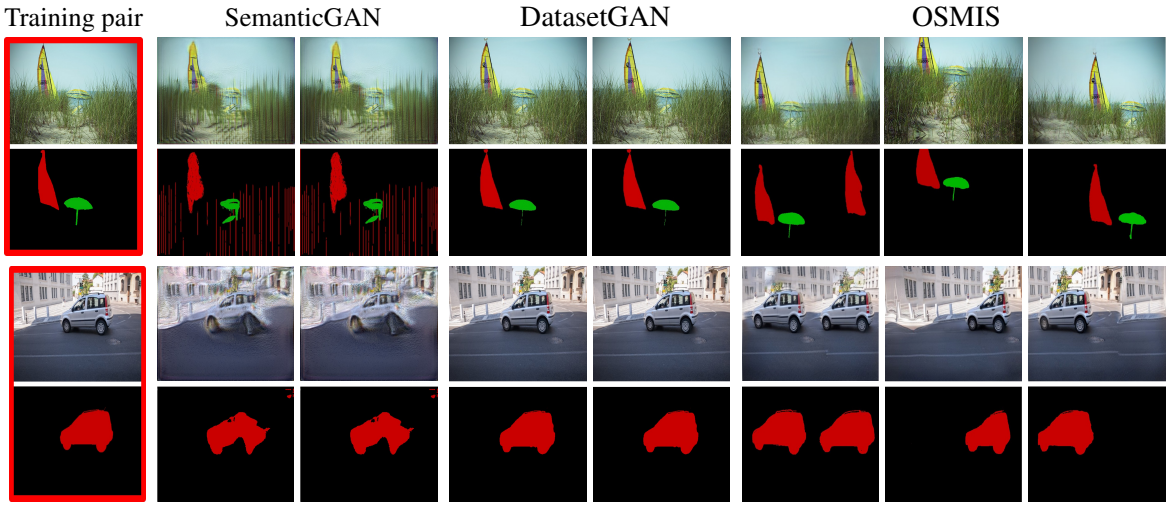

Figure 6.2: A comparison to SemanticGAN (*Li et al.*, 2021a), trained on a single image-mask pair (in red), and DatasetGAN (*Zhang et al.*, 2021b), pre-trained on a single image and trained on a single manual mask annotation. Both models suffer from memorization, while SemanticGAN also has poor quality due to training instabilities. In contrast, OSMIS avoids mode collapse.

Training a GAN from a single training sample is well known to be challenging due to the problem of memorization (*Nagarajan et al.*, 2018), as in many cases the generator converges to reproducing the exact copies of training data. For example, as shown in our experiments, this issue occurs in the prior image-mask GAN models from *Li et al.* (2021a); *Zhang et al.* (2021b) (see Fig. 6.2). Recently, the issue of memorization has been mitigated in the line of works on single-image GANs, which enabled diverse image synthesis from a single training image (*Shaham et al.*, 2019; *Hinz et al.*, 2021; *Sushko et al.*, 2021b). Inspired by these models, we aim to extend this ability to a joint synthesis of images and segmentation masks. To this end, we propose a new model, introducing two modifications to conventional GAN architectures. Firstly, we introduce a mask synthesis branch for the generator, enabling the synthesis of segmentation masks in addition to images. Secondly, to ensure that the produced segmentation masks are precisely aligned to the generated image content, we propose a masked content attention module for the discriminator, allowing it to judge the realism of different objects separately from each other. This way, to fool the discriminator, the generator is induced to label synthesized images accurately. In effect, our proposed model enables a structurally diverse, high-quality **o**ne-**s**hot joint **m**ask and **i**mage **s**ynthesis (see Fig. 6.1), and we thus name it **OSMIS**. As we show in our experiments, compared to prior single-image GANs (*Shaham et al.*, 2019; *Hinz et al.*, 2021; *Sushko et al.*, 2021b), OSMIS not only offers an additional ability to generate accurate segmentation masks, but also achieves higher quality and diversity of generated images.

Despite using only a single image-mask pair for training, OSMIS can generate a set of labelled samples of a high structural diversity, which sometimes cannot be achieved with standard data augmentation techniques (e.g., flipping, zooming, or rotation). For example, for a given scene, OSMIS can change the relative locations of foreground objects or edit the layout of backgrounds (see Fig. 6.1, 6.4, 6.5). Moreover, in contrast to *Li et al.* (2021a); *Zhang et al.* (2021b), OSMIS can successfully handle masks of different types, e.g., having class-wise (see Fig. 6.1) or instance-wise (see Fig. 6.4) annotations. This suggests a good potential of our model to serve as a source of additional labelled data augmentation for practical applications. We demonstrate this potential in Sec. 6.3.4, where we apply OSMIS at the test phase of one-shot video object segmentation (*Perazzi et al.*, 2016) and one-shot semantic image segmentation (*Amirreza et al.*, 2017). The results indicate that the data generated by OSMIS helps to improve the performance of state-of-the-art networks: OSVOS (*Caelles et al.*,

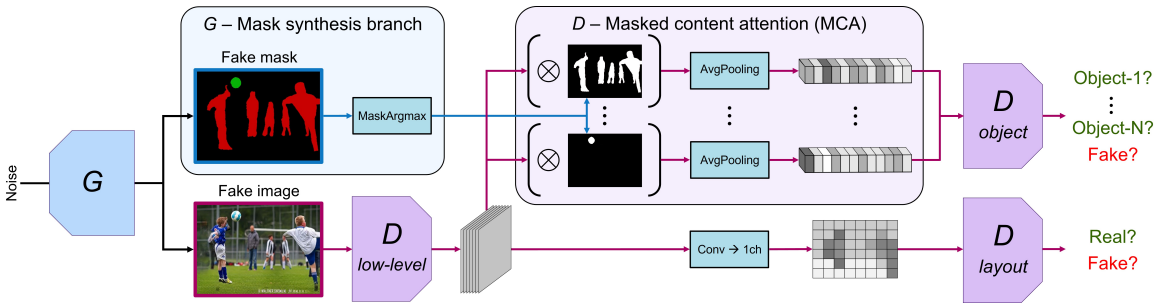

Figure 6.3: OSMIS model. A simple mask synthesis branch in the generator G allows the generation of segmentation masks of objects together with images. The precise alignment between the masks and the generated image content is enforced by a masked content attention (MCA) module in the discriminator D, designed to evaluate the realism of different objects separately from each other.

2017), STM (*Oh et al.*, 2019), and RePRI (*Boudiaf et al.*, 2021), providing complementary gains to standard data augmentation. We find these results promising for utilization of one-shot image-mask synthesis in future research.

6.2 Method

Given a single image with its pixel-level segmentation mask and assuming no access to any pre-training data, we aim to generate a diverse set of new image-mask pairs. In this section, we present OSMIS, our one-shot image-mask synthesis model. Adopting our SIV-GAN model, introduced in Chapter 5, as a state-of-the-art image synthesis baseline (Sec. 6.2.1), we propose modifications to the generator (Sec. 6.2.2) and discriminator (Sec. 6.2.3) architectures, enabling one-shot synthesis of segmentation masks that are precisely aligned with generated images.

6.2.1 SIV-GAN Baseline

As the baseline network architecture for OSMIS, we use SIV-GAN (*Sushko et al.*, 2021b), as it generates high-quality novel scene compositions with realistic objects, thereby achieving the highest quality and diversity of one-shot image synthesis among previous works. The detailed decription of the SIV-GAN architecture is provided in Chapter 5. To recap, SIV-GAN has a two-branch discriminator, in which an input image x is first transformed into a feature representation $F(x)$ by a low-level discriminator $\mathcal{D}_{low-level}$. Next, two separate discriminators assess different aspects of $F(x)$. The content discriminator $\mathcal{D}_{content}$ judges the realism of objects regardless of their spatial location by averaging out the spatial information contained in $F(x)$ via global average pooling. On the other hand, the layout discriminator $\mathcal{D}_{layout}$ evaluates the realism only of the spatial scene layouts by squeezing $F(x)$ with a one-channel convolution. In addition, the discriminator applies feature augmentation in the content and layout representations of $F(x)$ to further increase the high-level diversity among generated samples. The adversarial loss of the SIV-GAN model consists of three terms:

$$\mathcal{L}_{adv}(G, D) = \mathcal{L}_{D_{content}} + \mathcal{L}_{D_{layout}} + 2\mathcal{L}_{D_{low\text{-}level}}, \tag{6.1}$$

where each term is the mean of binary cross entropies obtained at different layers of respective discriminator parts.

In contrast to one-shot image synthesis, we assume that the single training image is provided with its pixel-level mask of objects, not assuming any fixed annotation type (e.g., class-wise or instance-wise). To incorporate it into the training process, we introduce two modifications to the architecture of the baseline

model. Firstly, we propose to generate segmentation masks simultaneously with images via an additional generator's mask synthesis branch. Secondly, to enforce the precise mask alignment to the generated image content, we re-formulate the objective of the content discriminator $\mathcal{D}_{content}$, designing it to judge the fidelity of different objects separately from each other. This is made possible by the introduced masked content attention module, which builds a separate content feature vector for each object considering the provided segmentation mask. The overview of our model architecture is shown in Fig. 6.3. Next, we describe the proposed modifications in detail.

6.2.2 Mask Synthesis Branch in the Generator

In line with *Tritrong et al.* (2021); *Zhang et al.* (2021b), we hypothesize that during training the generator should be able to learn discriminative features that completely describe the appearance of generated objects. Thus, while synthesizing an image, we collect feature activations of the generator layers and use them as input for the mask synthesis branch. In contrast to *Tritrong et al.* (2021); *Zhang et al.* (2021b), we use only the activations after the last generator block, as this simplest solution already performs well in our experiments. Using a simple convolution followed by a softmax activation, we transform these features into an N-channel soft probability map, where each channel corresponds to one of $N-1$ objects of interest in the segmentation mask or to the background. To obtain the final discrete mask prediction, an argmax operation T along the channel dimension is applied.

To enable the training of the mask synthesis branch with the discriminator loss, the generated masks should allow back-propagation of gradients, similarly to generated images. In our experiments, feeding the discriminator the continuous segmentation probability maps obtained before the non-differentiable argmax operation T impaired the GAN training, as the discriminator learnt to detect the continuous-discrete discrepancy between fake and real inputs. Thus, inspired from *Van den Oord et al.* (2017); *Bengio et al.* (2013), we enable back-propagation through argmax by developing a straight-through gradient estimator:

$$\text{MaskArgmax}(y) = y + \arg\max(y) - sg[y], \tag{6.2}$$

where sg denotes a stop-gradient operation. This way, the discriminator is provided with the generated masks in a discrete form $T(y)$, which enables its effective training, while the generator can be trained with the gradients passing through its probability map prediction y.

Yet, this solution can sometimes lead to degenerate solutions, e.g., when all the pixels are predicted as the background channel. This cannot be corrected during training, as in this case the gradient flow through all the other mask channels is blocked. We found that it can be mitigated by softening the argmax operation T at the beginning of training. For this, during the first P_0 epochs we regard each mask pixel as a random variable following Bernoulli distribution:

$$T(y) = \begin{cases} \sim \text{Bernoulli}(y) & \text{epoch} < P_0, \\ \text{argmax}(y) & \text{epoch} \geqslant P_0. \end{cases} \tag{6.3}$$

6.2.3 Masked Content Attention in the Discriminator

To provide a training signal to the generator's mask synthesis branch, we propose to incorporate the learning of the image-mask alignment to the objective of the content discriminator $\mathcal{D}_{content}$. In SIV-GAN, $\mathcal{D}_{content}$ was designed to judge the content distribution of the whole given image. Considering the provided segmentation mask, we can now select the image areas belonging to different objects, and require the discriminator to learn their appearance separately from each other. With this objective, as the discriminator can compare the appearance of the area belonging to the same object in real and fake images, it encourages the generator not only to synthesize realistic objects, but also to label them correctly.

To this end, we introduce a masked content attention (MCA) module. As shown in Fig. 6.3, MCA receives a downsampled segmentation mask y along with an intermediate feature representation $F(x) = \mathcal{D}_{low-level}(x)$ of an input image x, and thereout produces a set of N content vectors, corresponding to the masked content representations of each of the $N-1$ objects of interest and the background:

$$\text{MCA}(x, y) = \{\text{AvgPool}\,(F(x) \times \mathbb{1}_{y=i})\}_{i=1}^{N}. \tag{6.4}$$

Accordingly, we re-design the objective of the content discriminator (further denoted $\mathcal{D}_{object}$). For each of the obtained object representations, our proposed $\mathcal{D}_{object}$ is induced to predict a correct identity of each object or background of a real image, while all the identities of fake images should be categorized as an additional fake class:

$$\begin{aligned}\mathcal{L}_{D_{object}} = &- \mathbb{E}_{(x,y)}\left[\sum_{i=1}^{N} \alpha_i \log \mathcal{D}^i_{object}(\text{MCA}^i(x, y))\right] \\ &- \mathbb{E}_z\left[\sum_{i=1}^{N} \log(1 - \mathcal{D}^{fake}_{object}(\text{MCA}^i(G(z)))\right],\end{aligned} \tag{6.5}$$

where z is the noise vector used by the generator G to synthesize a fake image-mask pair $G(z) = \{G_x(z), G_y(z)\}$, (x, y) denotes the real image-mask pair, and $\mathcal{D}^i(*)$ is the discriminator logit for the object i. Considering that different objects or background can occupy different areas, we introduce a class balancing weight α_i, which is the inverse of the per-pixel class frequency in the segmentation mask y:

$$\alpha_i = \frac{(\text{sum}(\mathbb{1}_{y=i}))^{-1}}{\sum_{j=1}^{N}(\text{sum}(\mathbb{1}_{y=j}))^{-1}}. \tag{6.6}$$

Note that the balancing is applied only for real images, as in Eq. 6.5 all fake objects are considered as the same class.

Our $\mathcal{D}_{object}$ learns the content distribution of each object separately. The advantage of such a training scheme is two-fold. Firstly, a generator now needs to synthesize correct segmentation masks in order to fool the discriminator. The precise image-mask alignment is thus enforced directly by the adversarial loss, without the need for using additional networks or manual annotation. Secondly, as MCA provides representations only of separate objects, $\mathcal{D}_{object}$ has restricted access to the content distribution of the whole image. In effect, the discriminator memorization of the whole training sample becomes more difficult, which enables more diverse image synthesis (see Table 6.3).

6.3 Experiments

We evaluate OSMIS as follows. Firstly, we provide the qualitative and quantitative assessment of the achieved one-shot image-mask synthesis, evaluating the quality and diversity of generated images, as well as their alignment to the produced segmentation masks. Secondly, we apply OSMIS to two one-shot segmentation applications, demonstrating the potential of the generated image-mask pairs to be used as data augmentation.

6.3.1 Experimental Setup

Training details. We train our model with the loss from Eq. (6.5) for the object discriminator $\mathcal{D}_{object}$, setting P_0=15000. We employ differentiable augmentation (DA) of input images and masks while training the discriminator, using the whole set of transformations as proposed in *Karras et al.* (2020a). We use an exponential moving average of the generator weights with a decay of 0.9999, and follow SIV-GAN in setting all the other hyperparameters.

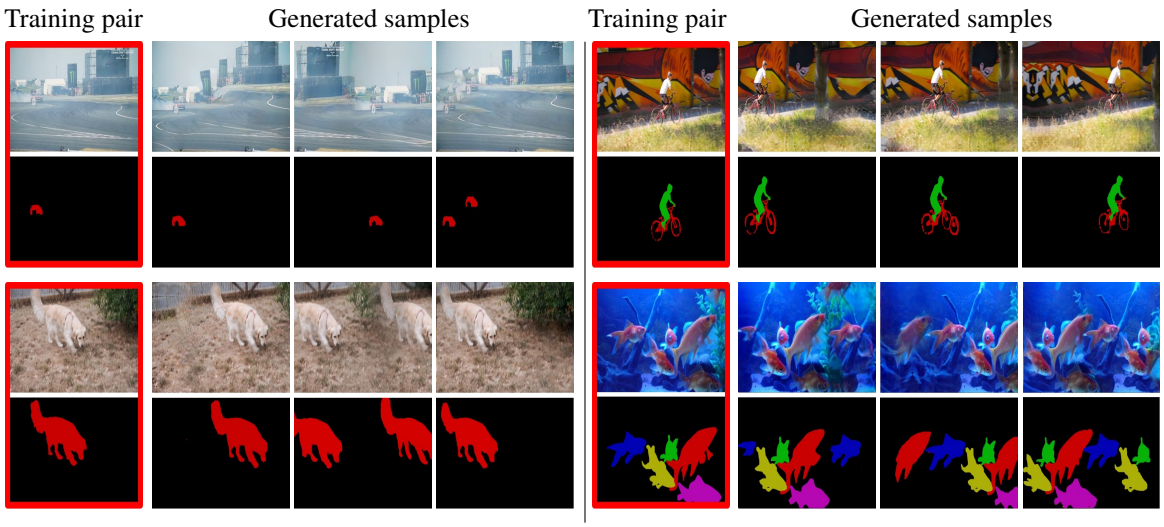

Figure 6.4: Qualitative results of OSMIS on DAVIS (*Perazzi et al.*, 2016). Given a single image-mask pair for training, our model achieves high-fidelity image synthesis with a high structural diversity, changing the positions of objects or editing the layout of backgrounds. For each synthesized image, it produces segmentation masks that accurately annotate the generated content. Training pairs are shown in red frames.

Datasets. To evaluate the synthesis, we use the DAVIS dataset (*Perazzi et al.*, 2016), originally introduced for video object segmentation. For each video from the DAVIS-17 validation split, we take the first frame and its segmentation mask of objects, which results in 30 image-mask pairs on which we train separate models. The resolution is set to 640x384. For additional visual results, we use samples from COCO (*Lin et al.*, 2014), trying to closely fit their resolution. Note that the datasets have different annotation types (class-wise and instance-wise).

Metrics. To mind a possible quality-diversity trade-off in our one-shot regime (*Robb et al.*, 2021; *Li et al.*, 2020), we assess the quality and diversity of generated images separately. For this, we report the average SIFID (*Shaham et al.*, 2019) as the measure of image quality, while the average LPIPS (*Zhang et al.*, 2018b) between the pairs of generated images is used to assess the diversity of synthesis.

On the other hand, evaluating the quality of generated masks is challenging, because generated images do not have ground truth segmentation annotations. To bypass this issue, we propose to evaluate the alignment between generated masks and synthetic images using an external segmentation network. For this, we take a UNet (*Ronneberger et al.*, 2015) and train it on the generated image-mask pairs for 500 epochs. After training, we compute its mIoU performance on the original real image, augmented with standard geometric transformations. Intuitively, a good performance on this test reveals that synthetic masks describe well the objects from the real data, indicating precise alignment between the generated images and their masks.

6.3.2 Evaluation of One-Shot Image-Mask Synthesis

Qualitative results. Fig. 6.4 and 6.5 show the image-mask pairs generated by OSMIS trained on samples from DAVIS and COCO. Given only a single image-mask pair, our model learns to generate new image-mask pairs, demonstrating a remarkable structural diversity among samples, photorealism of synthesized images, and a high quality of generated annotations. For example, OSMIS can re-synthesize the provided scene with a different number of foreground objects, e.g., more dogs (3^{rd} example in Fig. 6.4), less people (2^{nd} example

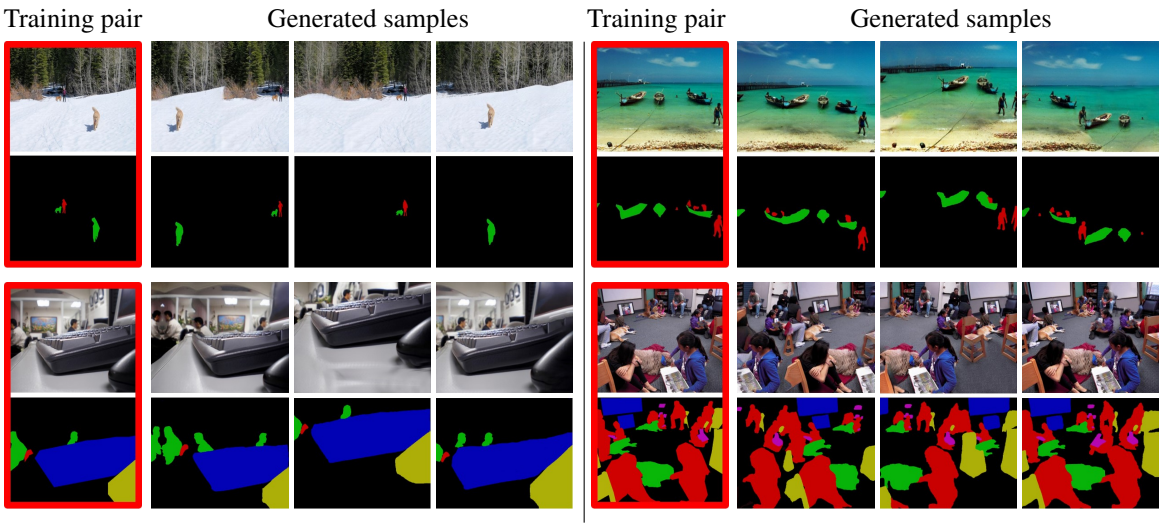

Figure 6.5: Qualitative results of OSMIS on COCO (*Lin et al.*, 2014). OSMIS successfully deals with different scene types and annotation styles. For example, it achieves high quality and diversity for both indoor and outdoor scenes, or sparse and dense annotations of foreground objects.

in Fig. 6.5), or edit layouts of backgrounds (1st examples in Fig. 6.4-6.5), in all cases providing accurate segmentation masks for the re-synthesized scenes. In addition, we remark that OSMIS successfully deals with very different scene types (e.g., both indoor and outdoor scenes), supports masks with both sparse and dense object annotations (e.g., foreground objects occupying small or large image areas), and can handle masks with many objects or even separate instances of the same semantic class (e.g., fish in 4th example in Fig. 6.4).

We note that reaching diverse synthesis from a single image-mask pair is extremely difficult from a single sample. For example, as can be seen in Fig. 6.2, in this regime, prior image-mask GAN methods DatasetGAN and SemanticGAN suffer from memorization issues and training instabilities. Like OSMIS, SemanticGAN was trained from scratch, using a single provided image-mask pair as real data. On the other hand, the training of DatasetGAN consisted of two stages: pre-training of the StyleGAN (*Karras et al.*, 2019) backbone architecture on the single provided training image, and training a label synthesis branch with manual segmentation annotations of generated images. Since StyleGAN typically collapsed to generating the same image, annotating a single generated sample was enough to train the label synthesis branch in our experiments. As seen from Fig. 6.2, both SemanticGAN and DatasetGAN suffer from memorization issues, always producing the same image that repeats the layout of the training sample. In addition, SemanticGAN showed unstable training, resulting in a low visual quality of generated images and noisy annotations. In contrast, OSMIS achieves high diversity and visual quality of generated image-masks at the same time. For example, in the examples from Fig. 6.2 our model changes the number of sails or cars, at the same time editing the layout of the backgrounds, while still preserving the realism of objects.

Quantitative results. We compare the quality and diversity of generated *images* to the single-image GAN models SinGAN (*Shaham et al.*, 2019), ConSinGAN (*Hinz et al.*, 2021) and SIV-GAN (*Sushko et al.*, 2021b). The image-mask synthesis is compared to DatasetGAN (*Zhang et al.*, 2021b) and SemanticGAN (*Li et al.*, 2021a). We use the official repositories provided by the authors.

The quantitative comparison of the image synthesis to single-image GAN models on DAVIS-17 is presented in Table 6.1. Compared to these models, OSMIS not only offers an additional ability to generate segmentation masks, but also achieves higher image quality and diversity. As seen in Table 6.1, despite a potential trade-off between SIFID and LPIPS, our model outperforms previously published baselines in both metrics by a notable margin. Further, Table 6.2 demonstrates that prior image-mask methods, DatasetGAN

Method	SIFID↓	LPIPS↑
SinGAN	0.131	0.267
ConSinGAN	0.103	0.296
SIV-GAN	0.091	0.347
OSMIS (ours)	**0.073**	**0.387**

Table 6.1: Comparison of image quality and diversity to single-image GANs on DAVIS-17. Bold denotes the best performance.

Method	SIFID↓	LPIPS↑	mIoU
DatasetGAN	0.118	0.007	91.1*
SemanticGAN	0.211	0.012	65.8
OSMIS (ours)	**0.073**	**0.387**	**86.6**

Table 6.2: Comparison to prior image-mask GANs on DAVIS-17. Bold denotes the best performance. Red indicates mode collapse. * Indicates manual annotation of masks for DatasetGAN training.

and SemanticGAN, suffer from instabilities and fail to achieve diverse synthesis, scoring very low in LPIPS.

6.3.3 Ablations

In Table 6.3 we compare the proposed masked content attention module (MCA) with three alternative discriminator mechanisms to provide supervision for the generator's mask synthesis branch. The simplest baseline is to concatenate the input masks to images, requiring the discriminator to judge their realism jointly. Another method is to use projection (*Miyato and Koyama*, 2018), by taking the inner product between the last linear layer output of $D_{\text{low-level}}$ and the pixel-wise linear projection of the input mask. Finally, we compare to the approach of SemanticGAN (*Li et al.*, 2021a), adding a separate discriminator network $\mathcal{D}_m$ which takes both segmentation masks and images, and propagate its gradients only to the generator's mask synthesis branch. While training these baselines, we preserve all the OSMIS hyperparameters, but remove the MCA and use the original $\mathcal{D}_{content}$ as in SIV-GAN. As seen from mIoU in Table 6.3, MCA enables the generation of segmentation masks with the best alignment to the generated image content, as measured by an external segmentation network. Notably, while all the alternative methods negatively affect diversity, MCA improves it (0.387 vs 0.368 LPIPS), highlighting its regularization effect which prevents the discriminator memorization of training data.

While enabling on average higher image diversity and mask quality, we found that MCA can struggle if the training sample contains annotations of fine-grained object details, due to downsampling of input masks. This is illustrated in Fig. 6.6 and Table 6.4, for which we train OSMIS with different numbers of low-level discriminator blocks $N_{\text{low-level}}$, corresponding to different degrees of mask downsampling. We observe a trade-off between the quality of images and masks: decreasing $N_{\text{low-level}}$ improves the image diversity and pixel-level mask fidelity, but harms image quality. We selected $N_{\text{low-level}} = 4$ as a compromise between the metrics in Table 6.4, even though this configuration sometimes fails to annotate small object details (as in Fig. 6.6). Note that despite this limitation, MCA still outperforms other methods that do not use downsampling on DAVIS-17 (see Table 6.3), and leads to image-mask pairs that are more useful as data augmentation, as discussed next.

Mask supervision	SIFID↓	LPIPS↑	mIoU
None	0.071	0.368	-
Projection	**0.071**	0.362	72.1
Input concat.	0.079	0.328	82.4
SemanticGAN D_m	0.074	0.351	83.3
MCA (ours)	0.073	**0.387**	**86.6**

Table 6.3: Comparison of MCA to other mask synthesis supervision mechanisms on DAVIS-17. Red indicates decreased diversity compared to the baseline. Bold denotes the best performance.

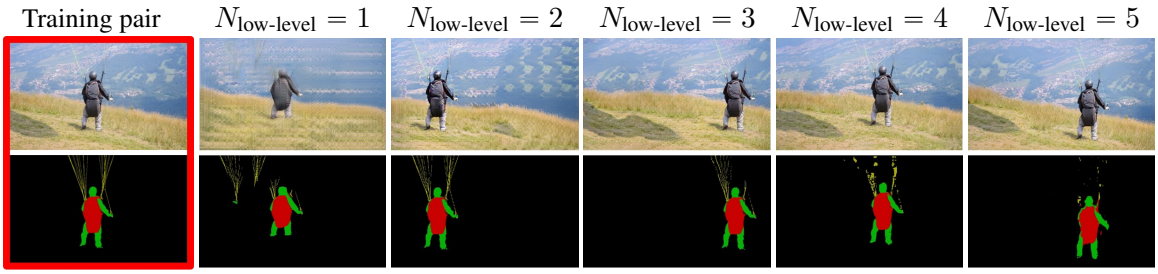

Figure 6.6: Trade-off between the image and mask quality when varying the number of $\mathcal{D}_{\text{low-level}}$ discriminator blocks. Increased number improves image quality, but harms the ability of masks to capture fine-grained object details due to stronger downsampling during training.

6.3.4 Application to One-Shot Segmentation Tasks

After training, OSMIS can augment the provided image-mask pair with novel diverse samples. As such diversity (edited backgrounds, objects changing relative locations) is difficult to achieve by means of standard data augmentation, we foresee a potential usage of our model as a source of labelled data augmentation. Thus, in what follows, we test the efficacy of OSMIS generations when applied at test phase of two one-shot segmentation applications.

One-shot video object segmentation. We apply our model to the semi-supervised one-shot video segmentation benchmark DAVIS (*Perazzi et al.*, 2016). At test phase, this task provides a video and the segmentation mask of objects only in the first frame, while a model is required to segment all the remaining video frames. We select two popular models from the literature: OSVOS (*Caelles et al.*, 2017), which fine-tunes the network weights on the first video frame and segments other frames independently, and STM (*Oh et al.*, 2019), which propagates the segmentation prediction sequentially using a space-time memory module. We conduct experiments on two DAVIS splits: *DAVIS-16*, having 20 videos with a single annotated object; and its extension *DAVIS-17*, having 30 videos with multi-instance annotations. To evaluate the video segmentation, we compute the average of the mean mIoU region similarity ($\mathcal{J}$) and the mean contour accuracy ($\mathcal{F}$) across all videos, which is a popular metric for this task (*Perazzi et al.*, 2016).

One-shot semantic image segmentation. The second setup is the one-shot image segmentation benchmark COCO-20^i (*Lin et al.*, 2014). In this task, a segmentation model is first trained on a large dataset. At test phase, the model is given a single image-mask pair (support set) with an object of a previously unseen test class, and is then required to segment another sample (query image) containing instances of the same class. We conduct experiments with the state-of-the-art RePRI network (*Boudiaf et al.*, 2021). COCO-20^i contains 80 classes, which are divided into 4 folds, with 60 base and 20 test classes in each fold. To test OSMIS, we randomly selected 5 support samples for each test class, resulting in 100 image-mask pairs in each of the folds,

$N_{\text{low-level}}$	SIFID↓	LPIPS↑	mIoU
1	0.262	0.395	82.4
2	0.165	**0.404**	**87.1**
3	0.102	0.394	86.9
4	0.073	0.387	86.6
5	**0.070**	0.321	83.9

Table 6.4: Ablation on the number of $\mathcal{D}_{\text{low-level}}$ discriminator blocks on DAVIS-17. Bold denotes the best performance.

and trained OSMIS on all of them separately. The performance of this task is evaluated separately for each fold, using the average mIoU across many different support-query examples.

Experimental setup. For both applications, we train OSMIS on the single given image-mask pair (the first video frame or support sample). We try to closely fit the resolution of each image from COCO, and set a fixed resolution of 640x384 for images from the DAVIS benchmark. After training, we generate a pool of synthetic image-mask pairs consisting of $n = 100$ samples. As OSMIS can occasionally fail and synthesize noisy examples, we compute the SIFID metric (*Shaham et al.*, 2019) for each generated image as a measure of its quality. Ranking the images by the average of SIFID ranks at different InceptionV3 layers, we exclude bad-quality samples by filtering out 15% lowest-ranked images. Finally, we add the remaining synthetic samples to the original image-mask pair as data augmentation. However, the exact method of utilizing data augmentation depends on the segmentation network, as described next.

OSVOS (*Caelles et al.*, 2017) fine-tunes weights of a pre-trained segmentation network on the image and mask of the first frame of a given video sequence. At each fine-tuning epoch, we double the batch size and randomly add generated image-mask pairs to the original data. Therefore, we keep the 50%-50% ratio between real and synthetic data, which we found to yield the best video segmentation performance.

STM (*Oh et al.*, 2019) scans a given video sequence frame-by-frame, starting from the first frame, for which a mask annotation is provided. This image-mask pair, as well as each K-th pair of a video frame and its segmentation prediction are added to a spatio-temporal memory bank. The memory bank is used to make the segmentation prediction of the latest video frames more accurate. To employ data augmentation, we added synthesized image-mask pairs to the STM memory bank at step 0, before processing the first video frame. To fit the memory bank into GPU memory, we had to limit the number of added samples to 10, which were sampled randomly from the synthetic pool.

RePRI (*Boudiaf et al.*, 2021) trains a small pixel-level classifier given a single support image-mask pair containing an object of a previously unseen class. We simply provide synthetic image-mask pairs as data augmentation for the original data. To fit the extended support set into GPU memory, we limited the number of added samples to 10. This way, the task of RePRI could be technically regarded as 11-shot semantic image segmentation, where all the available support data originates from a provided data sample.

Among the used segmentation models, only OSVOS (*Caelles et al.*, 2017) applies data augmentation at test phase (random combinations of image-mask flipping, zooming, and rotation). Thus, in experiments we compare our synthetic data augmentation to this pipeline (referred to as *standard* augmentation).

6.3.5 Effectiveness of Synthetic Data Augmentation

The performance of segmentation networks using different data augmentation is shown in Tables 6.5 and 6.6. To account for the variance between runs, all the results are averaged across 5 runs with different seeds for augmentation. We generally managed to reproduce the official reported numbers closely, with the exception

Network	Augmentation: Standard	Augmentation: **Ours**	DAVIS-16	DAVIS-17
OSVOS	✗	✗	76.9	51.3
	✓	✗	78.5 (80.2)	52.9 (52.8)
	✗	✓	78.2	52.6
	✓	✓	**79.8**	**54.2**
STM	✗	✗	89.7 (89.4)	72.4 (72.2)
	✓	✗	89.9	72.4
	✗	✓	90.1	72.6
	✓	✓	**90.2**	**72.7**

Table 6.5: Effect of data augmentation on the mean of mIoU and contour accuracy ($\mathcal{J}\&\mathcal{F}$) of one-shot video object segmentation. Bold denotes the best performance. Round brackets show the results reported in (*Caelles et al.*, 2017; *Oh et al.*, 2019). Reproduced and reported numbers for OSVOS differ as its official code lacks some model components.

Network	Augmentation: Standard	Augmentation: **Ours**	$COCO^0$	$COCO^1$	$COCO^2$	$COCO^3$
RePRI	✗	✗	31.2 (31.2)	38.3 (38.1)	32.9 (33.3)	33.2 (33.0)
	✓	✗	31.8	38.5	33.4	33.8
	✗	✓	32.4	38.7	33.7	34.3
	✓	✓	**32.8**	**39.0**	**34.1**	**34.6**

Table 6.6: Effect of synthesized data augmentation on mIoU of one-shot image segmentation. In each data split, support examples were sampled from a subset of 100 image-mask pairs, for which our model was trained. Bold denotes the best performance. The round brackets contain the numbers reported in (*Boudiaf et al.*, 2021).

of OSVOS, for which the official codebase[1] does not implement the model in full configuration. As seen in Tables 6.5 and 6.6, the synthetic data augmentation produced by OSMIS yields a notable increase in segmentation performance, on average improving the metrics of OSVOS and STM by 1.3 and 0.3 $\mathcal{J}\&\mathcal{F}$ points, and RePRI by 0.9 mIoU points compared to the models using no data augmentation. Despite a possible mismatch between OSMIS training resolution and target image size (e.g., 640x384 vs 854x480 for DAVIS) and the need for image resizing, our synthetic data augmentation consistently outperforms standard data augmentation for STM and RePRI, and is almost on par for OSVOS, which was originally tuned for training with standard data augmentation. These results validate the ability of OSMIS to generate structurally diverse data augmentation of sufficient quality in the one-shot regime. Finally, we note that the effect of OSMIS generations is complementary to standard data augmentation, as the best results for all models are observed when the two pipelines are used in combination.

Table 6.7 demonstrates the efficiency of synthetic data augmentation obtained with different GAN models. The previous image-mask models DatasetGAN and SemanticGAN both show poor applicability in the scenario of one-shot applications due to poor synthesis performance. Further, among the comparison methods for mask synthesis supervision, the strongest increase in performance is achieved with our proposed MCA module. This

[1] https://github.com/kmaninis/OSVOS-PyTorch

Synthesis method	OSVOS, DAVIS-16 $\mathcal{J}\&\mathcal{F}$	RePRI, COCO^0 mIoU
Reference w/o synth. augm.	78.5	31.8
SemanticGAN	73.1	29.4
DatasetGAN	77.8	30.9
Projection	78.4	30.9
Input concat.	79.3	31.9
SemanticGAN D_m	79.5	32.3
MCA (ours)	**79.8**	**32.8**

Table 6.7: Impact on the performance of synthesized data produced with different models and mask supervision methods. The reference performance is obtained using standard data augmentation. Bold denotes the best performance.

Data selection	η	OSVOS, DAVIS-16 $\mathcal{J}\&\mathcal{F}$	RePRI, COCO^0 mIoU
Reference w/o augmentations		78.5 (+0.0) $\pm$0.3	31.2 (+0.0) $\pm$0.1
No data selection	-	78.7 (+0.2) $\pm$0.6	30.7 (-0.5) $\pm$0.5
Only SIFID-pool$_1$	15%	79.3 (+0.8) $\pm$0.5	32.2 (+1.0) $\pm$0.4
SIFID-{1,2,3,4} (ours)	5%	79.3 (+0.8) $\pm$0.6	31.9 (+0.7) $\pm$0.4
	10%	79.6 (+1.1) $\pm$0.4	<u>32.6 (+1.4)</u> $\pm$0.2
	15%	**79.8 (+1.3)** $\pm$0.3	**32.8 (+1.6)** $\pm$0.2
	25%	<u>79.7 (+1.2)</u> $\pm$0.3	32.3 (+1.1) $\pm$0.2
	50%	79.5 (+1.0) $\pm$0.3	32.0 (+0.9) $\pm$0.1

Table 6.8: Impact of synthetic data selection strategies on one-shot segmentation performance. Bold and underlined show the first and second best performance.

indicates that the high synthesis diversity and precise image-mask alignment (see Table 6.3) are the keys to achieve useful data augmentation.

Finally, Table 6.8 shows that it is important to filter out noisy samples before forming a pool of synthetic augmentations. For example, if generated samples are used without any filtering, the performance of segmentation networks can decrease. On the contrary, a simple strategy to filter out 15% of lowest-ranked generated images by SIFID, computed after the first pooling layer of the InceptionV3 network, helps to reduce the impact of bad-quality augmentation and, in effect, substantially improves the final segmentation performance. However, we observed that the SIFID metric is biased towards low-level image statistics, such as color and texture distributions, and is not indicative of the quality of generated images at higher scales. To account for the quality of generated images at different scales, we ranked synthesized examples by a joint ranking at different layers (denoted as SIFID-1,2,3,4), taking the average of all ranks. As seen in Table 6.8, filtering out noisy examples using this strategy helps to boost the performance of one-shot segmentation networks. Furthermore, we observed that it helps to significantly decrease the performance variance between different runs, which generally increased while using synthetic data augmentation in our experiments. Lastly, Table 6.8 demonstrates that the filtering rate should be neither too low nor too high: filtering out only 5% or 10% leaves some low quality images that are harmful for the data augmentation efficiency, while filtering too many samples (25%, 50%) decreases the diversity of the synthetic data pool and thus also diminishes its effectiveness.

6.4 Conclusion

We presented OSMIS, an unconditional GAN model that can learn to generate new high-quality image-mask pairs from a single training pair, not relying on any pre-training data. In such a low-data regime, our model generates photorealistic scenes that structurally differ from the original samples, while the produced masks are precisely aligned to the generated image content. Without requiring any extra data for pre-training, it can serve as a source of useful data augmentation for one-shot segmentation applications, providing complementary gains to standard image augmentation. Thus, we find using one-shot image-mask synthesis in practical applications promising for future research.

Although OSMIS demonstrates impressive diversity of synthesis, it is important to note that its capabilities remain inherently constrained by the appearance of objects in the original training samples. For instance, when presented with white cars as depicted in 6.2, OSMIS cannot be expected to generate cars of varying colors or models. This limitation of OSMIS is shared with our model SIV-GAN, introduced in Chapter 5, where it is similarly not expected to change colors of the cars present in Fig. 5.1. Another common limitation of SIV-GAN and OSMIS is that these models exhibit limited interpolation ability. For example, when provided with multiple images featuring cars of different colors and shape, our models usually do not generate novel cars with intermediate sizes and shapes that lie between the displayed examples. In the next chapter, we will introduce an approach that aims to address these limitations by utilizing pre-trained GAN models.

CHAPTER 7

Smoothness Similarity Regularization for Few-Shot GAN Adaptation

In Chapters 5 and 6, we introduced GAN models for diverse synthesis in extremely low data regimes, trained from scratch. This chapter explores an alternative approach to improve few-shot GAN synthesis by leveraging pre-training. This task, known as few-shot GAN adaptation, involves fine-tuning GAN models that were pre-trained on large, diverse datasets using a small few-shot dataset consisting of 1-10 images. The use of pre-trained models comes with both advantages and disadvantages. On one hand, models trained without pre-training are inherently limited in capacity due to the narrow scope of features present in the training data, making generalization and extrapolation challenging. This way, pre-training has the potential to enhance the quality, diversity, and generalization of models. On the other hand, pre-training typically works well only when the structures of the pre-training dataset are similar to the few-shot target dataset, which restricts its usage in rare image domains. In this chapter, we aim to mitigate this limitation. To this end, we propose a new smoothness similarity regularization for the generator that transfers the inherently learned smoothness of the pre-trained GAN to the few-shot target domain. The advantage of this regularization is that it works well even if the source and target domains are very different. The proposed approach is evaluated by adapting both unconditional and class-conditional GANs to diverse few-shot target domains. Our method significantly improves the quality and diversity of few-shot GAN synthesis in rare domains.

Individual Contribution

This chapter is based on the following publication (*Sushko et al.*, 2023a):

Smoothness Similarity Regularization for Few-Shot GAN Adaptation
Vadim Sushko, Ruyu Wang, Juergen Gall
IEEE International Conference on Computer Vision (ICCV), 2023. DOI: 10.1109/ICCV51070.2023.00651

This publication is the result of a collaborative effort between Vadim Sushko and Juergen Gall. Vadim Sushko proposed the initial idea of regularizing the fine-tuning a GAN with the smoothness of the source generator, which was further refined and extended through joint discussions with Juergen Gall. Ruyu Wang greatly assisted this project with helpful discussions and implementations of the class-conditional setting. All co-authors contributed to the writing of the paper. Vadim Sushko, as the first author, made substantial contributions to all stages of the project, including the initial idea, implementation, and paper writing.

Contents

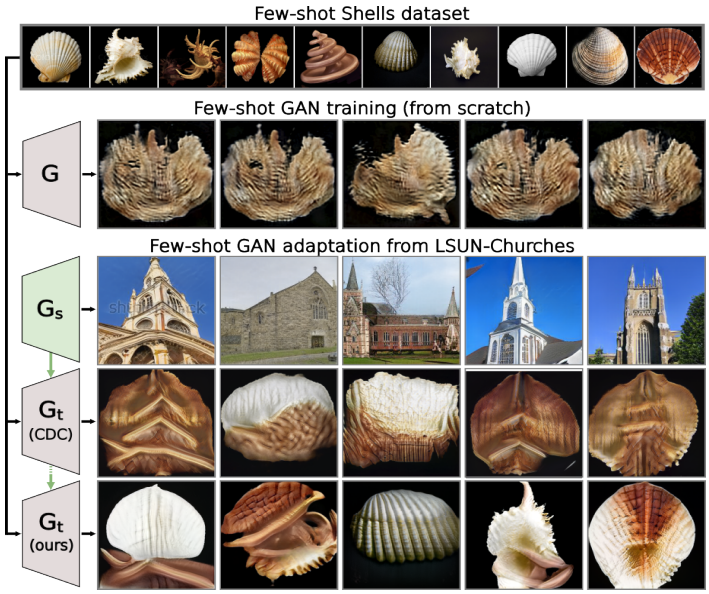

Figure 7.1: Training a GAN model G on a few-shot dataset (row 1) from scratch fails due to training instabilities (row 2). We thus aim to adapt a GAN G_s that has been pre-trained on a large dataset like LSUN-Church (row 3) to the target few-shot dataset (G_t). While fine-tuning (*Ojha et al.*, 2021) does not perform well either if source and target are dissimilar (row 4), our approach generates diverse and realistic images (row 5) by transferring the smoothness properties of G_s.

7.1 Introduction

Generative adversarial networks (GANs) have been shown powerful at various image synthesis tasks (*Choi et al.*, 2020; *Schönfeld et al.*, 2021; *Chan et al.*, 2022; *Karras et al.*, 2021; *Sauer et al.*, 2022, 2023). The success of these models is in large part enabled by the availability of large datasets for training, typically consisting of thousands of images. However, there are many applications and computer vision tasks such as one-shot or few-shot learning (*Boudiaf et al.*, 2021; *Tian et al.*, 2020b), out-of-distribution detection (*Ren et al.*, 2019), or long-tailed recognition tasks (*Gupta et al.*, 2019) where the number of available training images is very low.

Since training a GAN from scratch on very few samples does not perform well as shown in Fig. 7.1, a common strategy is to fine-tune a pre-trained GAN model on the few-shot dataset, typically employing additional regularization losses to penalize the degradation of the diversity (*Ojha et al.*, 2021; *Xiao et al.*, 2022). This approach, referred to as few-shot GAN adaptation, performs well when the target domain is structurally very similar to the dataset that has been used for pre-training, e.g., photographs vs. sketches of human faces. However, the performance drastically degrades in case of large dissimilarities between the

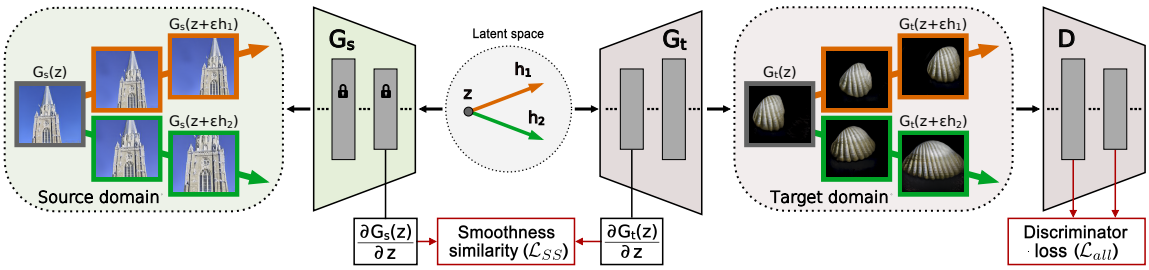

Figure 7.2: Given a pre-trained generator G_s, our smoothness similarity regularization preserves the learned smoothness of G_s while adapting it to a target domain with very few images. To mitigate overfitting to the target domain, the discriminator loss utilizes features at various layers and automatically adjusts the impact of different semantic scales to the similarity of source-target domains.

source and target domain as shown in Fig. 7.1. Such dissimilarities are a major bottleneck of using GANs in other disciplines like medicine, production, or crop science, where there is a lack of large datasets due to privacy, confidentiality, or simply lack of data. Motivated by this fact, we extend the protocol for few-shot GAN adaptation by investigating also pairs of datasets that are very different like churches and shells as shown in Fig. 7.1.

To improve few-shot GAN adaptation in the case of structurally dissimilar pairs, we propose a new GAN adaptation strategy. Firstly, we propose a new smoothness similarity regularization for the generator. Our key observation is that pre-trained GAN generators, regardless of the exact structure of objects in the pre-training dataset, learn well-structured and smooth latent spaces. For example, prior works demonstrated that various local shifts in the latent space can lead to interpretable and smooth transitions of output images, such as translation of objects in the scene or changing their size (*Voynov and Babenko*, 2020; *Härkönen et al.*, 2020; *Shen and Zhou*, 2021). As we show in our experiments, the proposed smoothness similarity regularization enables the transfer of this desirable property to other few-shot image domains without compromising the synthesis quality. Secondly, to overcome overfitting issues, we revisit the adversarial loss function of the discriminator and propose a simple yet efficient modification by computing the loss at different layers of the discriminator. This leads to the mitigation of overfitting and a more stabilized adaptation of the model to diverse target domains.

We evaluate our approach by adapting an unconditional (*Karras et al.*, 2020b) and a class-conditional GAN (*Brock et al.*, 2019) to diverse few-shot target domains. Our model significantly outperforms previous state-of-the-art methods in image quality and diversity in the challenging case of dissimilar source and target domains, while performing on par with the state of the art on structurally similar dataset pairs. In summary, our contributions are as follows: (i) We extend the evaluation protocol for few-shot GAN adaptation by including new dataset pairs that are structurally much less similar than was considered in prior work. (ii) We propose a new smoothness similarity regularization, which enables diverse synthesis in the target domain by transferring the learned smoothness of a pre-trained GAN. (iii) We revisit the adversarial loss function of the discriminator to stabilize few-shot GAN adaptation across diverse target domains. (iv) Our proposed model enables high-quality synthesis in the challenging case of dissimilar source and target domains, significantly outperforming prior methods. In addition, we show that our method can be applied to different classes of GAN architectures, including unconditional and class-conditional GAN models.

7.2 Method

In the task of few-shot GAN adaptation, we are given a small target dataset T and a pre-trained GAN model, consisting of a discriminator D and a generator G_s, which produces an image $x = G_s(z)$ from a continuous input variable z, such as a random noise vector or a continuous class embedding. The goal is to adapt the

generator to the target dataset such that it generates diverse and realistic images in the target domain as shown in Fig. 7.1. We denote the adapted target generator by G_t.

To achieve few-shot synthesis with a high image quality and diversity, our model should adhere to the following two properties. Firstly, the generator G_t should not only memorize and generate the target images, which will be addressed by the smoothness similarity regularization (Sec. 7.2.1). Secondly, the discriminator D must avoid overfitting to the few target images in order to provide useful supervision for G_t (Sec. 7.2.2). The overview of our method is shown in Fig. 7.2.

7.2.1 Smoothness Similarity Regularization for the Target Generator

In a low data regime like ours, G_t can easily overfit to the target dataset T and collapse to reproducing only the few modes represented in the training data. When walking in the latent space of such a generator, one would observe "staircase" patterns, where minor shifts in the latent space cause discontinuous transitions in the output image space (as shown in row 4 of Fig. 7.5). Naturally, to achieve a synthesis of high diversity, it is desirable for G_t to avoid such discontinuities, as having smoother image transitions allows to generate intermediate samples that can exhibit novel features. Therefore, in our model we aim to encourage G_t to produce smooth latent space interpolations, in which all the intermediate images are realistic.

Our approach is based on the observation that GANs trained on large datasets tend to have a well-structured latent space (*Voynov and Babenko*, 2020; *Härkönen et al.*, 2020; *Shen and Zhou*, 2021), in which different latent space directions can lead to smooth and interpretable image transitions. For example, in a generator pre-trained on a large dataset of churches, there can emerge latent directions causing smooth zooming or translation of churches (see Fig. 7.2). Our observation is that the nature of such image transitions (e.g., zooming or translation) is remarkably general. Thus, we propose a regularizer that utilizes this smoothness property of the source generator G_s as a cue while adapting it to another image domain, which can be very different from the domain that was used for pre-training. For example, as shown in Fig. 7.2, the same latent directions of churches can cause similar zooming or translation effects on shells.

Mathematically, the smoothness of the generator can be represented via a Jacobian matrix $J_{G^l}(z) = ||\partial G^l(z)/\partial z||$, quantifying how the generator's intermediate features after the l-th block change under local shifts in the latent space. As we want the same latent shift to cause perceptually similar image transitions in the source and target domains, we design a regularization term that brings the Jacobian matrices of G_s^l and G_t^l closer together. As the computation of full Jacobian matrices is expensive, we use an unbiased estimator of their products with a Gaussian vector (*Dauphin et al.*, 2015; *Karras et al.*, 2020b), which can be computed with standard back-propagation:

$$J_{G^l}^T(z) \cdot y = \mathbb{E}_{(y)\sim N(0,1)} \nabla_z \langle G^l(z), y \rangle, \tag{7.1}$$

where y is a Gaussian tensor of the same shape as G^l. Our smoothness similarity regularization is then expressed as:

$$\mathcal{L}_{SS} = \lambda_{SS} \cdot \mathbb{E}_{(z,y)\sim N(0,1)} ||\nabla_z \langle G_s^l(z), y \rangle - \nabla_z \langle G_t^l(z), y \rangle||_2, \tag{7.2}$$

where λ_{SS} steers the impact of the regularizer. As shown in Fig. 7.2, the smoothness similarity regularization depends on both generators, but only G_t is updated. It is interesting to note that the Jacobian matrix is also used for the path length regularization (*Karras et al.*, 2020b), which forces $J_G(z)$ to be orthogonal up to a global scale at any z. While this alternative regularizer also induces some form of smoothness, it does not transfer the inherently learned smoothness of a pre-trained GAN. In our experiments we show that it struggles to enforce the realism of intermediate images, as will be discussed in Sec. 7.3.4.

7.2.2 Revisiting the Adversarial Loss

To identify what kind of image transitions look realistic for the target domain, G_t requires strong supervision from the discriminator on image realism at different semantic scales. This includes the colors and textures of

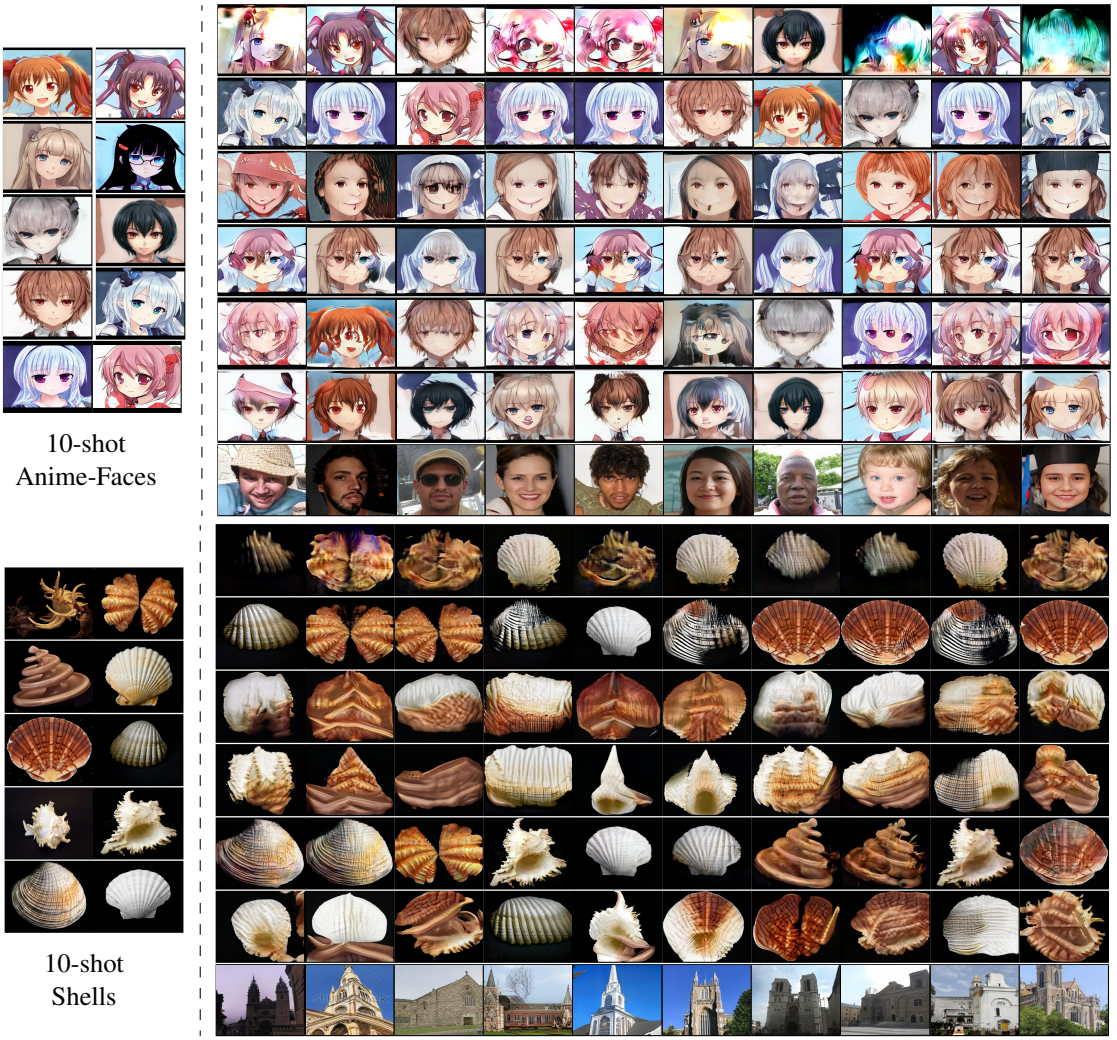

Figure 7.3: Visual comparison to prior methods on source-target dataset pairs with a dissimilar structure. In this challenging regime prior methods suffer from training instabilities, memorization issues, or inability to adapt the shapes of objects to the new domain. In contrast, our method generates realistic images that flexibly combine the features of different target images.

objects, as well as object shapes, especially if their distribution is different from the shapes of objects in the source domain. Learning the concept of image realism in low data regimes is, however, challenging due to the problem of overfitting.

Typically, a GAN discriminator consists of several consecutive blocks $\{D^i\}_{i=1}^N$ and for each given image x computes a real/fake logit after the last block $l = s^N \circ D^N(x)$, where s^N is a final processing layer such as a convolution. When adapting to a very small dataset, such a discriminator is prone to memorizing the training set (*Sushko et al.*, 2021b), leading to mode collapse and a synthesis of poor diversity (*Ojha et al.*, 2021). A possible solution (*Ojha et al.*, 2021; *Xiao et al.*, 2022) to overcome memorization is to use variants of the PatchGAN discriminator (*Isola et al.*, 2017), discarding the latest discriminator layers: $l = s^k \circ D^k(x), k < N$. This solution allows to adapt colors and textures of generated images to the target domain and still to avoid the memorization problem. However, it naturally has a limited capacity to learn more high-level semantic scene properties such as the shapes of objects, which we show in experiments.

In order to avoid memorization, and yet to balance the adaptation of colors, textures, and shapes of generated objects to a new domain, we hypothesize that a more flexible attention to different levels of image

realism is required by the discriminator. To this end, we perform a simple yet efficient modification to the loss function of the discriminator. Given a discriminator $\{D^i\}_{i=1}^N$ and its adversarial loss function $\mathcal{L}_\mathcal{D}(l)$ used for pre-training (e.g., cross-entropy or hinge loss), we design the discriminator to produce real/fake logits after *each* discriminator's block, and correspondingly compute the loss as the average across all blocks:

$$\mathcal{L}_{all}(x) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_\mathcal{D}[l^i(x)], \quad l^i(x) = s^i \circ D^i(x). \tag{7.3}$$

With the new objective, D is given more freedom to utilize the features extracted at different scales to compute the loss. Our finding is that D dynamically learns to use this freedom to identify the correct magnitude of each scale's loss contribution individually for each target domain, without explicit supervision. Consequently, we observe a strong overall stabilization effect on the adaptation performance across diverse source-target dataset pairs.

7.3 Experiments

To demonstrate that our approach for few-shot GAN adaptation can be applied to unconditional and class-conditional GANs, we selected for each category a popular GAN architecture: unconditional StyleGANv2 (*Karras et al.*, 2020b) and class-conditional BigGAN (*Brock et al.*, 2019). For both models, we test our approach on a variety of source-target domain pairs. For fair comparisons with prior works, most of our ablations and comparisons are conducted in the unconditional setting with StyleGANv2.

7.3.1 Experimental Setup

Datasets. In contrast to previous works that mostly considered pairs of similar datasets like *Face→Sketch* and *Face→Sunglasses*, we extend the protocol by including structurally dissimilar pairs of source and target domains, which is a more challenging task and is our primary interest. As source generators, we use StyleGANv2 checkpoints pre-trained on FFHQ (*Karras et al.*, 2019), LSUN-Church, and LSUN-Horse (*Yu et al.*, 2015). For the target datasets, we selected 10-shot subsets of various commonly used few-shot datasets, such as Anime-Face, Shells, or Pokemons (*Zhao et al.*, 2020a; *Liu et al.*, 2021).

Training details. We fine-tune StyleGANv2 using the $\mathcal{L}_{SS}$ and $\mathcal{L}_{all}$ loss terms as presented in Sec. 7.2. For the smoothness similarity regularization, we use the intermediate features G^l at resolution (32×32) and set $\lambda_{SS} = 5.0$. We follow *Ojha et al.* (2021) in choosing all the other hyperparameters, such as image resolution (256×256), learning rates, and batch size. Our experiments across all datasets use the same model configuration and set of hyperparameters.

Baselines. We compare our method to most recent few-shot GAN adaptation approaches: TGAN (*Wang et al.*, 2018b), FreezeD (*Mo et al.*, 2020), CDC (*Ojha et al.*, 2021), RSSA (*Xiao et al.*, 2022), and AdAM (*Yunqing et al.*, 2022). In addition, we compare our proposed smoothness similarity regularizer $\mathcal{L}_{SS}$ to other regularization techniques: path length regularization (PPL) (*Karras et al.*, 2020b) and MixDL (*Kong et al.*, 2022).

Evaluation. We assess the diversity and the quality of the generated images in the target domains. Following *Ojha et al.* (2021), we evaluate diversity with the intra-LPIPS, a measure that clusters generated images around the nearest training samples and computes the average LPIPS (*Zhang et al.*, 2018b) of all the clusters. To measure the image quality, we use FID (*Heusel et al.*, 2017) computed between a held-out validation set and a generated set of the same size. We train all models for 30k epochs in case of dissimilar source-target domains and for 5k on close domain pairs, evaluating metrics every 1k epochs. The final checkpoints in all experiments are selected by best FID.

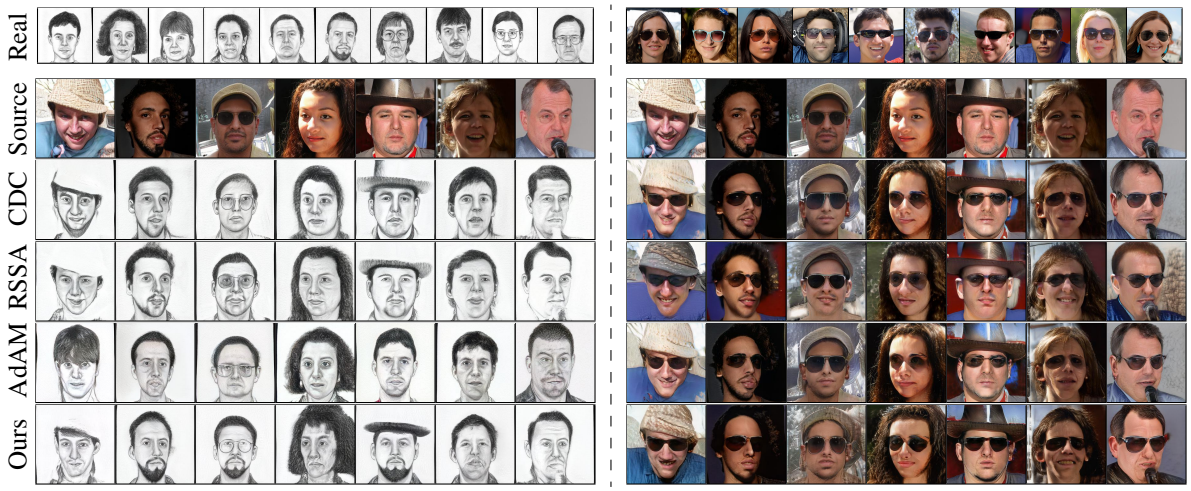

Figure 7.4: Visual comparison to most recent prior methods on *Face→Sketch* and *Church→Sunglasses*, the dataset pairs depicting similar image domains. In this regime, our method performs on par with previous state of the art. See Table 7.2 for a quantitative comparison.

Method	Face→Anime		Church→Shells		Horse→Pokemons	
	FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑
TGAN	153.2	0.29	205.3	0.22	115.0	0.52
FreezeD	112.4	0.22	180.8	0.27	123.3	0.49
CDC	140.2	0.50	187.9	0.48	109.5	0.55
RSSA	133.2	0.37	182.4	0.44	117.3	0.54
AdAM	116.4	0.42	152.4	0.28	106.5	0.55
Ours	**97.3**	**0.57**	**140.5**	**0.53**	**84.1**	**0.61**

Table 7.1: Comparison of the adaptation performance in case of dissimilar source-target domains. Bold denotes best performance.

7.3.2 Results With Dissimilar Source-Target Domains

We first present our results on the source-target domain pairs with dissimilar structure: *Face→Anime*, *Church→Shells*, and *Horse→Pokemon* (see Fig. 7.3). Our general observation from Fig. 7.3 is that in this challenging regime prior methods suffer either from training instabilities, memorization issues, or inability to adapt the shape of objects to the new domain. For example, for *Face→Anime*, despite an apparent correspondence between the two domains, none of the prior methods successfully transfers the distribution of head poses to the anime style, e.g., overfitting too strongly to the 10 provided samples (FreezeD), failing to adapt the shape of faces to the style of anime (CDC), or not generating high-quality anime-faces due to instabilities (TGAN, RSSA, AdAM). Similarly, for *Church→Shells* we observe that prior methods produce only copies of the example shells (FreezeD, AdAM), generate shells of unrealistic church-like shapes (CDC, RSSA), or suffer from instabilities (TGAN). In contrast, our method achieves high-quality synthesis, in which the generated images (i) look like realistic anime-faces and shells; (ii) flexibly combine features observed in different target images (e.g., anime hair color can be combined with various eye colors or background styles); and (iii) meaningfully transfer the variation of images from the source domain (e.g., generated shells adjust to the positions and shapes of churches).

Method	Face→Sketch		Face→Sunglasses	
	FID↓	LPIPS↑	FID↓	LPIPS↑
TGAN	54.2	0.38	36.8	0.56
FreezeD	48.8	0.32	32.0	0.59
CDC	54.2	0.40	30.5	0.59
RSSA	61.4	**0.45**	36.3	0.58
AdAM	56.3	0.37	31.1	**0.60**
Ours	**45.2**	0.44	**27.5**	**0.60**

Table 7.2: Comparison in case of close source-target domains. Bold denotes best performance.

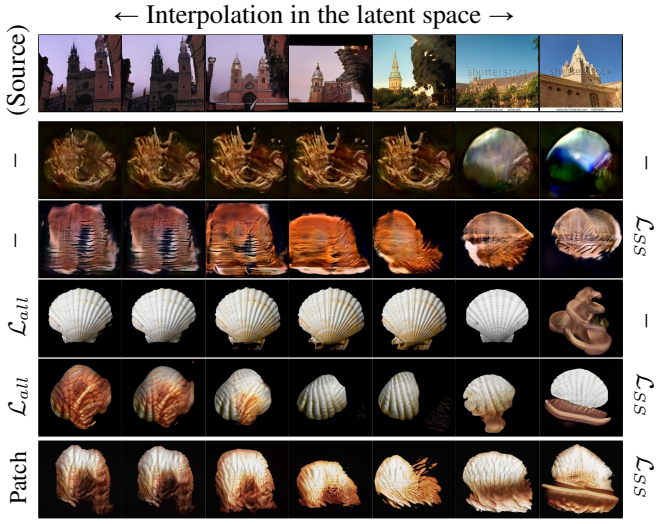

Figure 7.5: Latent space interpolations of the source generator and the ablation models from Table 7.3. Leftmost and rightmost columns show the used D loss and G smoothness regularization.

The quantitative comparison in Table 7.1 confirms our analysis, where our method achieves the best quality and diversity scores across all datasets. We note a high average relative improvement of more than 18% and 11% in FID and LPIPS compared to the highest scores achieved by prior methods. Overall, we conclude that our method significantly improves over prior works on few-shot GAN adaptation with dissimilar source and target domains.

7.3.3 Results With Close Source-Target Domains

Next, we follow the evaluation of prior works and compare the models on similar source and target domains, such as adaptation of human faces to a different style. The visual results for *Face→Sketch* and *Face→Sunglasses* are shown in Fig. 7.4. Our method successfully performs the few-shot adaptation in this setting, adapting the colors and textures of faces to the gray-scale sketch domain, or adding a novel attribute (sunglasses). We note that our method is not explicitly designed to transfer all the details of a face from the source domain, thus changes in the generated images like facial hair are expected. Yet, we observe that our method generally does not lose distinctive features of faces in source images, performing on par with previous state-of-the-art methods. The quantitative comparison is provided in Table 7.2: on both datasets our method achieves the best FID scores and performs on par with the best performer in LPIPS.

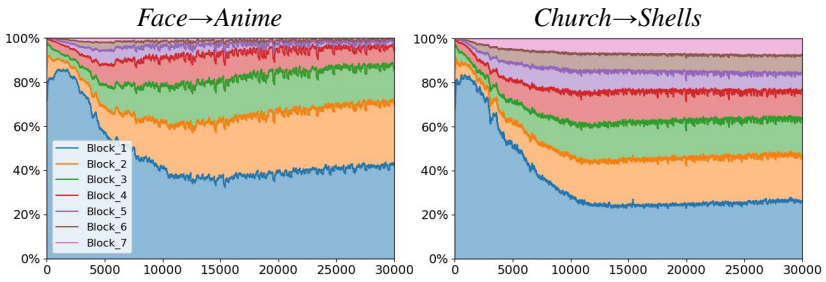

Figure 7.6: The contribution of features at different D blocks to the adversarial loss function $\mathcal{L}_{all}$. For two closer image domains (the left plot), the network concentrates mostly on earlier layers to compute the loss, while for less similar domains the network learns to use the later layers representing more high-level D features.

D loss	Smooth reg. for G	**Face→Anime** FID↓	LPIPS↑	**Church→Shells** FID↓	LPIPS↑
StyleGANv2	-	178.0	0.21	243.8	0.17
StyleGANv2	SS (ours)	180.7	**0.61**	252.8	**0.62**
PatchGAN	-	145.2	0.37	183.1	0.31
PatchGAN	SS (ours)	132.2	0.55	184.2	0.56
$\mathcal{L}_{all}$ (ours)	-	116.4	0.36	175.4	0.43
$\mathcal{L}_{all}$ (ours)	SS (ours)	**97.3**	0.57	**140.5**	0.53

Table 7.3: Impact of $\mathcal{L}_{all}$ and $\mathcal{L}_{SS}$. Bold denotes best performance.

7.3.4 Ablations

We demonstrate the importance of our proposed loss terms in Fig. 7.5, which shows latent space interpolations of trained models and their similarity to the pre-trained source model G_s (row 1). Firstly, we note that the plain StyleGANv2 model (row 2) suffers from instabilities in our low data regime, achieving poor image quality and diversity and having "staircase"-like latent space interpolations. Applying $\mathcal{L}_{SS}$ without $\mathcal{L}_{all}$ (row 3) helps to achieve diverse synthesis with smooth interpolations, but is not enough to achieve good image quality. On the other hand, using $\mathcal{L}_{all}$ (row 4) helps to overcome instabilities and improve image quality, but it cannot maintain smooth interpolations and high diversity. Finally, our full model (row 5) allows a higher-quality, diverse synthesis with smooth and realistic latent space interpolations. Note how the image transitions mimic the behaviour of the source model (churches and shells change shapes and positions similarly), allowing to achieve diverse and realistic synthesis.

The effect of $\mathcal{L}_{all}$ is further demonstrated in Fig. 7.6, where we show the contribution of different D blocks to the adversarial loss at different epochs. We note the ability of the discriminator to identify correct loss contributions adaptively for different source-target domain pairs. For example for *Face→Anime*, the network concentrates mostly on the earliest D blocks to adapt the colors and textures of faces to a new style. In contrast, for the more distant domains *Church→Shells*, the network learns to attribute a higher weight to the later blocks to also adapt higher-level features, such as shapes of objects. In effect, we observe a stabilized adaptation of colors, textures, and shapes of objects across diverse source-target pairs. Using PatchGAN (*Ojha et al.*, 2021) as discriminator loss does not achieve such a balance as it focuses mostly on lower-scale features (row 6 in Fig. 7.5).

Our observations are confirmed by the quantitative study in Table 7.3: without $\mathcal{L}_{SS}$ the model does not achieve high diversity (high LPIPS), while $\mathcal{L}_{all}$ is necessary for high image quality (low FID). We conclude

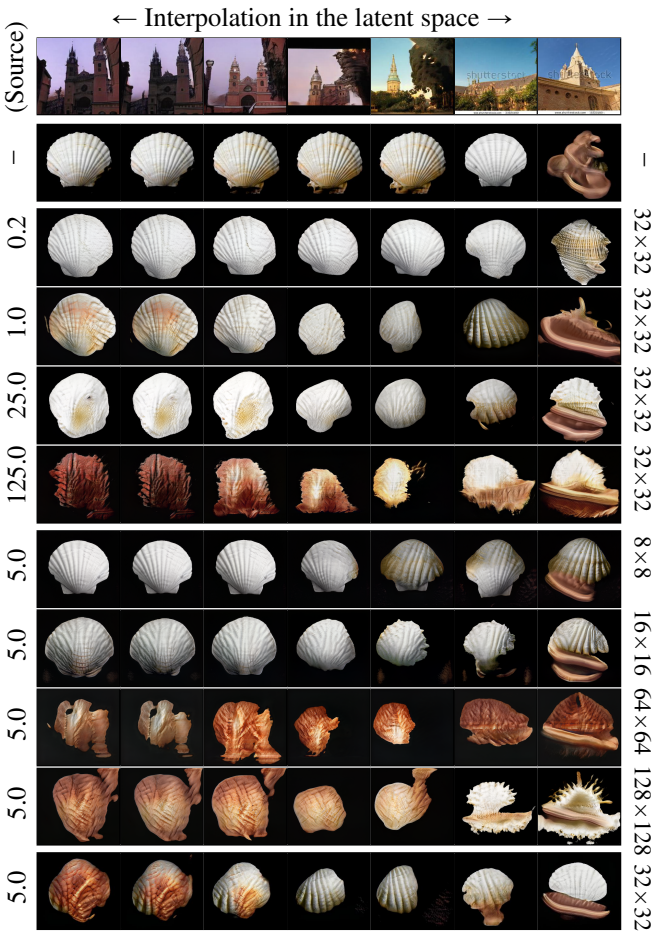

Figure 7.7: Latent space interpolations of the source generator and the ablation models from Table 7.4. Leftmost and rightmost columns show the used λ_{SS} and the resolution of G^l.

that both our proposed loss terms are important to achieve high-quality synthesis.

An ablation on the parameters of $\mathcal{L}_{SS}$ (λ_{SS} and the resolution of features G^l)) is provided in Fig. 7.7 and Table 7.4. Firstly, we observe the effect of λ_{SS} (rows 3-6 in Fig. 7.7 and rows 2-5 in Table 7.4). As seen from the ablation study, compared to the model without any regularization, our smoothness similarity regularization helps to overcome memorization and achieve diverse synthesis. The effect of $\mathcal{L}_{SS}$ is, as expected, higher when λ_{SS} is increased, which is indicated by increasing LPIPS scores. Yet, we find that setting a high λ_{SS} starts to compromise the image quality, as the loss starts to overtake the adversarial loss supervision. We found that $\lambda = 5.0$ consistently achieves a good trade-off between image quality and diversity across many source-target domains. Furthermore, we observe the effect of using features at different resolutions, corresponding to different generator blocks (rows 7-10 in Fig. 7.7 and rows 6-9 in Table 7.4). We find that using later generator blocks at higher resolution increases the impact of the regularization. However, we also observe that using a very high resolution leads to the transfer of image transitions from the source domain at more fine-grained level, which can compromise image quality, for example transferring minor details that do not look realistic in the target domain. Based on the results in Table 7.4, we concluded that the resolution (32×32) provides a good quality-diversity trade-off as it transfers high-level, more interpretable image variations without compromising the high-level coherency in the target domain.

Lastly, Table 7.5 provides a comparison of our proposed $\mathcal{L}_{SS}$ loss term to other regularizers: path length regularization (PPL) (*Karras et al.*, 2019) and MixDL (*Kong et al.*, 2022). While all regularizers help to

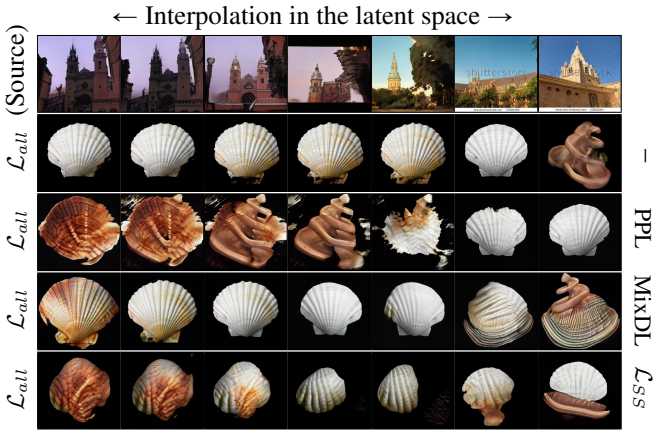

Figure 7.8: Latent space interpolations of the source generator and the ablation models from Table 7.5. Leftmost and rightmost columns show the used D loss and G smoothness regularization.

λ_{SS}	Res. of G^l	**Face→Anime** FID↓	LPIPS↑	**Church→Shells** FID↓	LPIPS↑
-	-	116.4	0.36	175.4	0.43
0.2	32×32	110.0	0.41	160.2	0.44
1.0	32×32	**96.4**	0.51	144.5	0.50
25.0	32×32	105.2	0.58	171.0	0.55
125.0	32×32	131.3	**0.64**	188.5	**0.57**
5.0	8×8	104.1	0.44	156.6	0.45
5.0	16×16	101.4	0.55	150.2	0.48
5.0	64×64	114.7	0.59	165.5	0.54
5.0	128×128	128.2	0.60	182.2	**0.57**
5.0	32×32	97.3	0.57	**140.5**	0.53

Table 7.4: Ablation on the λ_{SS} and the resolution of G^l for the smoothness similarity regularization.

achieve smoother latent space interpolations and thus improve the quality and diversity metrics, our smoothness similarity regularization enables the highest performance in both FID and LPIPS. While our approach transfers the learned smoothness of the source generator to the target domain, PPL and MixDL resort to gradually interpolating between the provided training samples, which leads to latent space interpolations that either look unrealistic or lack diversity (rows 7-8 in Fig. 7.5). This demonstrates that transferring smoothness from a pre-trained generator is beneficial to enforce image transitions that are realistic and diverse.

7.3.5 Experiments in the 1-shot and 5-shot settings and comparison to SIV-GAN

In addition to the above experiments in the 10-shot regime, it is interesting to study the behaviour of the models in more extreme data settings, such as the tasks of 1-shot and 5-shot image generation. Consistent with the main goal of this Chapter, our main focus is on the challenging case of rare target domains that do not have a structurally similar pre-training dataset. The results for 1-shot and 5-shot image generation setups are presented in Fig. 7.9. In this setting, we compare our proposed method to CDC (*Ojha et al.*, 2021), a popular baseline from the literature. Our observations from Fig. 7.9 are consistent with Sec. 7.3.2: while the

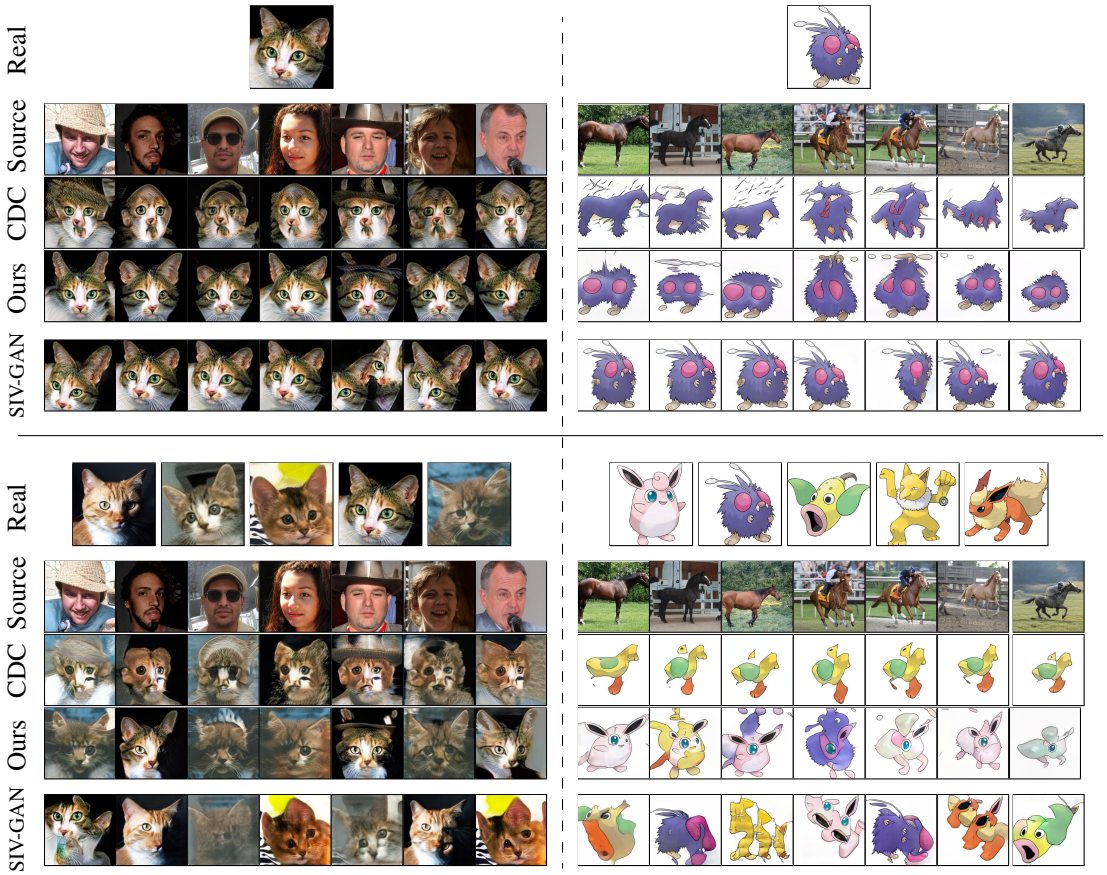

Figure 7.9: 1-shot and 5-shot image generation results using two adaptation techniques (CDC (*Ojha et al.*, 2021) and ours) and SIV-GAN training from scratch.

prior method CDC cannot learn the shapes of objects in the new domain, our method achieves more realistic synthesis, successfully transferring meaningful high-level image variations even from structurally dissimilar datasets.

It is interesting to note that our model SIV-GAN from Chapter 5 can be also applied to such extremely low data regimes as learning from 1 or 5 images. The results of SIV-GAN are therefore also included to Fig. 7.9. The difference of SIV-GAN from models in this Chapter is that it is trained from scratch, without using any pre-training. As seen from the figure, SIV-GAN also overcomes memorization and generates new images with noticeable diversity. Yet, in comparison to few-shot GAN adaptation, SIV-GAN demonstrates more copying of objects' shapes and, in case of 5-shot synthesis, rarely generates objects that combine features from multiple training images. This makes the usage of pre-training preferable in practical scenarios requiring image synthesis with higher diversity.

7.3.6 Adaptation of Class-Conditional GAN

Our approach is not limited to unconditional GANs, but it can also be applied to a class-conditional GAN model. We selected BigGAN (*Brock et al.*, 2019) for our experiments as it is a popular backbone architecture for class-conditional synthesis on ImageNet (*Deng et al.*, 2009). We make two modifications to enable the adaptation of the model to unconditional few-shot datasets. Firstly, we remove the conditioning of the discriminator via the projection layer (*Miyato and Koyama*, 2018). Secondly, we treat the generator's learned continuous class embedding as part of the latent space, thus sampling a Gaussian vector in the joint noise-class

D loss	Smooth reg. for G	Face→Anime		Church→Shells	
		FID↓	LPIPS↑	FID↓	LPIPS↑
$\mathcal{L}_{all}$ (ours)	-	116.4	0.36	175.4	0.43
$\mathcal{L}_{all}$ (ours)	PPL	107.8	0.46	179.4	0.44
$\mathcal{L}_{all}$ (ours)	MixDL	105.9	0.50	150.4	0.51
$\mathcal{L}_{all}$ (ours)	SS (ours)	**97.3**	**0.57**	**140.5**	**0.53**

Table 7.5: Comparison of smoothness similarity regularization $\mathcal{L}_{SS}$ with other regularizers. Bold denotes best performance.

D loss	Smooth reg. for G	ImageNet→Flowers		ImageNet→Pokemons	
		FID↓	LPIPS↑	FID↓	LPIPS↑
BigGAN	-	213.3	0.29	226.8	0.15
BigGAN	SS (ours)	225.6	0.47	208.3	0.47
$\mathcal{L}_{all}$ (ours)	-	123.9	0.28	129.4	0.27
$\mathcal{L}_{all}$ (ours)	SS (ours)	**106.4**	**0.55**	**89.6**	**0.56**

Table 7.6: Ablation on the performance when adapting the class-conditional BigGAN model (*Brock et al.*, 2019) pre-trained on ImageNet.

space at each fine-tuning epoch. This way, the generator produces an image based on a single input vector in an unconditional fashion. We then fine-tune the pre-trained model using our loss terms $\mathcal{L}_{SS}$ and $\mathcal{L}_{all}$ as presented in Sec. 7.2. We use image resolution 256×256 and batch size of 32. The hyperparameters for $\mathcal{L}_{SS}$ are the same as for StyleGANv2: G^l features at resolution (32×32) and $\lambda_{SS} = 5.0$. We train for 30k epochs and select checkpoints by best FID.

Datasets. As the source generator, we use the BigGAN checkpoint pre-trained on class-conditional ImageNet at resolution 256×256. We demonstrate 10-shot adaptation results with two commonly used few-shot generation datasets: Oxford-Flowers (*Nilsback and Zisserman*, 2006) and Pokemons (*Liu et al.*, 2021). We use the same model configuration for both datasets.

Results. Fig. 7.10 demonstrates latent space interpolations of the source and target generators. We note that a simple fine-tuning of BigGAN suffers from training instabilities and mode collapse. In contrast, our method successfully adapts BigGAN to generate diverse images in the target domains. We highlight that our method transfers smooth and realistic image transitions from the well-learned BigGAN's noise-class space, despite significant dissimilarities between ImageNet and the few-shot datasets, in particular Pokemons. For example, it can be noticed how the latent space interpolations in the target domains mimic the source domain, e.g., the generated flowers and pokemons change their position and size similarly to dogs and wolves (5^{th}-10^{th} columns in Fig. 7.10) or stretch their shape to mimic the proportions of busses (11^{th}-14^{th} columns).

Table 7.6 shows the importance of our proposed loss terms. Our observations are consistent with the ablations with StylGANv2: $\mathcal{L}_{all}$ is necessary to avoid instabilities and achieve a good image quality (low FID), while $\mathcal{L}_{SS}$ is required to achieve smooth latent space interpolations and good diversity (high LPIPS). We conclude that our method successfully extends to the adaptation of class-conditional models, where target domains benefit from the rich noise-class space learned on a multi-class dataset such as ImageNet.

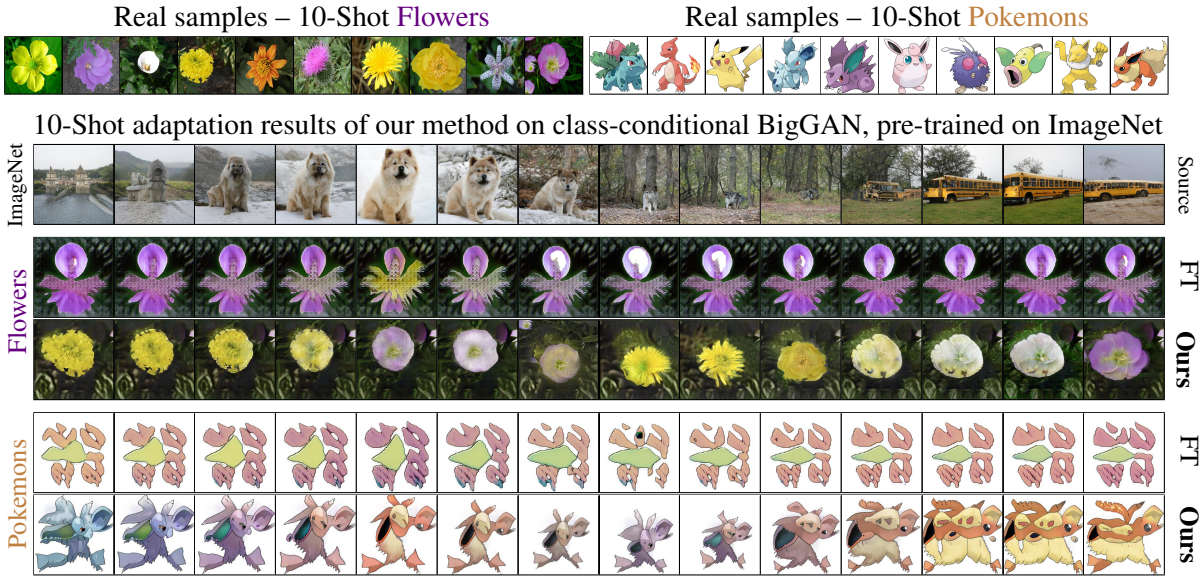

Figure 7.10: Results for the class-conditional BigGAN (*Brock et al.*, 2019) pre-trained on ImageNet. While simple fine-tuning (FT) suffers from training instabilities and mode collapse, our method helps to achieve much higher image quality and diversity, transferring smooth and realistic image transitions from the source domain, e.g., objects smoothly changing their locations, size, and shape.

7.4 Conclusion

In this work, we presented a new method for few-shot adaptation of GAN models. It transfers the smooth latent space of a pre-trained GAN, which was trained on a large dataset, to a new domain with very few images. We addressed the case of few-shot GAN adaptation when the source and target domains are structurally dissimilar, which is a common issue in applications. Our extensive results demonstrate that in this setting our approach outperforms previous works in terms of image quality and diversity. These results show the potential of few-shot GAN synthesis for applications in other disciplines where there is a lack of large datasets due to privacy, confidentiality, or simply lack of data.

Although our models presented in this chapter are significantly less reliant on the structure of the pre-training datasets compared to previous approaches, it is important to acknowledge that our method may still have inherent limitations. One such limitation arises from the assumption that the knowledge gained from large datasets of natural images regarding image transitions can be effectively applied to target domains. However, this assumption can be violated if common image variations like zooming or translation cannot be readily applied to objects within the target domain. In such cases, our models SIV-GAN and OSMIS, which do not rely on any pre-training, may be preferred. Nevertheless, our experiments have demonstrated remarkably high levels of generalization even for distinct image domains (e.g., shells on black background) for our new method. Compared to SIV-GAN and OSMIS, it enables the generation of images with significantly greater diversity and, notably, the ability to interpolate between different images, resulting in objects with novel shapes and intermediate sizes that were not observed in the training set. This enhanced synthesis capability can become of great value across a wide range of image domains and applications.

CHAPTER 8

Conclusion

Contents

8.1 Overview

In this thesis, we have addressed several image generation tasks with generative adversarial networks (GANs). These tasks can be categorized into two areas: semantic image synthesis and unconditional image synthesis in low data regimes. In semantic image synthesis, the goal is to produce realistic and diverse images that adhere to provided semantic label maps. Historically, to achieve good performance in this task, GAN models relied on the perceptual VGG loss, which was used to train the generator in addition to the adversarial discriminator's loss. As we demonstrated in Chapter 4, this loss limited the progress of semantic image synthesis GANs, as it biased the color and texture distributions of generated images and severely constrained diversity. To mitigate these issues, in Chapter 4, we have introduced a new OASIS model that needs only adversarial supervision for successful training. Our model outperforms previous models in terms of both image quality and diversity while being more lightweight and offering new capabilities.

The subsequent three chapters were concerned with different regimes for limited-data unconditional GAN training. Potentially, applications dealing with restricted image domains are the ones that require synthetic images the most, as there is the highest potential for synthetic data augmentation to be useful. However, conventional GAN models did not perform well with insufficient training data due to training instabilities, mode collapse, and memorization issues. In Chapter 5, we have addressed these problems by introducing the SIV-GAN model, which enables diverse synthesis of realistic images even in very low data regimes. Notably, it excels in a new training regime: few-shot datasets comprising very similar images. In contrast, as was demonstrated in our experiments, previous single-image and few-shot GAN methods struggled in this challenging regime. In Chapter 6, we have extended the impressive capabilities of SIV-GAN to the joint synthesis of images and segmentation masks. While previous single-image GAN methods failed to generate segmentation masks for synthesized images, and image-mask GANs encountered memorization and training instability issues when trained on a single image-mask pair, our proposed model, OSMIS, enables the generation of diverse and high-quality paired image-mask data from a single training example. These results facilitate the successful

use of generated samples as synthetic data augmentation in downstream one-shot segmentation applications. Lastly, in Chapter 7, we have explored the task of few-shot GAN adaptation, which leverages pre-training knowledge to enhance the quality and diversity of few-shot image synthesis. As demonstrated in Chapter 7, conventional adaptation methods struggled when the source and target domains exhibited similar structures, severely limiting their applicability in rare image domains. To address this limitation, we have developed a new regularization loss that transfers the smoothness of a pre-trained GAN generator to the target domain. Our method performs well even when the source and target domains are not closely similar, outperforming prior methods by a significant margin in such cases.

8.2 Discussion of Contributions

The overall goal of this thesis was to improve the performance of GANs in various image generation tasks, expanding their potential for new applications, datasets, and different data types. The main contributions of this thesis are summarized and reviewed next.

8.2.1 Semantic Image Synthesis with Only Adversarial Supervision

As discussed in Sec. 2.2, to achieve good performance in semantic image synthesis, GAN models require optimal conditioning mechanisms for both the generator and discriminator. The networks need to be effectively conditioned: on label maps, for the discriminator, and on both noise and label maps for the generator. While the conditioning of the generator on label maps has received considerable attention in the literature, leading to the development of mechanisms like spatially-adaptive normalization layers (SPADE) (*Park et al.*, 2019b), other forms of conditioning have been largely underexplored. Conventionally, noise has been used simply as input to the first layer of the generator, while the discriminator naively concatenated the label maps with the input images. In Chapter 4, we demonstrated that these approaches are suboptimal. We introduced a segmentation-based discriminator that uses the provided semantic label maps as targets for the loss function, rather than as input. This discriminator provides stronger and more detailed pixel-level feedback, which is further enhanced through the introduction of LabelMix regularization. By leveraging the improved training signal from the discriminator, we eliminate the need for additional losses such as the VGG perceptual loss. Consequently, this enabled us to achieve the synthesis with color and texture distributions that are closer to real data. In addition, we introduced a novel 3D noise injection scheme that spatially modulates noise at different layers of the generator. This scheme eliminates the need for image encoders to achieve multi-modality, enables both global and local image resampling, and significantly improves the overall diversity of synthesis. Moreover, as demonstrated in Sec. 4.3.3, the proposed architectural improvements, make it possible to train a semantic image synthesis GAN even on datasets with severe class imbalance. Our model, OASIS, achieves high-quality and diverse synthesis on the LVIS dataset, which consists of over 1000 classes, the majority of which are underrepresented. Unlike the baseline SPADE, OASIS also effectively handles sparsely annotated label maps and avoids mode collapse in such scenarios.

Overall, with our proposed OASIS model, we mitigate the issues of prior GAN models for semantic image synthesis outlined in Sec. 1.2.1: overreliance on the perceptual loss, insensitivity to noise, and learning from imbalanced datasets.

8.2.2 Diverse Unconditional Synthesis in Extremely Low Data Regimes

GAN models are notorious for their instability, frequently displaying collapsed training due to escalated gradients. This behaviour is more prevalent when dealing with very limited data, as the discriminator is more prone to overfitting to the limited training data. In fact, training GAN models with only a single image or a few very similar images has proven to be challenging, as these models frequently suffered from mode collapse and

training instabilities. These issues significantly limited the application of GANs in domains where collecting data remains a challenge.

In Chapter 5, we introduced a new two-branch SIV-GAN discriminator architecture that effectively mitigates the memorization of the whole training data in extremely low data regimes. The core idea of this discriminator is to disentangle the learning of the appearance of objects from their spatial arrangements, through two distinct branches. As a result, the discriminator avoids rapid overfitting to the entire training data and encourages the generator to produce novel scene compositions with realistic content and layouts. On the generator side, we introduced a new diversity regularization technique specifically designed for extremely limited data scenarios. This regularization maximizes the difference in generator's outputs at different layers for different noise inputs. In effect, our SIV-GAN model overcomes training instabilities and memorization problems when provided with just one or a few similar images. Our model also reduces the limitation of prior single-image GAN models that suffered from distorting object appearances and struggled to learn from multiple images.

8.2.3 One-Shot Image-Mask Synthesis

Diverse image generation from very small datasets has great potential to be used as synthetic data augmentation in applications. However, previous GAN models struggled to produce sufficiently high-quality and diverse images for synthetic data augmentation to be successful, mainly for two reasons. Firstly, as discussed previously, GANs were generally not successful in low data regimes, commonly suffering from training instabilities and mode collapse. In addition to this problem, computer vision applications often require annotations in addition to new images, which are non-trivial to obtain with GANs without using expensive and non-automatic labelling procedures.

In Chapter 6, we mitigated the above issues by extending SIV-GAN to joint synthesis of images and segmentation masks. Our new model, called OSMIS, produces new diverse image-mask pairs when given only a single image-mask pair for training. This is achieved via the introduced masked content attention mechanism, which allowed the discriminator to compare the content of different objects in the real and fake images separately from each other, thereby encouraging the generator to produce accurate segmentation masks. In contrast to previous image-mask GANs, this solution is remarkably inexpensive as it does not require expensive labelling procedures, such as manual annotation or external pre-trained segmentation networks. In effect, OSMIS became an effective tool for producing useful data augmentation, as demonstrated in Chapter 6 with the example of one-shot segmentation applications.

It is interesting to note that semantic image synthesis models, like OASIS, can be also be applied to segmentation applications. For example, in Chapter 4, we demonstrated how OASIS improves the performance of a segmentation network on rare classes in case of large-scale datasets. However, OASIS and SIV-GAN are tailored for very different data regimes and are thus unlikely to have the same use cases. At inference stage, OASIS requires semantic label maps, which suggests its usage in domains where annotations are easily available (e.g., self-driving datasets like Cityscapes). On the other hand, OSMIS creates images and masks with new semantic layouts without any conditioning, making it better suited for restricted domains where new segmentation masks along with new images should be produced (such as the generation of defects on manufacturing details in new locations).

8.2.4 Few-Shot GAN Adaptation with Dissimilar Source-Target Domains

Transfer learning is a widely used technique in machine learning that improves the performance of neural networks when trained on limited data. However, prior to our work, transfer learning in the context of GANs had only been successful between highly similar image domains, where the object shapes in the source and target domains were very similar. In contrast, for rare image domains where a large pre-training dataset with a similar structure is not available, existing methods for few-shot GAN adaptation failed to achieve satisfactory performance.

In Chapter 7, we addressed this issue by introducing a novel few-shot GAN adaptation method specifically designed for diverse pairs of source and target domains. Our approach involves fine-tuning a GAN model using a few-shot dataset while preserving the smoothness properties of the original pre-trained generator. By doing so, our method effectively mitigates the problem of memorization while maintaining the realism of the generated images through the transfer of smooth and realistic latent space interpolations from the pre-trained GAN. Additionally, we revisited the adversarial loss of the discriminator and enabled the computation of the loss across different layers. This change not only helps to mitigate overfitting but also allows the discriminator to automatically identify the appropriate contributions for a given source-target pair of domains. As a result, our new model significantly enhances the quality and diversity of synthesized images compared to prior GAN adaptation methods, opening up possibilities for applying GANs in new few-shot image domains.

Remarkably, the approach of transferring the smoothness of a pre-trained GAN extends even to 1-shot image domains, which was also addressed by our SIV-GAN model in Chapter 5. Interestingly, our adaptation approach in Chapter 7 outperformed SIV-GAN in terms of synthesis diversity, addressing the difficulty of generating new objects combining features from different training images. It is therefore preferable to make use of pre-trained GANs over training from scratch in applications that benefit from increased synthesis diversity.

8.3 Outlook and Future Perspectives

In the subsequent sections, we will outline limitations and potential areas for improvement in the two primary research areas of this thesis: GANs for semantic image synthesis and unconditional GAN training in low data regimes. Additionally, we will provide a broader outlook on the development of generative adversarial networks and explore alternative directions for image synthesis models beyond GANs.

8.3.1 GANs for Semantic Image Synthesis

Semantic image synthesis is the focus of the Chapter 4. In this section, we discuss possible research directions to improve the performance and extend the capabilities of semantic image synthesis GANs.

Alternative segmentation losses. In Chapter 4, we demonstrated that a segmentation-based discriminator is a reasonable choice for semantic image synthesis GANs. Our OASIS uses a simple balanced multi-class cross entropy loss, while the alternative segmentation losses were left unexplored. In recent years, the literature on semantic segmentation has seen higher-performing losses, including the focal loss (*Lin et al.*, 2017), lovasz-softmax loss (*Berman et al.*, 2018), region mutual information loss (*Zhao et al.*, 2019), or poly loss (*Leng et al.*, 2022). Improved ways to train a segmentation-based discriminator can lead to clearer feedback for the generator, having a potential to speed up the training and improve the final synthesis performance.

Semi-supervised learning. In addition to exploring alternative segmentation losses, the discriminator of OASIS can benefit from the advancements in the field of semi-supervised semantic segmentation. Semi-supervised segmentation often involves incorporating large collections of unlabeled images in addition to the standard training datasets. While these unlabeled images may not directly contribute to the learning of semantic classes, they can play a role in improving the discriminator's understanding of image realism and reducing overfitting. Consequently, by incorporating semi-supervised learning techniques, we can anticipate a significant improvement in the generation of more realistic images.

GAN inversion for image editing. In Sec. 4.3.4, we demonstrated that the OASIS discriminator can be used to predict the semantic label map of a given unlabelled image, and then to re-synthesize it in many other styles while preserving its global layout. However, preserving also the style of a given image remains not straightforward. One possible approach for capturing the style of an image is to use an image encoder, as done in prior works, such as SPADE (*Park et al.*, 2019b). Another approach could be to use existing GAN inversion techniques, which aim to find noise vectors that leads to the generation of images of interest. When such noise vectors are found, the generator can be used to apply minor image edits, both in the style (*Schönfeld*

et al., 2023) and layout aspects (*Richardson et al.*, 2021). Therefore, exploring ways to invert semantic image synthesis generators can assist many applications related to image editing.

The usage of pre-training. In Chapter 7, we explored the usage of pre-trained GANs in few-shot image synthesis. In the context of semantic image synthesis, however, the usage of pre-training remains unexplored. Like in other disciplines, pre-training the models on larger sets of images can contribute to the improvement of the training speed, prevention of overfitting, and overall performance. As large-scale collections of images with pixel-wise annotations cannot be found in all domains and for all possible semantic classes, it would be beneficial to design transfer techniques from GANs pre-trained with a set of other semantic classes, or even from unconditional GANs. In addition, it would be interesting to couple semantic image synthesis with the recent findings on using pre-trained feature extractors in discriminators, e.g., as proposed in ProjectedGAN (*Sauer et al.*, 2021).

Data augmentation. Our models in Chapters 5-7 incorporate data augmentation techniques, both during pre-processing (e.g., doubling the dataset size through horizontal flipping) and training. However, semantic image synthesis models, including OASIS, are typically trained without any augmentations. Nevertheless, widely used techniques like differentiable data augmentation (DA) can be applied to semantic image synthesis GANs too, presenting the opportunity to reduce the amount of data required for successful training. It is worth mentioning that our proposed LabelMix consistency regularization can also benefit from advanced transformations and augmentations of label maps, which have the potential to further enhance the effectiveness of the discriminator in providing localized feedback on image realism.

8.3.2 GANs in Low Data Regimes

In Chapters 5-7, we presented new models for unconditional GAN training in low data regimes. In the following, we discuss possible future steps for this research direction.

Universal techniques between data regimes. The literature review in Sec. 2.3 highlights that various low data regimes, including limited-data, few-shot, and one-shot learning, are typically addressed using different models with their own specific training techniques. Having divergent research branches working on very similar problems leads to drawbacks such as confusion and duplicated effort of researchers. In Chapter 5, SIV-GAN takes a step towards the unification of one-shot and few-shot models, being able to train from a single image or a small collection of very similar images. Moreover, all our models from Chapters 5-7 use differentiable augmentation (DA), demonstrating that DA is generally beneficial in all low data regimes. Yet, more research is needed to develop more training strategies that cover broader ranges of limited-data training regimes.

Continual learning. Continual learning aims to learn from a stream of data that comes in sequences. The objective of such learning is to learn about new data while retaining the knowledge about the old data, thus avoiding catastrophic forgetting. In the context of GANs, continual learning can aim to adapt to a new limited set of image building upon the previously shown larger dataset. This has an advantage of potentially using the same model for different limited data domains, unlike our models in Chapters 5-7. It is worth noting that our few-shot adaptation method from Chapters 7 does not support continual learning, as the target generator is not incentivized to remember the exact appearance of object in the source domain. Yet, it remains to be seen whether popular continual learning techniques can be applied to GANs to improve their learning from limited data.

Controllability. While our models in Chapters 5-7 significantly improve the synthesis quality and diversity, more research is yet to be conducted towards making this generation controllable. For example, SIV-GAN and OSMIS can re-generate the provided scene with different locations or number of objects, but synthesizing the exact desired combination can take several attempts with different input noise vectors. Improving the controllability of our models, for example through spatial conditioning or interpretable latent directions (*Voynov and*

Babenko, 2020), is an interesting direction for future work.

Advanced filtering of data augmentation. In Chapter 6, we explored the application of OSMIS to synthetic data generation. As was discussed in Sec. 6.3.4, filtering out noisy generated examples plays a crucial role for the overall effectiveness of generated images for synthetic data augmentation. In the future, more systematic studies on how to select the images that are most useful for downstream applications should be conducted. An interesting question would be to outline theoretical and practical limits of possible performance gains depending on the dataset size, image diversity, and application type.

8.3.3 Broader Outlook and Other Image Generation Models

With the current pace of progress in deep learning, breakthroughs in generative models occur at a very impressive rate. In recent years, generative adversarial networks have made tremendous progress, improving the abilities for image synthesis in numerous aspects. These advances include the overall image quality and diversity, ability to scale up to high-resolution images and large-scale datasets, addressing the challenges of unstable training, and significantly improving the synthesis invertibility, explainability, and controllability. As a result, GANs have found numerous applications in areas such as artistic content creation, image editing, and synthetic data augmentation.

The progress of both semantic image synthesis and unconditional image synthesis models heavily relies on the general development of neural network architectures and training algorithms in other fields of deep learning. In the following, we identify the three trends in deep learning that are likely to shape the development of generative modeling of images in the nearest future.

Transformer-based architectures. As discussed in Sec. 3.2.4, GAN models inherit a lot of architectural advances from discriminative deep learning. For example, most of the models used in this thesis are based on convolutional neural networks (CNNs), mostly ResNets (*He et al.*, 2016). While ResNets are strong models that enable good performance in many computer vision tasks, they have been recently outperformed by transformer-based architectures (*Dosovitskiy et al.*, 2021). In particular, in (*Dosovitskiy et al.*, 2021), it was shown that replacing CNN-based residual blocks with transformer-based attention blocks, originally introduced for natural language processing (*Vaswani et al.*, 2017), allows to achieve higher performance on standard image classification benchmarks such as ImageNet. Since then, vision transformers have quickly conquered state of the art in semantic segmentation (*Xie et al.*, 2021), super-resolution (*Liang et al.*, 2021), video recognition (*Liu et al.*, 2022), image restoration (*Zamir et al.*, 2022), self-supervised representation learning (*Chen et al.*, 2021b), and many other computer vision tasks. In the context of GANs, vision transformers have been already implemented in the generator and discriminator of several models (*Lee et al.*, 2022; *Jiang et al.*, 2021; *Hudson and Zitnick*, 2022), improving image synthesis quality over respective CNN-based baselines. Transformers may appear a suitable choice for image synthesis, as they are well suited for learning the correlations between distant image patches, which can significantly improve the global coherency in generated images. Therefore, it can be expected that more GAN models will adopt transformers in their design to improve the performance, and more research will follow on their application to semantic image synthesis and restricted image domains.

Synergies between data modalities. Recently, the community has seen impressive results in the task of conditional image generation from text prompts, also known as text-to-image translation. These results were in large part enabled by the availability of powerful text embeddings, which were pre-trained on datasets consisting of more than 400 million samples (*Schuhmann et al.*, 2021). The synergy between images and texts will likely be extended to other modalities as well, and there are already signs that training GANs with different conditionings (e.g., text, sketches, and semantic label maps) at the same time leads to higher performance (*Huang et al.*, 2022). Training with multiple modalities presents distinct challenges in data curation, as collecting annotations of multiple types can be prohibitive in practice. Overcoming these challenges, for example via novel semi-supervised training schemes or algorithms that deal with unbalanced or missing annotations,

may give a boost to new content creation tools or producing synthetic data augmentation in more controlled ways.

Diffusion models. Finally, there remains a question whether GANs will be replaced by more recent model designs. In recent times, numerous advancements in image synthesis have been made not with GANs, but with diffusion models (DMs). In Chapter 4, for instance, we demonstrated that DMs already outperform OASIS and other GAN models in semantic image synthesis, as evidenced by higher FID and LPIPS scores. Diffusion models have also demonstrated superior performance compared to GANs in many other tasks (*Nichol and Dhariwal*, 2021; *Dhariwal and Nichol*, 2021; *Rombach et al.*, 2022; *Xue et al.*, 2023). Therefore, it is likely that diffusion models will become the dominant class of image generation models in the near future. These models have several advantages, including stable training, absence of mode collapse issues, and effective scalability to multiple modalities. However, it is improbable that GANs will be completely replaced. The primary limitation of diffusion models lies in their notably slow inference speed, which makes GANs still more favorable for real-time applications or situations with hardware constraints. In semantic image synthesis, GANs still achieve much better alignment between images and label maps, as measured by mIoU, making them more suitable for tasks like synthetic data augmentation. Furthermore, similar to GANs, the early successes of DMs have been enabled by the availability of large training datasets, while their performance in low data regimes remains suboptimal. Therefore, we can expect the future of image generation to be shaped by concurrent advancement of GANs and DMs, with these models specializing in different dataset types, use cases, and applications.

Bibliography

Alaa, Ahmed; Van Breugel, Boris; Saveliev, Evgeny S, and van der Schaar, Mihaela. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning (ICML)*, 2022. (Cited on page 30.)

Alharbi, Yazeed and Wonka, Peter. Disentangled image generation through structured noise injection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on page 18.)

Amirreza, Shaban; Shray, Bansal; Zhen, Liu; Irfan, Essa, and Byron, Boots. One-shot learning for semantic segmentation. In *British Machine Vision Conference (BMVC)*, 2017. (Cited on pages 86 and 87.)

Arjovsky, Martin and Bottou, Léon. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2017. (Cited on pages 5, 14 and 58.)

Arjovsky, Martin; Chintala, Soumith, and Bottou, Léon. Wasserstein generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. (Cited on pages 14, 29 and 35.)

Badrinarayanan, Vijay; Kendall, Alex, and Cipolla, Roberto. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2016. (Cited on page 50.)

Bengio, Yoshua; Léonard, Nicholas, and Courville, Aaron. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv:1308.3432*, 2013. (Cited on page 89.)

Bergmann, Urs; Jetchev, Nikolay, and Vollgraf, Roland. Learning texture manifolds with the periodic spatial gan. In *International Conference on Machine Learning (ICML)*, 2017. (Cited on page 22.)

Berman, Maxim; Triki, Amal Rannen, and Blaschko, Matthew B. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. (Cited on page 116.)

Borji, Ali. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding*, 2022. (Cited on page 30.)

Boudiaf, Malik; Kervadec, Hoel; Masud, Ziko Imtiaz; Piantanida, Pablo; Ben Ayed, Ismail, and Dolz, Jose. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (Cited on pages 9, 88, 94, 95, 96 and 100.)

Brock, Andrew; Donahue, Jeff, and Simonyan, Karen. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019. (Cited on pages 16, 17, 37, 39, 46, 47, 68, 73, 101, 104, 110, 111 and 112.)

Bruna, Joan; Sprechmann, Pablo, and LeCun, Yann. Super-resolution with deep convolutional sufficient statistics. In *International Conference on Learning Representations (ICLR)*, 2016. (Cited on page 19.)

Caelles, Sergi; Maninis, Kevis-Kokitsi; Pont-Tuset, Jordi; Leal-Taixé, Laura; Cremers, Daniel, and Van Gool, Luc. One-shot video object segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. (Cited on pages 9, 87, 94, 95 and 96.)

Caesar, Holger; Uijlings, Jasper, and Ferrari, Vittorio. Coco-stuff: Thing and stuff classes in context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. (Cited on pages 8, 48, 53 and 86.)

Casanova, Arantxa; Careil, Marlène; Verbeek, Jakob; Drozdzal, Michal, and Romero Soriano, Adriana. Instance-conditioned GAN. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on pages 16, 17 and 46.)

Chan, Eric R; Lin, Connor Z; Chan, Matthew A; Nagano, Koki; Pan, Boxiao; De Mello, Shalini; Gallo, Orazio; Guibas, Leonidas J; Tremblay, Jonathan, and Khamis, Sameh. Efficient geometry-aware 3d generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (Cited on page 100.)

Chen, Liang-Chieh; Papandreou, George; Kokkinos, Iasonas; Murphy, Kevin, and Yuille, Alan L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *International Conference on Learning Representations (ICLR)*, 2015. (Cited on page 54.)

Chen, Liang-Chieh; Zhu, Yukun; Papandreou, George; Schroff, Florian, and Adam, Hartwig. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision (ECCV)*, 2018. (Cited on pages 50 and 86.)

Chen, Qifeng and Koltun, Vladlen. Photographic image synthesis with cascaded refinement networks. In *International Conference on Computer Vision (ICCV)*, 2017. (Cited on pages 19 and 20.)

Chen, Tianlong; Cheng, Yu; Gan, Zhe; Liu, Jingjing, and Wang, Zhangyang. Data-efficient GAN training beyond (just) augmentations: A lottery ticket perspective. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a. (Cited on page 21.)

Chen, Ting; Lucic, Mario; Houlsby, Neil, and Gelly, Sylvain. On self modulation for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019. (Cited on page 39.)

Chen, Xinlei; Xie, Saining, and He, Kaiming. An empirical study of training self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021b. (Cited on page 118.)

Child, Rewon. Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations (ICLR)*, 2022. (Cited on pages 24 and 25.)

Choi, Yunjey; Uh, Youngjung; Yoo, Jaejun, and Ha, Jung-Woo. StarGAN v2: Diverse image synthesis for multiple domains. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on pages 15, 16, 73, 79 and 100.)

Chong, Min Jin and Forsyth, David. Effectively unbiased FID and inception score and where to find them. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on page 32.)

Cordts, Marius; Omran, Mohamed; Ramos, Sebastian; Rehfeld, Timo; Enzweiler, Markus; Benenson, Rodrigo; Franke, Uwe; Roth, Stefan, and Schiele, Bernt. The cityscapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. (Cited on pages 48 and 53.)

Cubuk, Ekin Dogus; Zoph, Barret; Shlens, Jon, and Le, Quoc. Randaugment: Practical automated data augmentation with a reduced search space. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. (Cited on page 62.)

Cui, Kaiwen; Huang, Jiaxing; Luo, Zhipeng; Zhang, Gongjie; Zhan, Fangneng, and Lu, Shijian. Genco: generative co-training for generative adversarial networks with limited data. In *Conference on Artificial Intelligence (AAAI)*, 2022. (Cited on page 21.)

Dauphin, Yann; De Vries, Harm, and Bengio, Yoshua. Equilibrated adaptive learning rates for non-convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. (Cited on page 102.)

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. (Cited on pages 5, 17, 19, 31, 41, 47 and 110.)

Dhariwal, Prafulla and Nichol, Alexander. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on pages 20, 25 and 119.)

Dong, Qiaole; Cao, Chenjie, and Fu, Yanwei. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (Cited on pages 69 and 82.)

Dosovitskiy, Alexey and Brox, Thomas. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. (Cited on page 74.)

Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg, and Gelly, Sylvain. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. (Cited on page 118.)

Gatys, L. A.; Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. (Cited on page 19.)

Gatys, Leon; Ecker, Alexander S, and Bethge, Matthias. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPs)*. 2015. (Cited on page 19.)

Goetschalckx, Lore; Andonian, Alex; Oliva, Aude, and Isola, Phillip. Ganalyze: Toward visual definitions of cognitive image properties. In *International Conference on Computer Vision (ICCV)*, 2019. (Cited on page 30.)

Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. (Cited on pages 2, 13, 14, 15, 16 and 52.)

Gulrajani, Ishaan; Ahmed, Faruk; Arjovsky, Martin; Dumoulin, Vincent, and Courville, Aaron C. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. (Cited on pages 15 and 35.)

Gupta, Agrim; Dollar, Piotr, and Girshick, Ross. Lvis: A dataset for large vocabulary instance segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. (Cited on pages 3, 8, 48, 49, 53, 54 and 100.)

Härkönen, Erik; Hertzmann, Aaron; Lehtinen, Jaakko, and Paris, Sylvain. Ganspace: Discovering interpretable GAN controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. (Cited on pages 101 and 102.)

Hazami, Louay; Mama, Rayhane, and Thurairatnam, Ragavan. Efficient-VDVAE: Less is more. *arXiv:2203.13751*, 2022. (Cited on page 24.)

He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. (Cited on page 118.)

He, Kaiming; Gkioxari, Georgia; Dollár, Piotr, and Girshick, Ross. Mask R-CNN. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. (Cited on page 86.)

Hermosilla, Gabriel; Tapia, Diego-Ignacio Henríquez; Allende-Cid, Hector; Castro, Gonzalo Farías, and Vera, Esteban. Thermal face generation using stylegan. *IEEE Access*, 2021. (Cited on page 14.)

Heusel, Martin; Ramsauer, Hubert; Unterthiner, Thomas; Nessler, Bernhard, and Hochreiter, Sepp. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. (Cited on pages 30, 31, 54 and 104.)

Hinz, Tobias; Fisher, Matthew; Wang, Oliver, and Wermter, Stefan. Improved techniques for training single-image GANs. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021. (Cited on pages 6, 22, 69, 70, 71, 74, 75, 87 and 92.)

Ho, Jonathan; Jain, Ajay, and Abbeel, Pieter. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. (Cited on page 25.)

Hou, Liang; Cao, Qi; Shen, Huawei; Pan, Siyuan; Li, Xiaoshuang, and Cheng, Xueqi. Conditional GANs with auxiliary discriminative classifier. In *International Conference on Machine Learning (ICML)*, 2022. (Cited on page 17.)

Huang, Xun and Belongie, Serge. Arbitrary style transfer in real-time with adaptive instance normalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. (Cited on page 38.)

Huang, Xun; Liu, Ming-Yu; Belongie, Serge, and Kautz, Jan. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*, 2018. (Cited on pages 16 and 46.)

Huang, Xun; Mallya, Arun; Wang, Ting-Chun, and Liu, Ming-Yu. Multimodal conditional image synthesis with product-of-experts GANs. In *European Conference on Computer Vision (ECCV)*, 2022. (Cited on page 118.)

Hudson, Drew A and Zitnick, Larry. Generative adversarial transformers. In *International conference on machine learning (ICLR)*, 2022. (Cited on page 118.)

Isola, Phillip; Zhu, Jun-Yan; Zhou, Tinghui, and Efros, Alexei A. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. (Cited on pages 16, 17, 18, 19, 40, 46, 47, 50, 69, 71 and 103.)

Jain, Himalaya; Vu, Tuan-Hung; Pérez, Patrick, and Cord, Matthieu. CSG0: Continual urban scene generation with zero forgetting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (Cited on page 19.)

Jeong, Jaebong; Jo, Janghun; Wang, Jingdong; Cho, Sunghyun, and Park, Jaesik. Realistic image synthesis with configurable 3d scene layouts. *arXiv:2108.10031*, 2021. (Cited on page 19.)

Jiang, Yifan; Chang, Shiyu, and Wang, Zhangyang. TransGAN: Two pure transformers can make one strong GAN, and that can scale up. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on page 118.)

Johnson, Justin; Alahi, Alexandre, and Fei-Fei, Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016. (Cited on page 19.)

Karnewar, Animesh and Wang, Oliver. MSG-GAN: multi-scale gradient GAN for stable image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on pages 16, 38 and 74.)

Karras, Tero; Aila, Timo; Laine, Samuli, and Lehtinen, Jaakko. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. (Cited on page 15.)

Karras, Tero; Laine, Samuli, and Aila, Timo. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. (Cited on pages 15, 16, 38, 47, 92, 104 and 108.)

Karras, Tero; Aittala, Miika; Hellsten, Janne; Laine, Samuli; Lehtinen, Jaakko, and Aila, Timo. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a. (Cited on pages 20, 41, 47, 69, 74, 78 and 90.)

Karras, Tero; Laine, Samuli; Aittala, Miika; Hellsten, Janne; Lehtinen, Jaakko, and Aila, Timo. Analyzing and improving the image quality of StyleGAN. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020b. (Cited on pages 2, 15, 16, 36, 38, 39, 47, 68, 101, 102 and 104.)

Karras, Tero; Aittala, Miika; Laine, Samuli; Härkönen, Erik; Hellsten, Janne; Lehtinen, Jaakko, and Aila, Timo. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on pages 16, 38, 47, 68 and 100.)

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. (Cited on pages 33 and 54.)

Kingma, Diederik P. and Welling, Max. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. (Cited on page 24.)

Kirillov, Alexander; He, Kaiming; Girshick, Ross; Rother, Carsten, and Dollár, Piotr. Panoptic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. (Cited on page 86.)

Kong, Chaerin; Kim, Jeesoo; Han, Donghoon, and Kwak, Nojun. Few-shot image generation with mixup-based distance learning. In *European Conference on Computer Vision (ECCV)*, 2022. (Cited on pages 15, 104 and 108.)

Krizhevsky, Alex and Hinton, Geoffrey. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*, 2009. (Cited on page 13.)

Kurach, Karol; Lučić, Mario; Zhai, Xiaohua; Michalski, Marcin, and Gelly, Sylvain. The GAN landscape: Losses, architectures, regularization, and normalization. In *International Conference on Machine learning (ICML)*, 2019. (Cited on page 14.)

LeCun, Yann; Bottou, Léon; Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. (Cited on page 13.)

Lee, Kwonjoon; Chang, Huiwen; Jiang, Lu; Zhang, Han; Tu, Zhuowen, and Liu, Ce. ViTGAN: Training GANs with vision transformers. In *International Conference on Learning Representations (ICLR)*, 2022. (Cited on page 118.)

Leng, Zhaoqi; Tan, Mingxing; Liu, Chenxi; Cubuk, Ekin Dogus; Shi, Jay; Cheng, Shuyang, and Anguelov, Dragomir. Polyloss: A polynomial expansion perspective of classification loss functions. In *International Conference on Learning Representations (ICLR)*, 2022. (Cited on page 116.)

Li, Daiqing; Yang, Junlin; Kreis, Karsten; Torralba, Antonio, and Fidler, Sanja. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a. (Cited on pages 23, 86, 87, 92 and 93.)

Li, Yijun; Zhang, Richard; Lu, Jingwan Cynthia, and Shechtman, Eli. Few-shot image generation with elastic weight consolidation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. (Cited on pages 22 and 91.)

Li, Yuheng; Li, Yijun; Lu, Jingwan; Shechtman, Eli; Lee, Yong Jae, and Singh, Krishna Kumar. Collaging class-specific GANs for semantic image synthesis. In *International Conference on Computer Vision (ICCV)*, 2021b. (Cited on pages 18 and 19.)

Li, Ziqiang; Wang, Chaoyue; Zheng, Heliang; Zhang, Jing, and Li, Bin. FakeCLR: Exploring contrastive learning for solving latent discontinuity in data-efficient GANs. In *European Conference on Computer Vision (ECCV)*, 2022. (Cited on page 21.)

Liang, Jingyun; Cao, Jiezhang; Sun, Guolei; Zhang, Kai; Van Gool, Luc, and Timofte, Radu. Swinir: Image restoration using swin transformer. In *International Conference on Computer Vision*, 2021. (Cited on page 118.)

Lim, Jae Hyun and Ye, Jong Chul. Geometric GAN. *ArXiv:1705.02894*, 2017. (Cited on page 35.)

Lin, Tsung-Yi; Maire, Michael; Belongie, Serge J.; Bourdev, Lubomir D.; Girshick, Ross B.; Hays, James; Perona, Pietro; Ramanan, Deva; Dollár, Piotr, and Zitnick, C. Lawrence. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. (Cited on pages 43, 91, 92 and 94.)

Lin, Tsung-Yi; Goyal, Priya; Girshick, Ross B.; He, Kaiming, and Dollár, Piotr. Focal loss for dense object detection. *International Conference on Computer Vision (ICCV)*, 2017. (Cited on page 116.)

Ling, Huan; Kreis, Karsten; Li, Daiqing; Kim, Seung Wook; Torralba, Antonio, and Fidler, Sanja. EditGAN: High-precision semantic image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on page 23.)

Liu, Bingchen; Zhu, Yizhe; Song, Kunpeng, and Elgammal, Ahmed. Towards faster and stabilized GAN training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. (Cited on pages 21, 47, 69, 70, 75, 104 and 111.)

Liu, Xihui; Yin, Guojun; Shao, Jing, and Wang, Xiaogang. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. (Cited on pages 18, 19, 46, 47, 54 and 57.)

Liu, Ze; Ning, Jia; Cao, Yue; Wei, Yixuan; Zhang, Zheng; Lin, Stephen, and Hu, Han. Video swin transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (Cited on page 118.)

Mao, Xudong; Li, Qing; Xie, Haoran; Lau, Raymond YK; Wang, Zhen, and Paul Smolley, Stephen. Least squares generative adversarial networks. In *International Conference on Computer Vision (ICCV)*, 2017. (Cited on pages 14, 15 and 35.)

Mathieu, Michael F; Zhao, Junbo Jake; Zhao, Junbo; Ramesh, Aditya; Sprechmann, Pablo, and LeCun, Yann. Disentangling factors of variation in deep representation using adversarial training. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. (Cited on page 16.)

Mescheder, Lars; Geiger, Andreas, and Nowozin, Sebastian. Which training methods for GANs do actually converge? In *International Conference on Machine learning (ICML)*, 2018. (Cited on pages 5, 14 and 15.)

Mirza, Mehdi and Osindero, Simon. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014. (Cited on pages 16 and 46.)

Miyato, Takeru and Koyama, Masanori. cGANs with projection discriminator. In *International Conference on Learning Representations (ICLR)*, 2018. (Cited on pages 17, 18, 19, 52, 64, 93 and 110.)

Miyato, Takeru; Kataoka, Toshiki; Koyama, Masanori, and Yoshida, Yuichi. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. (Cited on pages 15, 35, 47 and 74.)

Mo, Sangwoo; Cho, Minsu, and Shin, Jinwoo. Freeze discriminator: A simple baseline for fine-tuning GANs. In *CVPR AI for Content Creation Workshop*, 2020. (Cited on pages 23 and 104.)

Mroueh, Youssef and Sercu, Tom. Fisher GAN. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. (Cited on page 14.)

Mroueh, Youssef; Li, Chun-Liang; Sercu, Tom; Raj, Anant, and Cheng, Yu. Sobolev GAN. *arXiv:1711.04894*, 2017a. (Cited on page 14.)

Mroueh, Youssef; Sercu, Tom, and Goel, Vaibhava. MCGAN: Mean and covariance feature matching GAN. In *International Conference on Machine learning (ICML)*, 2017b. (Cited on page 14.)

Musat, Valentina; De Martini, Daniele; Gadd, Matthew, and Newman, Paul. Depth-SIMS: Semi-parametric image and depth synthesis. In *International Conference on Robotics and Automation (ICRA)*, 2022. (Cited on page 19.)

Nagarajan, Vaishnavh; Raffel, Colin, and Goodfellow, Ian J. Theoretical insights into memorization in GANs. In *Advances in Neural Information Processing Systems (NeurIPS) Workshops*, 2018. (Cited on page 87.)

Nichol, Alexander Quinn and Dhariwal, Prafulla. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2021. (Cited on pages 25 and 119.)

Nilsback, Maria-Elena and Zisserman, Andrew. A visual vocabulary for flower classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. (Cited on page 111.)

Nilsson, David and Sminchisescu, Cristian. Semantic video segmentation by gated recurrent flow propagation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. (Cited on page 86.)

Noguchi, Atsuhiro and Harada, T. Image generation from small datasets via batch statistics adaptation. In *International Conference on Computer Vision (ICCV)*, 2019. (Cited on page 22.)

Nowozin, Sebastian; Cseke, Botond, and Tomioka, Ryota. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. (Cited on page 14.)

Ntavelis, Evangelos; Romero, Andrés; Kastanis, Iason; Van Gool, Luc, and Timofte, Radu. SESAME: Semantic editing of scenes by adding, manipulating or erasing objects. *arXiv:2004.04977*, 2020. (Cited on pages 18, 19, 46, 47 and 55.)

Odena, Augustus; Olah, Christopher, and Shlens, Jonathon. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Learning Representations (ICLR)*, 2017. (Cited on pages 17 and 36.)

Odena, Augustus; Buckman, Jacob; Olsson, Catherine; Brown, Tom; Olah, Christopher; Raffel, Colin, and Goodfellow, Ian. Is generator conditioning causally related to GAN performance? In *International Conference on Machine Learning (ICML)*, 2018. (Cited on page 15.)

Oh, Seoung Wug; Lee, Joon-Young; Xu, Ning, and Kim, Seon Joo. Video object segmentation using space-time memory networks. In *International Conference on Computer Vision (ICCV)*, 2019. (Cited on pages 88, 94, 95 and 96.)

Ojala, Timo; Pietikäinen, Matti, and Harwood, David. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996. (Cited on page 55.)

Ojha, Utkarsh; Li, Yijun; Lu, Jingwan; Efros, Alexei A; Lee, Yong Jae; Shechtman, Eli, and Zhang, Richard. Few-shot image generation via cross-domain correspondence. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (Cited on pages 23, 100, 103, 104, 107, 109 and 110.)

Park, Jongchan; Woo, Sanghyun; Lee, Joon-Young, and Kweon, In So. BAM: Bottleneck attention module. In *British Machine Vision Conference (BMVC)*, 2018. (Cited on page 71.)

Park, T.; Zhu, Jun-Yan; Wang, Oliver; Lu, Jingwan; Shechtman, E.; Efros, Alexei A., and Zhang, Richard. Swapping autoencoder for deep image manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a. (Cited on page 16.)

Park, Taesung; Liu, Ming-Yu; Wang, Ting-Chun, and Zhu, Jun-Yan. GauGAN: semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH*, 2019a. (Cited on page 46.)

Park, Taesung; Liu, Ming-Yu; Wang, Ting-Chun, and Zhu, Jun-Yan. Semantic image synthesis with spatially-adaptive normalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019b. (Cited on pages 16, 18, 19, 20, 39, 46, 47, 48, 49, 54, 57, 114 and 116.)

Park, Taesung; Efros, Alexei A; Zhang, Richard, and Zhu, Jun-Yan. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision (ECCV)*, 2020b. (Cited on pages 16 and 46.)

Perazzi, Federico; Pont-Tuset, Jordi; McWilliams, Brian; Van Gool, Luc; Gross, Markus, and Sorkine-Hornung, Alexander. A benchmark dataset and evaluation methodology for video object segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. (Cited on pages 74, 76, 86, 87, 91 and 94.)

Qi, Xiaojuan; Chen, Qifeng; Jia, Jiaya, and Koltun, Vladlen. Semi-parametric image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. (Cited on pages 20 and 53.)

Radford, Alec; Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016. (Cited on page 15.)

Reed, Scott E.; Akata, Zeynep; Yan, Xinchen; Logeswaran, Lajanugen; Schiele, Bernt, and Lee, Honglak. Generative adversarial text to image synthesis. In *International Conference on Machine learning (ICML)*, 2016. (Cited on page 46.)

Ren, Jie; Liu, Peter J; Fertig, Emily; Snoek, Jasper; Poplin, Ryan; Depristo, Mark; Dillon, Joshua, and Lakshminarayanan, Balaji. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. (Cited on page 100.)

Richardson, Elad; Alaluf, Yuval; Patashnik, Or; Nitzan, Yotam; Azar, Yaniv; Shapiro, Stav, and Cohen-Or, Daniel. Encoding in style: a stylegan encoder for image-to-image translation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (Cited on pages 18, 19 and 117.)

Robb, Esther; Chu, Wen-Sheng; Kumar, Abhishek, and Huang, Jia-Bin. Few-shot adaptation of generative adversarial networks. *arXiv:2010.11943*, 2021. (Cited on pages 22, 74 and 91.)

Rombach, Robin; Blattmann, Andreas; Lorenz, Dominik; Esser, Patrick, and Ommer, Björn. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (Cited on pages 20, 25 and 119.)

Ronneberger, Olaf; Fischer, Philipp, and Brox, Thomas. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. (Cited on pages 47, 50 and 91.)

Rubner, Yossi; Tomasi, Carlo, and Guibas, Leonidas J. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision (IJCV)*, 2000. (Cited on page 55.)

Saha, Oindrila; Cheng, Zezhou, and Maji, Subhransu. GANORCON: Are generative models useful for few-shot segmentation? *arXiv:2112.00854*, 2021. (Cited on page 86.)

Sajjadi, Mehdi SM; Bachem, Olivier; Lucic, Mario; Bousquet, Olivier, and Gelly, Sylvain. Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. (Cited on page 32.)

Salimans, Tim; Goodfellow, Ian; Zaremba, Wojciech; Cheung, Vicki; Radford, Alec; Chen, Xi, and Chen, Xi. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. (Cited on page 31.)

Sauer, Axel; Chitta, Kashyap; Müller, Jens, and Geiger, Andreas. Projected GANs converge faster. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on pages 47 and 117.)

Sauer, Axel; Schwarz, Katja, and Geiger, Andreas. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *ACM SIGGRAPH*, 2022. (Cited on pages 17 and 100.)

Sauer, Axel; Karras, Tero; Laine, Samuli; Geiger, Andreas, and Aila, Timo. StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. *arXiv:2301.09515*, 2023. (Cited on pages 16 and 100.)

Schönfeld, Edgar; Schiele, Bernt, and Khoreva, Anna. A U-Net based discriminator for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on pages 16, 19, 40, 52 and 68.)

Schönfeld, Edgar; Sushko, Vadim; Zhang, Dan; Gall, Juergen; Schiele, Bernt, and Khoreva, Anna. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. (Cited on pages 7, 45 and 100.)

Schönfeld, Edgar; Borges, Julio; Sushko, Vadim; Schiele, Bernt, and Khoreva, Anna. Discovering class-specific GAN controls for semantic image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. (Cited on page 116.)

Schuhmann, Christoph; Vencu, Richard; Beaumont, Romain; Kaczmarczyk, Robert; Mullis, Clayton; Katta, Aarush; Coombes, Theo; Jitsev, Jenia, and Komatsuzaki, Aran. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*, 2021. (Cited on page 118.)

Shaham, Tamar Rott; Dekel, Tali, and Michaeli, T. SinGAN: Learning a generative model from a single natural image. In *International Conference on Computer Vision (ICCV)*, 2019. (Cited on pages 6, 22, 32, 69, 70, 71, 74, 75, 87, 91, 92 and 95.)

Shannon, Matt; Poole, Ben; Mariooryad, Soroosh; Bagby, Tom; Battenberg, Eric; Kao, David; Stanton, Daisy, and Skerry-Ryan, RJ. Non-saturating GAN training as divergence minimization. *arXiv:2010.08029*, 2021. (Cited on page 14.)

Shen, Yujun and Zhou, Bolei. Closed-form factorization of latent semantics in GANs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (Cited on pages 101 and 102.)

Shijie, Li; Ming-Ming, Cheng, and Juergen, Gall. Dual pyramid generative adversarial networks for semantic image synthesis. In *British Machine Vision Conference (BMCV)*, 2022. (Cited on page 16.)

Shocher, Assaf; Cohen, N., and Irani, M. Zero-shot super-resolution using deep internal learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. (Cited on page 22.)

Shocher, Assaf; Bagon, S.; Isola, Phillip, and Irani, M. Ingan: Capturing and retargeting the "dna" of a natural image. In *International Conference on Computer Vision (ICCV)*, 2019. (Cited on pages 22 and 69.)

Simard, Patrice Y; Steinkraus, David, and Platt, John C. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, 2003. (Cited on page 20.)

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. (Cited on pages 41 and 47.)

Souly, Nasim; Spampinato, Concetto, and Shah, Mubarak. Semi supervised semantic segmentation using generative adversarial network. In *International Conference on Computer Vision (ICCV)*, 2017. (Cited on pages 19 and 40.)

Srivastava, Nitish; Hinton, Geoffrey; Krizhevsky, Alex; Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 2014. (Cited on page 72.)

Sudre, Carole H; Li, Wenqi; Vercauteren, Tom; Ourselin, Sebastien, and Cardoso, M Jorge. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, 2017. (Cited on pages 5, 48 and 58.)

Sushko, Vadim; Gall, Juergen, and Khoreva, Anna. One-Shot GAN: Learning to generate samples from single images and videos. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021a. (Cited on pages 8 and 67.)

Sushko, Vadim; Zhang, Dan; Gall, Juergen, and Khoreva, Anna. Learning to generate novel scene compositions from single images and videos. *arXiv:2103.13389*, 2021b. (Cited on pages 87, 88, 92 and 103.)

Sushko, Vadim; Schönfeld, Edgar; Zhang, Dan; Gall, Juergen; Schiele, Bernt, and Khoreva, Anna. OASIS: Only adversarial supervision for semantic image synthesis. *International Journal of Computer Vision (IJCV)*, 2022. (Cited on pages 7 and 45.)

Sushko, Vadim; Wang, Ruyu, and Gall, Juergen. Smoothness similarity regularization for few-shot GAN adaptation. In *International Conference on Computer Vision (ICCV)*, 2023a. (Cited on pages 10 and 99.)

Sushko, Vadim; Zhang, Dan; Gall, Juergen, and Khoreva, Anna. One-shot synthesis of images and segmentation masks. In *Winter Conference on Applications of Computer Vision (WACV)*, 2023b. (Cited on pages 9 and 85.)

Szegedy, Christian; Vanhoucke, Vincent; Ioffe, Sergey; Shlens, Jon, and Wojna, Zbigniew. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. (Cited on page 31.)

Tan, Zhentao; Chen, Dongdong; Chu, Qi; Chai, Menglei; Liao, Jing; He, Mingming; Yuan, Lu, and Yu, Nenghai. Rethinking spatially-adaptive normalization. *arXiv:2004.02867*, 2020. (Cited on page 19.)

Tang, Hao; Bai, Song, and Sebe, Nicu. Dual attention GANs for semantic image synthesis. *arXiv:2008.13024*, 2020a. (Cited on page 19.)

Tang, Hao; Qi, Xiaojuan; Xu, Dan; Torr, Philip HS, and Sebe, Nicu. Edge guided GANs with semantic preserving for semantic image synthesis. *arXiv:2003.13898*, 2020b. (Cited on page 18.)

Tang, Hao; Xu, Dan; Yan, Yan; Torr, Philip HS, and Sebe, Nicu. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020c. (Cited on pages 18 and 19.)

Thomee, Bart; Shamma, David A; Friedland, Gerald; Elizalde, Benjamin; Ni, Karl; Poland, Douglas; Borth, Damian, and Li, Li-Jia. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2016. (Cited on pages 74, 76 and 77.)

Tian, Yonglong; Wang, Yue; Krishnan, Dilip; Tenenbaum, Joshua B, and Isola, Phillip. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, 2020a. (Cited on page 69.)

Tian, Zhuotao; Zhao, Hengshuang; Shu, Michelle; Yang, Zhicheng; Li, Ruiyu, and Jia, Jiaya. Prior guided feature enrichment network for few-shot segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2020b. (Cited on page 100.)

Tritrong, Nontawat; Rewatbowornwong, Pitchaporn, and Suwajanakorn, Supasorn. Repurposing GANs for one-shot semantic part segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (Cited on pages 23, 86 and 89.)

Tseng, Hung-Yu; Jiang, Lu; Liu, Ce; Yang, Ming-Hsuan, and Yang, Weilong. Regularizing generative adversarial networks under limited data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (Cited on page 21.)

Ulyanov, D.; Vedaldi, A., and Lempitsky, V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. (Cited on page 22.)

Ulyanov, D.; Vedaldi, A., and Lempitsky, V. Deep image prior. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. (Cited on page 22.)

Vahdat, Arash and Kautz, Jan. NVAE: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. (Cited on page 24.)

Van Den Oord, Aäron; Kalchbrenner, Nal, and Kavukcuoglu, Koray. Pixel recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2016. (Cited on page 30.)

Van den Oord, Aaron; Vinyals, Oriol, and Kavukcuoglu, Koray. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. (Cited on page 89.)

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz, and Polosukhin, Illia. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. (Cited on page 118.)

Voynov, Andrey and Babenko, Artem. Unsupervised discovery of interpretable directions in the GAN latent space. In *International conference on machine learning (ICML)*, 2020. (Cited on pages 101, 102 and 117.)

Wan, Li; Zeiler, Matthew; Zhang, Sixin; Le Cun, Yann, and Fergus, Rob. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning (ICML)*, 2013. (Cited on page 20.)

Wang, Jiaqi; Zhang, Wenwei; Zang, Yuhang; Cao, Yuhang; Pang, Jiangmiao; Gong, Tao; Chen, Kai; Liu, Ziwei; Loy, Chen Change, and Lin, Dahua. Seesaw loss for long-tailed instance segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a. (Cited on page 54.)

Wang, Kaixin; Liew, Jun Hao; Zou, Yingtian; Zhou, Daquan, and Feng, Jiashi. Panet: Few-shot image semantic segmentation with prototype alignment. In *International Conference on Computer Vision (ICCV)*, 2019. (Cited on page 86.)

Wang, Tengfei; Zhang, Ting; Zhang, Bo; Ouyang, Hao; Chen, Dong; Chen, Qifeng, and Wen, Fang. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022a. (Cited on pages 20 and 56.)

Wang, Ting-Chun; Liu, Ming-Yu; Zhu, Jun-Yan; Tao, Andrew; Kautz, Jan, and Catanzaro, Bryan. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018a. (Cited on pages 5, 18, 19 and 47.)

Wang, Weilun; Bao, Jianmin; Zhou, Wengang; Chen, Dongdong; Chen, Dong; Yuan, Lu, and Li, Houqiang. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022b. (Cited on pages 20 and 56.)

Wang, Yaxing; Wu, Chenshen; Herranz, L.; van de Weijer, Joost; Gonzalez-Garcia, Abel, and Raducanu, B. Transferring GANs: generating images from limited data. In *European Conference on Computer Vision (ECCV)*, 2018b. (Cited on pages 22 and 104.)

Wang, Yaxing; Gonzalez-Garcia, Abel; Berga, David; Herranz, L.; Khan, F., and van de Weijer, Joost. MineGAN: Effective knowledge transfer from GANs to target domains with few images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on page 22.)

Wang, Yi; Qi, Lu; Chen, Ying-Cong; Zhang, Xiangyu, and Jia, Jiaya. Image synthesis via semantic composition. In *International Conference on Computer Vision (ICCV)*, 2021b. (Cited on pages 16, 18, 19, 46 and 47.)

Wang, Zhe; Chi, Ziqiu, and Zhang, Yanbing. FreGAN: Exploiting frequency components for training gans under limited data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022c. (Cited on page 21.)

Wang, Zhou; Simoncelli, Eero P, and Bovik, Alan C. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers (ACSSC)*, 2003. (Cited on pages 55 and 56.)

Weng, Lilian. What are diffusion models?, 2021. URL `https://lilianweng.github.io/posts/2021-07-11-diffusion-models/`. (Cited on page 24.)

Woo, Sanghyun; Park, Jongchan; Lee, Joon-Young, and Kweon, In So. CBAM: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)*, 2018. (Cited on page 71.)

Wu, Wayne; Cao, Kaidi; Li, Cheng; Qian, Chen, and Loy, Chen Change. Disentangling content and style via unsupervised geometry distillation. In *International Conference on Learning Representations (ICLR) workshops*, 2019. (Cited on page 16.)

Xiao, Jiayu; Li, Liang; Wang, Chaofei; Zha, Zheng-Jun, and Huang, Qingming. Few shot generative model adaption via relaxed spatial structural alignment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (Cited on pages 23, 100, 103 and 104.)

Xiao, Tete; Liu, Yingcheng; Zhou, Bolei; Jiang, Yuning, and Sun, Jian. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, 2018. (Cited on page 54.)

Xie, Enze; Wang, Wenhai; Yu, Zhiding; Anandkumar, Anima; Alvarez, Jose M, and Luo, Ping. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on page 118.)

Xu, Jianjin and Zheng, Changxi. Linear semantics in generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (Cited on page 23.)

Xue, Han; Huang, Zhiwu; Sun, Qianru; Song, Li, and Zhang, Wenjun. Freestyle layout-to-image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. (Cited on pages 20, 57 and 119.)

Yang, Ceyuan; Shen, Yujun; Xu, Yinghao, and Zhou, Bolei. Data-efficient instance generation from instance discrimination. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on page 21.)

Yang, Dingdong; Hong, Seunghoon; Jang, Y.; Zhao, T., and Lee, H. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019. (Cited on pages 15, 36, 70, 73 and 79.)

Yazici, Yasin; Foo, Chuan-Sheng; Winkler, Stefan; Yap, Kim-Hui; Piliouras, Georgios, and Chandrasekhar, Vijay. The unusual effectiveness of averaging in GAN training. *arXiv:1806.04498*, 2018. (Cited on page 55.)

Yu, Fisher; Zhang, Yinda; Song, Shuran; Seff, Ari, and Xiao, Jianxiong. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365*, 2015. (Cited on page 104.)

Yu, Fisher; Koltun, Vladlen, and Funkhouser, Thomas. Dilated residual networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. (Cited on page 54.)

Yun, Sangdoo; Han, Dongyoon; Oh, Seong Joon; Chun, Sanghyuk; Choe, Junsuk, and Yoo, Youngjoon. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. (Cited on pages 51, 64 and 72.)

Yunqing, ZHAO; Chandrasegaran, Keshigeyan; Abdollahzadeh, Milad, and Cheung, Ngai-man. Few-shot image generation via adaptation-aware kernel modulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. (Cited on pages 23 and 104.)

Zamir, Syed Waqas; Arora, Aditya; Khan, Salman; Hayat, Munawar; Khan, Fahad Shahbaz, and Yang, Ming-Hsuan. Restormer: Efficient transformer for high-resolution image restoration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (Cited on page 118.)

Zhang, Dan and Khoreva, Anna. PA-GAN: Improving GAN training by progressive augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. (Cited on page 47.)

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X., and Metaxas, D. N. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *Transactions on Pattern Analysis and Machine Intelligence*, 2018a. (Cited on pages 16 and 46.)

Zhang, Han; Goodfellow, Ian J.; Metaxas, Dimitris N., and Odena, Augustus. Self-attention generative adversarial networks. In *International Conference on Machine learning (ICML)*, 2019. (Cited on page 16.)

Zhang, Han; Zhang, Zizhao; Odena, Augustus, and Lee, Honglak. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2020a. (Cited on page 20.)

Zhang, Han; Koh, Jing Yu; Baldridge, Jason; Lee, Honglak, and Yang, Yinfei. Cross-modal contrastive learning for text-to-image generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a. (Cited on pages 16 and 46.)

Zhang, Hang; Wu, Chongruo; Zhang, Zhongyue; Zhu, Yi; Zhang, Zhi; Lin, Haibin; Sun, Yue; He, Tong; Mueller, Jonas, and Manmatha, R. Resnest: Split-attention networks. *arXiv:2004.08955*, 2020b. (Cited on pages 55 and 60.)

Zhang, Lingzhi; Wen, Tarmily, and Shi, Jianbo. Deep image blending. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020c. (Cited on pages 69 and 82.)

Zhang, Richard; Isola, Phillip; Efros, Alexei A; Shechtman, Eli, and Wang, Oliver. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018b. (Cited on pages 30, 55, 56, 91 and 104.)

Zhang, Yuxuan; Ling, Huan; Gao, Jun; Yin, Kangxue; Lafleche, Jean-Francois; Barriuso, Adela; Torralba, Antonio, and Fidler, Sanja. DatasetGAN: Efficient labeled data factory with minimal human effort. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b. (Cited on pages 23, 86, 87, 89 and 92.)

Zhao, Junbo; Mathieu, Michael, and LeCun, Yann. Energy-based generative adversarial network. In *International Conference on Learning Representations (ICLR) workshops*, 2017. (Cited on page 14.)

Zhao, Shengyu; Liu, Zhijian; Lin, Ji; Zhu, Jun-Yan, and Han, Song. Differentiable augmentation for data-efficient GAN training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020a. (Cited on pages 20, 23, 40, 41, 74, 78 and 104.)

Zhao, Shuai; Wang, Yang; Yang, Zheng, and Cai, Deng. Region mutual information loss for semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. (Cited on page 116.)

Zhao, Yang; Li, Chunyuan; Yu, Ping; Gao, Jianfeng, and Chen, Changyou. Feature quantization improves GAN training. In *International Conference on Machine Learning (ICML)*, 2020b. (Cited on page 17.)

Zhao, Yunqing; Ding, Henghui; Huang, Houjing, and Cheung, Ngai-Man. A closer look at few-shot image generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (Cited on page 23.)

Zhao, Zhengli; Zhang, Zizhao; Chen, Ting; Singh, Sameer, and Zhang, Han. Image augmentations for GAN training. *arXiv:2006.02595*, 2020c. (Cited on pages 20 and 74.)

Zhao, Zhengli; Singh, Sameer; Lee, Honglak; Zhang, Zizhao; Odena, Augustus, and Zhang, Han. Improved consistency regularization for GANs. In *Conference on Artificial Intelligence (AAAI)*, 2021. (Cited on pages 15, 73 and 79.)

Zhengsu, Tian; Jianwei, Chen, and Qi, Niu. Dropfilter: Dropout for convolutions. *arXiv:1810.09849*, 2018. (Cited on page 72.)

Zhou, Bolei; Lapedriza, Agata; Khosla, Aditya; Oliva, Aude, and Torralba, Antonio. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017a. (Cited on pages 74 and 76.)

Zhou, Bolei; Zhao, Hang; Puig, Xavier; Fidler, Sanja; Barriuso, Adela, and Torralba, Antonio. Scene parsing through ADE20K dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017b. (Cited on pages 48 and 53.)

Zhou, Peng; Xie, Lingxi; Ni, Bingbing; Geng, Cong, and Tian, Qi. Omni-GAN: On the secrets of cGANs and beyond. In *International Conference on Computer Vision (ICCV)*, 2021. (Cited on page 17.)

Zhou, Yang; Zhu, Zhen; Bai, X.; Lischinski, Dani; Cohen-Or, D., and Huang, Hui. Non-stationary texture synthesis by adversarial expansion. In *ACM Transactions on Graphics (TOG)*, 2018. (Cited on page 22.)

Zhu, Zhen; Xu, Zhiliang; You, Ansheng, and Bai, Xiang. Semantically multi-modal image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. (Cited on page 18.)

Appendices

Parts of this thesis are based on several publications as indicated in the corresponding chapters. We provide these publications in the appendices. Namely, Appendix A, B, C, D, and E provide our publications in the following order:

- **You Only Need Adversarial Supervision for Semantic Image Synthesis**
  Edgar Schönfeld*, Vadim Sushko*, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva
  International Conference on Learning Representations (ICLR), 2021.
- **OASIS: Only Adversarial Supervision for Semantic Image Synthesis**
  Vadim Sushko*, Edgar Schönfeld*, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva
  International Journal of Computer Vision (IJCV), 2022.
  DOI: 10.1007/s11263-022-01673-x
- **One-Shot GAN: Learning to Generate Samples from Single Images and Videos**
  Vadim Sushko, Juergen Gall, Anna Khoreva
  IEEE Computer Vision and Pattern Recognition Conference (CVPR) Workshops, 2021.
  DOI: 10.1109/CVPRW53098.2021.00293
- **One-Shot Synthesis of Images and Segmentation Masks**
  Vadim Sushko, Dan Zhang, Juergen Gall, Anna Khoreva
  IEEE Winter Conference on Applications of Computer Vision (WACV), 2023.
  DOI: 10.1109/WACV56688.2023.00622
- **Smoothness Similarity Regularization for Few-Shot GAN Adaptation**
  Vadim Sushko, Ruyu Wang, Juergen Gall
  IEEE International Conference on Computer Vision (ICCV), 2023.
  DOI: 10.1109/ICCV51070.2023.00651

(* denotes equal contribution)

Contents

A You Only Need Adversarial Supervision for Semantic Image Synthesis

In this appendix, we provide the conference publication that Chapter 4 of the thesis is based on:

- **You Only Need Adversarial Supervision for Semantic Image Synthesis**
  Edgar Schönfeld*, Vadim Sushko*, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva
  International Conference on Learning Representations (ICLR), 2021.

YOU ONLY NEED ADVERSARIAL SUPERVISION FOR SEMANTIC IMAGE SYNTHESIS

Edgar Schönfeld *
Bosch Center for Artificial Intelligence

Vadim Sushko *
Bosch Center for Artificial Intelligence

Dan Zhang
Bosch Center for Artificial Intelligence

Jürgen Gall
University of Bonn

Bernt Schiele
Max Planck Institute for Informatics

Anna Khoreva
Bosch Center for Artificial Intelligence

ABSTRACT

Despite their recent successes, GAN models for semantic image synthesis still suffer from poor image quality when trained with only adversarial supervision. Historically, additionally employing the VGG-based perceptual loss has helped to overcome this issue, significantly improving the synthesis quality, but at the same time limiting the progress of GAN models for semantic image synthesis. In this work, we propose a novel, simplified GAN model, which needs only adversarial supervision to achieve high quality results. We re-design the discriminator as a semantic segmentation network, directly using the given semantic label maps as the ground truth for training. By providing stronger supervision to the discriminator as well as to the generator through spatially- and semantically-aware discriminator feedback, we are able to synthesize images of higher fidelity with better alignment to their input label maps, making the use of the perceptual loss superfluous. Moreover, we enable high-quality multi-modal image synthesis through global and local sampling of a 3D noise tensor injected into the generator, which allows complete or partial image change. We show that images synthesized by our model are more diverse and follow the color and texture distributions of real images more closely. We achieve an average improvement of 6 FID and 5 mIoU points over the state of the art across different datasets using only adversarial supervision.

Figure 1: Existing semantic image synthesis models heavily rely on the VGG-based perceptual loss to improve the quality of generated images. In contrast, our model can synthesize diverse and high-quality images while only using an adversarial loss, without any external supervision.

*Equal contribution. Correspondence to {edgar.schoenfeld, vadim.sushko}@bosch.com.

1 INTRODUCTION

Conditional generative adversarial networks (GANs) (Mirza & Osindero, 2014) synthesize images conditioned on class labels (Zhang et al., 2019; Brock et al., 2019), text (Reed et al., 2016; Zhang et al., 2018a), other images (Isola et al., 2017; Huang et al., 2018), or semantic label maps (Wang et al., 2018; Park et al., 2019). In this work, we focus on the latter, addressing semantic image synthesis. Semantic image synthesis enables rendering of realistic images from user-specified layouts, without the use of an intricate graphic engine. Therefore, its applications range widely from content creation and image editing to generating training data that needs to adhere to specific semantic requirements (Wang et al., 2018; Chen & Koltun, 2017). Despite the recent progress on stabilizing GANs (Gulrajani et al., 2017; Miyato et al., 2018; Zhang & Khoreva, 2019) and developing their architectures (Zhang et al., 2019; Karras et al., 2019), state-of-the-art GAN-based semantic image synthesis models (Park et al., 2019; Liu et al., 2019) still greatly suffer from training instabilities and poor image quality when trained only with adversarial supervision (see Fig. 1). An established practice to overcome this issue is to employ a perceptual loss (Wang et al., 2018) to train the generator, in addition to the discriminator loss. The perceptual loss aims to match intermediate features of synthetic and real images, that are estimated via an external perception network. A popular choice for such a network is VGG (Simonyan & Zisserman, 2015), pre-trained on ImageNet (Deng et al., 2009). Although the perceptual loss substantially improves the accuracy of previous methods, it comes with the computational overhead introduced by utilizing an extra network for training. Moreover, it usually dominates over the adversarial loss during training, which can have a negative impact on the diversity and quality of generated images, as we show in our experiments. Therefore, in this work we propose a novel, simplified model that achieves state-of-the-art results without requiring a perceptual loss.

A fundamental question for GAN-based semantic image synthesis models is how to design the discriminator to efficiently utilize information from the given semantic label maps. Conventional methods (Park et al., 2019; Wang et al., 2018; Liu et al., 2019; Isola et al., 2017) adopt a multi-scale classification network, taking the label map as input along with the image, and making a global image-level real/fake decision. Such a discriminator has limited representation power, as it is not incentivized to learn high-fidelity pixel-level details of the images and their precise alignment with the input semantic label maps. To mitigate this issue, we propose an alternative architecture for the discriminator, re-designing it as an encoder-decoder semantic segmentation network (Ronneberger et al., 2015), and directly exploiting the given semantic label maps as ground truth via a (N+1)-class cross-entropy loss (see Fig. 3). This new discriminator provides semantically-aware pixel-level feedback to the generator, partitioning the image into segments belonging to one of the N real semantic classes or the fake class. Enabled by the discriminator per-pixel response, we further introduce a LabelMix regularization, which fosters the discriminator to focus more on the semantic and structural differences of real and synthetic images. The proposed changes lead to a much stronger discriminator, that maintains a powerful semantic representation of objects, giving more meaningful feedback to the generator, and thus making the perceptual loss supervision superfluous (see Fig. 1).

Next, we propose to enable multi-modal synthesis of the generator via 3D noise sampling. Previously, directly using 1D noise as input was not successful for semantic image synthesis, as the generator tended to mostly ignore it or synthesized images of poor quality (Isola et al., 2017; Wang et al., 2018). Thus, prior work (Wang et al., 2018; Park et al., 2019) resorted to using an image encoder to produce multi-modal outputs. In this work, we propose a lighter solution. Empowered by our stronger discriminator, the generator can effectively synthesize different images by simply re-sampling a 3D noise tensor, which is used not only as the input but also combined with intermediate features via conditional normalization at every layer. Such noise is spatially sensitive, so we can re-sample it both globally (channel-wise) and locally (pixel-wise), allowing to change not only the appearance of the whole scene, but also of specific semantic classes or any chosen areas (see Fig. 2). We call our model OASIS, as it needs **o**nly **a**dversarial supervision for **s**emantic **i**mage **s**ynthesis.

In summary, our main contributions are: (1) We propose a novel segmentation-based discriminator architecture, that gives more powerful feedback to the generator and eliminates the necessity of the perceptual loss supervision. (2) We present a simple 3D noise sampling scheme, notably increasing the diversity of multi-modal synthesis and enabling complete or partial change of the generated image. (3) With the OASIS model, we achieve high quality results on the ADE20K, Cityscapes and COCO-stuff datasets, on average improving the state of the art by 6 FID and 5 mIoU points, while

Figure 2: OASIS multi-modal synthesis results. The 3D noise can be sampled globally (first 2 rows), changing the whole scene, or locally (last 2 rows), partially changing the image. For the latter, we sample different noise per region, like the bed segment (in red) or arbitrary areas defined by shapes.

relying only on adversarial supervision. We show that images synthesized by OASIS exhibit much higher diversity and more closely follow the color and texture distributions of real images. Our code and pretrained models are available at `https://github.com/boschresearch/OASIS`.

2 Related work

Semantic image synthesis. Pix2pix (Isola et al., 2017) first proposed to use conditional GANs (Mirza & Osindero, 2014) for semantic image synthesis, adopting an encoder-decoder generator which takes semantic label maps as input, and employing a PatchGAN discriminator. Since then, various generator and discriminator modifications have been introduced (Wang et al., 2018; Park et al., 2019; Liu et al., 2019; Tang et al., 2020c;b; Ntavelis et al., 2020). Besides GANs, Chen & Koltun (2017) proposed to use a cascaded refinement network (CRN) for high-resolution semantic image synthesis, and SIMS (Qi et al., 2018) extended it with a non-parametric component, serving as a memory bank of source material to assist the synthesis. Further, Li et al. (2019) employed implicit maximum likelihood estimation (Li & Malik, 2018) to increase the variety of the CRN model. However, these approaches still underperform in comparison to state-of-the-art GAN models. Therefore, next we focus on the recent GAN architectures for semantic image synthesis.

Discriminator architectures. Pix2pix (Isola et al., 2017), Pix2pixHD (Wang et al., 2018) and SPADE (Park et al., 2019) all employed a multi-scale PatchGAN discriminator, that takes an image and its semantic label map as input. CC-FPSE (Liu et al., 2019) proposed a feature-pyramid discriminator, embedding both images and label maps into a joint feature map, and then consecutively upsampling it in order to classify it as real/fake at multiple scales. LGGAN (Tang et al., 2020c) introduced a classification-based feature learning module to learn more discriminative and class-specific features. In this work, we propose to use a pixel-wise semantic segmentation network as a discriminator instead of multi-scale image classifiers as in the above approaches, and to directly exploit the semantic label maps for its supervision. Segmentation-based discriminators have been shown to improve semantic segmentation (Souly et al., 2017) and unconditional image synthesis (Schönfeld et al., 2020), but to the best of our knowledge have not been explored for semantic image synthesis and our work is the first to apply adversarial semantic segmentation loss for this task.

Generator architectures. Conventionally, the semantic label map is provided to the image generation pipeline via an encoder (Isola et al., 2017; Wang et al., 2018; Tang et al., 2020c;b; Ntavelis et al., 2020). However, it is shown to be suboptimal at preserving the semantic information until the later stages of image generation. Therefore, SPADE introduced a spatially-adaptive normalization layer

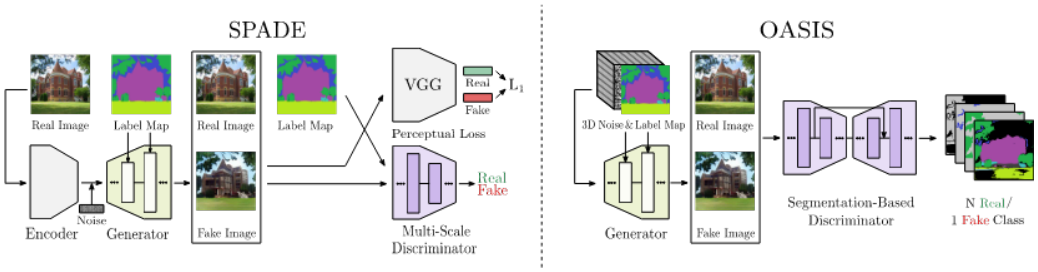

Figure 3: SPADE (left) vs. OASIS (right). OASIS outperforms SPADE, while being simpler and lighter: it uses only adversarial loss supervision and a single segmentation-based discriminator, without relying on heavy external networks. Furthermore, OASIS learns to synthesize multi-modal outputs by directly re-sampling the 3D noise tensor, instead of using an image encoder as in SPADE.

that directly modulates the label map onto the generator's hidden layer outputs at various scales. Alternatively, CC-FPSE proposed to use spatially-varying convolution kernels conditioned on the label map. Struggling with generating diverse images from noise, both Pix2pixHD and SPADE resorted to having an image encoder in the generator design to enable multi-modal synthesis. The generator then combines the extracted image style with the label map to reconstruct the original image. By alternating the style vector, one can generate multiple outputs conditioned on the same label map. However, using an image encoder is a resource demanding solution. In this work, we enable multi-modal synthesis directly through sampling of a 3D noise tensor injected at every layer of the network. Differently from structured noise injection of Alharbi & Wonka (2020) and class-specific latent codes of Zhu et al. (2020), we inject the 3D noise along with label maps and adjust it to image resolution, also enabling re-sampling of selected semantic segments (see Fig. 2).

Perceptual losses. Gatys et al. (2015); Gatys et al. (2016); Johnson et al. (2016) and Bruna et al. (2016) were pioneers at exploiting perceptual losses to produce high-quality images for super-resolution and style transfer using convolutional networks. For semantic image synthesis, the VGG-based perceptual loss was first introduced by CRN, and later adopted by Pix2pixHD. Since then, it has become a default for training the generator (Park et al., 2019; Liu et al., 2019; Tan et al., 2020; Tang et al., 2020a). As the perceptual loss is based on a VGG network pre-trained on ImageNet (Deng et al., 2009), methods relying on it are constrained by the ImageNet domain and the representational power of VGG. With the recent progress on GAN training, e.g. by architecture designs and regularization techniques, the actual necessity of the perceptual loss requires a reassessment. We experimentally show that such loss imposes unnecessary constraints on the generator, significantly limiting sample diversity. While our model, trained without the VGG loss, achieves improved image diversity while not compromising image quality.

3 OASIS MODEL

In this section, we present our OASIS model, which, in contrast to other semantic image synthesis methods, needs only adversarial supervision for generator training. Using SPADE as a starting point (Sec. 3.1), we first propose to re-design the discriminator as a semantic segmentation network, directly using the given semantic label maps as ground truth (Sec. 3.2). Empowered by spatially- and semantically-aware feedback of the new discriminator, we next re-design the SPADE generator, enabling its effective multi-modal synthesis via 3D noise sampling (Sec. 3.3).

3.1 THE SPADE BASELINE

We choose SPADE as our baseline as it is a state-of-the-art model and a relatively simple representative of conventional semantic image synthesis models. As depicted in Fig. 3, the discriminator of SPADE largely follows the PatchGAN multi-scale discriminator (Isola et al., 2017), adopting two image classification networks operating at different resolutions. Both of them take the channel-wise concatenation of the semantic label map and the real/synthesized image as input, and produce true/-fake classification scores. On the generator side, SPADE adopts spatially-adaptive normalization layers to effectively integrate the semantic label map into the synthesis process from low to high scales. Additionally, the image encoder is used to extract the style vector from the reference image and then combine it with a 1D noise vector for multi-modal synthesis. The training loss of SPADE

consists of three terms, namely, an adversarial loss, a feature matching loss and the VGG-based perceptual loss: $\mathcal{L} = \max_G \min_D \mathcal{L}_{\text{adv}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}}$. Overall, SPADE is a resource demanding model at both training and test time, i.e., with two PatchGAN discriminators, an image encoder in addition to the generator, and the VGG loss. In the following, we revisit its architecture and introduce a simpler and more efficient model that offers better performance with less complexity.

3.2 OASIS DISCRIMINATOR

For the generator to learn to synthesize images that are well aligned with the input semantic label maps, we need a powerful discriminator that coherently captures discriminative semantic features at different image scales. While classification-based discriminators, such as PatchGAN, take label maps as input concatenated to images, they can afford to ignore them and make the decision solely on image patch realism. Thus, we propose to cast the discriminator task as a multi-class semantic segmentation problem to directly utilize label maps for supervision, and accordingly alter its architecture to an encoder-decoder segmentation network (see Fig. 3). Encoder-decoder networks have proven to be effective for semantic segmentation (Badrinarayanan et al., 2016; Chen et al., 2018). Thus, we build our discriminator architecture upon U-Net (Ronneberger et al., 2015), which consists of the encoder and decoder connected by skip connections. This discriminator architecture is multi-scale through its design, integrating information over up- and down-sampling pathways and through the encoder-decoder skip connections. For details on the architecture see App. C.1.

The segmentation task of the discriminator is formulated to predict the per-pixel class label of the real images, using the given semantic label maps as ground truth. In addition to the N semantic classes from the label maps, all pixels of the fake images are categorized as one extra class. Overall, we have $N+1$ classes in the semantic segmentation problem, and thus propose to use a (N+1)-class cross-entropy loss for training. Considering that the N semantic classes are usually imbalanced and that the per-pixel size of objects varies for different semantic classes, we weight each class by its inverse per-pixel frequency, giving rare semantic classes more weight. In doing so, the contributions of each semantic class are equally balanced, and, thus, the generator is also encouraged to adequately synthesize less-represented classes. Mathematically, the new discriminator loss is expressed as:

$$\mathcal{L}_D = -\mathbb{E}_{(x,t)}\left[\sum_{c=1}^{N}\alpha_c \sum_{i,j}^{H\times W} t_{i,j,c}\log D(x)_{i,j,c}\right] - \mathbb{E}_{(z,t)}\left[\sum_{i,j}^{H\times W}\log D(G(z,t))_{i,j,c=N+1}\right], \quad (1)$$

where x denotes the real image; (z,t) is the noise-label map pair used by the generator G to synthesize a fake image; and the discriminator D maps the real or fake image into a per-pixel (N+1)-class prediction probability. The ground truth label map t has three dimensions, where the first two correspond to the spatial position $(i,j) \in H \times W$, and the third one is a one-hot vector encoding the class $c \in \{1,..,N{+}1\}$. The class balancing weight α_c is the inverse of the per-pixel class frequency

$$\alpha_c = \frac{H \times W}{\sum_{i,j}^{H\times W} E_t\left[\mathbb{1}[t_{i,j,c}=1]\right]}. \quad (2)$$

LabelMix regularization. In order to encourage our discriminator to focus on differences in content and structure between the fake and the real classes, we propose a LabelMix regularization. Based on the semantic layout, we generate a binary mask M to mix a pair $(x,\hat{x})$ of real and fake images conditioned on the same label map: $\text{LabelMix}(x,\hat{x},M) = M \odot x + (1-M)\odot\hat{x}$, as visualized in Fig. 4. Given the mixed image, we further train the discriminator to be equivariant under the LabelMix operation. This is achieved by adding a consistency loss term $\mathcal{L}_{cons}$ to Eq. 1:

$$\mathcal{L}_{cons} = \Big\| D_{\text{logits}}\Big(\text{LabelMix}(x,\hat{x},M)\Big) - \text{LabelMix}\Big(D_{\text{logits}}(x), D_{\text{logits}}(\hat{x}), M\Big)\Big\|^2, \quad (3)$$

where D_{logits} are the logits attained before the last softmax activation layer, and $\|\cdot\|$ is the L_2 norm. This consistency loss compares the output of the discriminator on the LabelMix image with the LabelMix of its outputs, penalizing the discriminator for inconsistent predictions. LabelMix is different to CutMix (Yun et al., 2019), which randomly samples the binary mask M. A random mask will introduce inconsistency between the pixel-level classes and the scene layout provided by the label map. For an object with the semantic class c, it will contain pixels from both real and fake images, resulting in two labels, i.e. c and $N+1$. To avoid such inconsistency, the mask of LabelMix is generated according to the label map, providing natural borders between semantic regions, see Fig. 4 (Mask M). Under LabelMix regularization, the generator is encouraged to respect the natural semantic boundaries, improving pixel-level realism while also considering the class segment shapes.

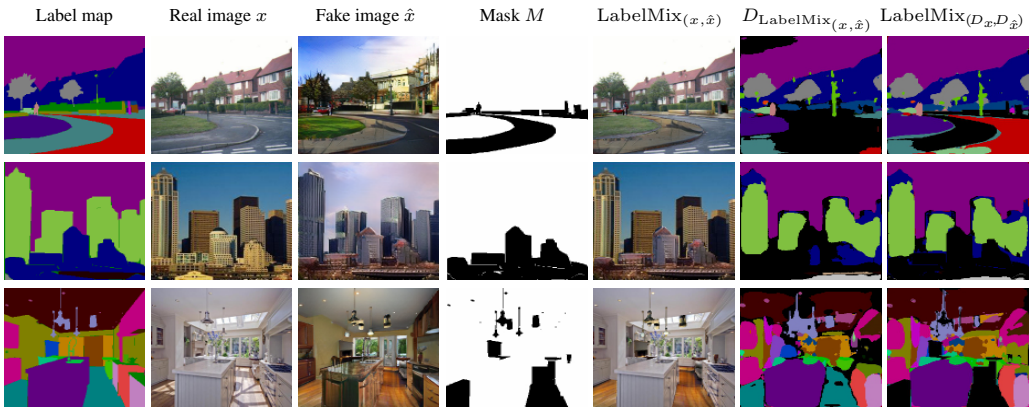

Figure 4: LabelMix regularization. Real x and fake $\hat{x}$ images are mixed using a binary mask M, sampled based on the label map, resulting in $\text{LabelMix}_{(x,\hat{x})}$. The consistency regularization then minimizes the L2 distance between the logits of $D_{\text{LabelMix}_{(x,\hat{x})}}$ and $\text{LabelMix}_{(D_x,D_{\hat{x}})}$. In this visualization, **black** corresponds to the fake class in the N+1 segmentation output.

Other variants. Besides the proposed (N+1)-class cross entropy loss, there are other ways to train the segmentation-based discriminator with the label map. One can concatenate the label map to the input image, analogous to SPADE. Another option is to use projection, by taking the inner product between the last linear layer output and the embedded label map, analogous to class-label conditional GANs (Miyato & Koyama, 2018). For both alternatives, the training loss is pixel-level real/fake binary cross-entropy (Schönfeld et al., 2020). From the label map encoding perspective, these two variants use labels map as input (concatenated to image or at last linear layer), propagating it *forward* through the network. The (N+1)-setting uses the label map for loss computation, so it is propagated *backward* via gradient updates. Backward propagation ensures that the discriminator learns semantic-aware features, in contrast to forward propagation, where the label map alignment is not as strongly enforced. Performance comparison of the label map encodings is shown in Table 5.

3.3 OASIS GENERATOR

To stay in line with the OASIS discriminator design, the training loss for the generator is changed to

$$\mathcal{L}_G = -\mathbb{E}_{(z,t)}\left[\sum_{c=1}^{N} \alpha_c \sum_{i,j}^{H\times W} t_{i,j,c} \log D(G(z,t))_{i,j,c}\right], \tag{4}$$

which is a direct outcome of the non-saturation trick (Goodfellow et al., 2014) to Eq. 1. We next re-design the generator to enable multi-modal synthesis through noise sampling. SPADE is deterministic in its default setup, but can be trained with an extra image encoder to generate multi-modal outputs. We introduce a simpler version, that enables synthesis of diverse outputs directly from input noise. For this, we construct a noise tensor of size $64\times H\times W$, matching the spatial dimensions of the label map $H\times W$. Channel-wise concatenation of the noise and label map forms a 3D tensor used as input to the generator and also as a conditioning at every spatially-adaptive normalization layer. In doing so, intermediate feature maps are conditioned on both the semantic labels and the noise (see Fig. 3). With such a design, the generator produces diverse, noise-dependent images. As the 3D noise is channel- and pixel-wise sensitive, at test time, one can sample the noise globally, per-channel, and locally, per-segment or per-pixel, for controlled synthesis of the whole scene or of specific semantic objects. For example, when generating a scene of a bedroom, one can re-sample the noise locally and change the appearance of the bed alone (see Fig. 2). Note that for simplicity during training we sample the 3D noise tensor globally, i.e. per-channel, replicating each channel value spatially along the height and width of the tensor. We analyse alternative ways of sampling 3D noise during training in App. A.7. Using image styles via an encoder, as in SPADE, is also possible in our setting, by replacing noise with encoder features. Lastly, to further reduce the complexity, we remove the first residual block in the generator, reducing the number of parameters from 96M to 72M (see App. C.2) without a noticeable performance loss (see Table 3).

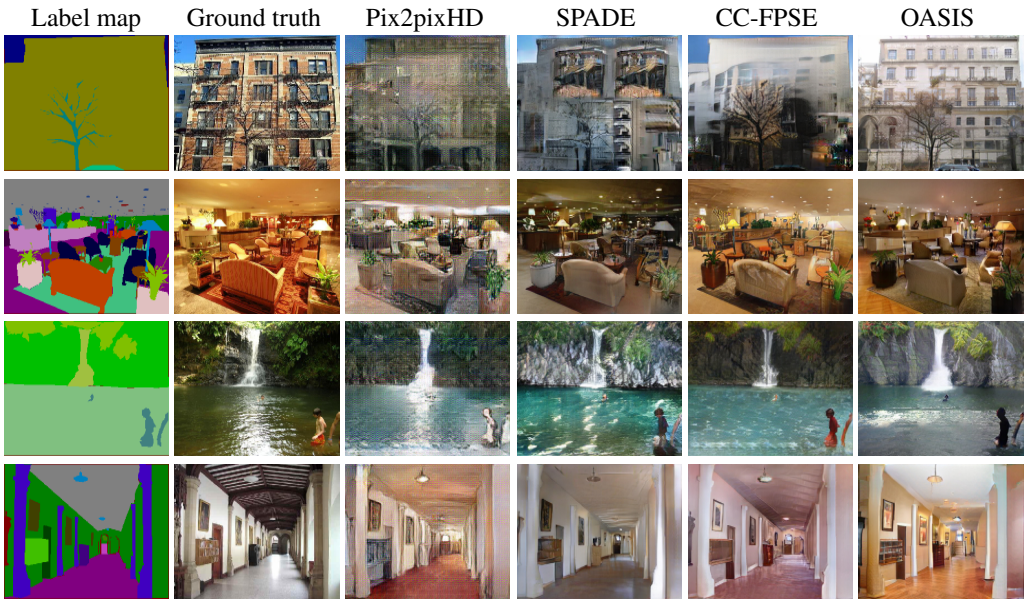

Figure 5: Qualitative comparison of OASIS with other methods on ADE20K. Trained with only adversarial supervision, our model generates images with better perceptual quality and structure.

4 EXPERIMENTS

We conduct experiments on three challenging datasets: ADE20K (Zhou et al., 2017), COCO-stuff (Caesar et al., 2018) and Cityscapes (Cordts et al., 2016). Following Qi et al. (2018), we also evaluate OASIS on ADE20K-outdoors, a subset of ADE20K containing outdoor scenes. We follow the experimental setting of Park et al. (2019). We did not use the GAN feature matching loss for OASIS, as we did not observe any improvement with it (see App. A.5), and used the VGG loss only for ablations with $\lambda_{\text{VGG}} = 10$. We did not experience any training instabilities and, thus, did not employ any extra stabilization techniques. All our models use an exponential moving average (EMA) of the generator weights with 0.9999 decay. For further training details refer to App. C.3.

Following prior work (Isola et al., 2017; Wang et al., 2018; Park et al., 2019; Liu et al., 2019), we evaluate models quantitatively on the validation set using the Fréchet Inception Distance (FID) (Heusel et al., 2017) and mean Intersection-over-Union (mIoU). FID is known to be sensitive to both quality and diversity and has been shown to be well aligned with human judgement (Heusel et al., 2017). We show additional evaluation of quality and diversity with "improved precision and recall" in App. A.9. Mean IoU is used to assess the alignment of the generated image with the ground truth label map, computed via a pre-trained semantic segmentation network. We use UperNet101 (Xiao et al., 2018) for ADE20K, multi-scale DRN-D-105 (Yu et al., 2017) for Cityscapes, and DeepLabV2 (Chen et al., 2015) for COCO-Stuff. We additionally propose to compare color and texture statistics between generated and real images on ADE20K to better understand how the perceptual loss influences performance. For this, we compute color histograms in LAB space and measure the earth mover's distance between the real and generated sets (Rubner et al., 2000). We measure the texture similarity to the real data as the χ^2-distance between Local Binary Patterns histograms (Ojala et al., 1996). As different classes have different color and texture distributions, we aggregate histogram distances separately per class and then take the mean. Lower values for the texture and color distances indicate a closer similarity to real data.

4.1 MAIN RESULTS

We use SPADE as our baseline, using the authors' implementation[1]. For a fair comparison, we train this model without the feature matching loss and using EMA (Yaz et al., 2018) at test phase, which

[1] `github.com/NVlabs/SPADE`

Table 1: Comparison with other methods across datasets.Bold denotes the best performance.

Method	# param	VGG	ADE20K		ADE-outd.		Cityscapes		COCO-stuff	
			FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
CRN	84M	✓	73.3	22.4	99.0	16.5	104.7	52.4	70.4	23.7
SIMS	56M	✓	n/a	n/a	67.7	13.1	49.7	47.2	n/a	n/a
Pix2pixHD	183M	✓	81.8	20.3	97.8	17.4	95.0	58.3	111.5	14.6
LGGAN	n/a	✓	31.6	41.6	n/a	n/a	57.7	68.4	n/a	n/a
CC-FPSE	131M	✓	31.7	43.7	n/a	n/a	54.3	65.5	19.2	41.6
SPADE	102M	✓	33.9	38.5	63.3	30.8	71.8	62.3	22.6	37.4
SPADE+	102M	✓	32.9	42.5	51.1	32.1	47.8	64.0	21.7	38.8
		✗	60.7	21.0	65.4	22.7	61.4	47.6	99.1	16.1
OASIS	94M	✗	**28.3**	**48.8**	**48.6**	**40.4**	**47.7**	**69.3**	**17.0**	**44.1**

Table 2: Multi-modal synthesis evaluation on ADE20K. Bold and red denote the best and the worst performance, respectively.

Method	Multi-mod.	VGG	MS-SSIM↓	LPIPS↑	FID↓	mIoU↑
SPADE+	Encoder	✓	0.85	0.16	33.4	40.2
SPADE+	3D noise	✗	**0.35**	**0.50**	58.4	18.7
		✓	0.53	0.36	34.4	36.2
OASIS	3D noise	✗	0.65	0.35	**28.3**	48.8
		✓	0.88	0.15	31.6	**50.8**

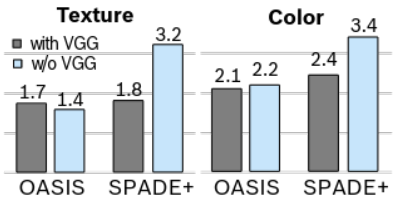

Figure 6: Histogram distances to real data.

we further refer to as SPADE+. We found that the feature matching loss has a negligible impact (see App. A.5), while EMA significantly increases the performance for all metrics (see Table 1).

OASIS outperforms the current state of the art on all datasets with an average improvement of 6 FID and 5 mIoU points (Table 1). Importantly, OASIS achieved the improvement via adversarial supervision alone. On the contrary, SPADE+ does not produce images of high visual quality without the perceptual loss, and struggles to learn the color and texture distribution of real images (Fig. 6). A strong discriminator is the key factor for good performance: without a rich training signal from the discriminator, the SPADE+ generator has to learn through minimizing the VGG loss. With the stronger OASIS discriminator, the perceptual loss does not overtake the generator supervision (see App. A.2), allowing to produce images with the color and texture distribution closer to the real data.

Fig. 5 shows a qualitative comparison of our results to previous models. Our approach noticeably improves image quality, synthesizing finer textures and more natural colors. With the powerful feedback from the discriminator, OASIS is able to learn the appearance of small or rarely occurring semantic classes (which is reflected in the per-class IoU scores presented in App. A.3), producing plausible results even for complex scenes with rare classes and reducing unnatural artifacts.

Multi-modal image synthesis. In contrast to previous work, OASIS can produce diverse images by directly re-sampling input 3D noise. As 3D noise modulates features directly at every layer of the generator at different scales, matching their resolution, it affects both global and local characteristics of the image. Thus, the noise can be sampled globally, varying the whole image, or locally, resulting in the selected object change while preserving the rest of the scene (see Fig. 2).

To measure the variation in the multi-modal generation, we evaluate MS-SSIM (Wang et al., 2003) and LPIPS (Zhang et al., 2018b) between images generated from the same label map. We generate 20 images and compute the mean pairwise scores, and then average over all label maps. The lower the MS-SSIM and the higher the LPIPS scores, the more diverse the generated images are. To assess the effect of the perceptual loss and the noise sampling on diversity, we train SPADE+ with 3D noise or the image encoder, and with or without the perceptual loss. Table 2 shows that OASIS, without perceptual loss, improves over SPADE+ with the image encoder, both in terms of image diversity (MS-SSIM, LPIPS) and quality (mean FID, mIoU across 20 realizations). Using 3D noise further increases diversity for SPADE+. However, a strong quality-diversity tradeoff exists for SPADE+: 3D noise improves diversity at the cost of quality, and the perceptual loss improves quality at the cost of diversity. For OASIS, the VGG loss also reduces diversity but does not noticeably affect quality. Note that in our experiments LabelMix does not notably affect diversity (see App. A.1).

4.2 ABLATIONS

We conduct ablations on ADE20K to evaluate our proposed changes. The main ablation shows the impact of our new discriminator, lighter generator, LabelMix and 3D noise. Further ablations are concerned with architecture changes and the label map encodings in the discriminator, where for fair comparison we use no 3D noise and LabelMix.

Main ablation. Table 3 shows that SPADE+ scores low on the image quality metrics without the perceptual loss. Replacing the SPADE+ discriminator with the OASIS discriminator, while keeping the generator fixed, improves FID and mIoU by more than 30 points. Changing the SPADE+ generator to the lighter OASIS generator leads to a negligible degradation of 0.3 in FID and 0.5 in mIoU. With LabelMix FID improves further by ~ 1 point (more ablations on LabelMix in App. A.4). Adding 3D noise improves FID but degrades mIoU, as diversity complicates the task of the pre-trained semantic segmentation network used to compute the score. For OASIS the perceptual loss deteriorates FID by more than 2 points, but improves mIoU. Overall, without the perceptual loss the new discriminator is the key to the performance boost over SPADE+.

Table 3: OASIS ablation on ADE20K. Bold denotes the best performance.

G	D	VGG	LabelMix	FID↓	mIoU↑
SPADE+	SPADE+	✗	✗	60.7	21.0
SPADE+	OASIS	✗	✗	29.0	**52.1**
OASIS	OASIS	✗	✗	29.3	51.6
		✗	✓	28.4	50.6
OASIS	OASIS	✗	✓	**28.3**	48.8
+3D noise		✓	✓	31.6	50.8

Ablation on the discriminator architecture. We train the OASIS generator with three alternative discriminators: the original multi-scale PatchGAN consisting of two networks, a single-scale PatchGAN, and a ResNet-based discriminator, corresponding to the encoder of the U-Net shaped OASIS discriminator. Table 4 shows that the alternative discriminators only perform well with perceptual supervision, while the OASIS discriminator achieves superior performance independent of it. The single-scale discriminators even collapse without the perceptual loss (highlighted in red in Table 4).

Table 4: Ablation on the D architecture. Bold denotes the best performance, red highlights collapsed runs.

D architecture	w/o VGG		with VGG	
	FID↓	mIoU↑	FID↓	mIoU↑
MS-PatchGAN (2x)	60.7	21.0	32.9	42.5
PatchGAN	197	0.62	34.2	42.2
ResNet-PatchGAN	147	0.42	32.4	45.1
OASIS	**29.3**	**51.6**	**29.2**	**51.1**

Ablation on the label map encoding. We study four different label map encodings: input concatenation, as in SPADE, projection conditioned on the label map (Miyato & Koyama, 2018), employing label maps as ground truth for the N+1 segmentation loss, or for the class-balanced N+1 loss (see Sec. 3.2). As shown in Table 5, input concatenation is not sufficient without additional perceptual loss supervision, leading to training collapse. Without perceptual loss, the N+1 loss outperforms the input concatenation and the projection in both the FID and mIoU metrics. The class balancing noticeably improves mIoU due to better supervision for rarely occurring semantic classes. More ablations can be found in App. A.

Table 5: Ablation on the label map encoding. Bold denotes the best performance, red highlights collapsed runs.

Label encoding	w/o VGG		with VGG	
	FID↓	mIoU↑	FID↓	mIoU↑
Input concatenation	280	0.02	30.0	43.9
Projection	32.4	44.9	**28.0**	46.9
N+1 loss	**28.3**	47.2	28.6	49.8
Balanced N+1 loss	29.3	**51.6**	29.2	**51.1**

5 CONCLUSION

In this work we propose OASIS, a semantic image synthesis model that only relies on adversarial supervision to achieve high fidelity image synthesis. In contrast to previous work, our model eliminates the need for a perceptual loss, which often imposes extra constraints on image quality and diversity. This is achieved via detailed spatial and semantic-aware supervision from our novel segmentation-based discriminator, which uses semantic label maps as ground truth for training. With this powerful discriminator, OASIS can easily generate diverse multi-modal outputs by re-sampling the 3D noise, both globally and locally, allowing to change the appearance of the whole scene and of individual objects. OASIS significantly improves over the state of the art in terms of image quality and diversity, while being simpler and more lightweight than previous methods.

ACKNOWLEDGEMENT

Jürgen Gall has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2070 -390732324.

REFERENCES

Yazeed Alharbi and Peter Wonka. Disentangled image generation through structured noise injection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2016.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. In *International Conference on Learning Representations (ICLR)*, 2016.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *International Conference on Learning Representations (ICLR)*, 2015.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *International Conference on Computer Vision (ICCV)*, 2017.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPs)*. 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, pp. 3927–3936, 2019.

Ke Li and Jitendra Malik. Implicit maximum likelihood estimation. *arXiv preprint arXiv:1809.09087*, 2018.

Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image synthesis from semantic layouts via conditional imle. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4220–4229, 2019.

Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations (ICLR)*, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. *arXiv:2004.04977*, 2020.

Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In *Advances in Neural Information Processing Systems*, pp. 12268–12279, 2019.

Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine learning (ICML)*, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision (IJCV)*, 2000.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pp. 5228–5237, 2018.

Edgar Schönfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *International Conference on Computer Vision (ICCV)*, 2017.

Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, and Nenghai Yu. Rethinking spatially-adaptive normalization. *arXiv:2004.02867*, 2020.

Hao Tang, Song Bai, and Nicu Sebe. Dual attention gans for semantic image synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1994–2002, 2020a.

Hao Tang, Xiaojuan Qi, Dan Xu, Philip HS Torr, and Nicu Sebe. Edge guided gans with semantic preserving for semantic image synthesis. *arXiv:2003.13898*, 2020b.

Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020c.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, 2003.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, 2018.

Yasin Yaz, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, Vijay Chandrasekhar, et al. The unusual effectiveness of averaging in gan training. In *International Conference on Learning Representations*, 2018.

Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019.

Dan Zhang and Anna Khoreva. PA-GAN: Improving GAN training by progressive augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *Transactions on Pattern Analysis and Machine Intelligence*, 2018a.

Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine learning (ICML)*, 2019.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018b.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

B OASIS: Only Adversarial Supervision for Semantic Image Synthesis

In this appendix, we provide the journal publication that Chapter 4 of the thesis is based on:

- **OASIS: Only Adversarial Supervision for Semantic Image Synthesis**
  Vadim Sushko*, Edgar Schönfeld*, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva
  International Journal of Computer Vision (IJCV), 2022.
  DOI: 10.1007/s11263-022-01673-x

International Journal of Computer Vision (2022) 130:2903–2923
https://doi.org/10.1007/s11263-022-01673-x

OASIS: Only Adversarial Supervision for Semantic Image Synthesis

Vadim Sushko[1] · Edgar Schönfeld[1] · Dan Zhang[1,2] · Juergen Gall[3] · Bernt Schiele[4] · Anna Khoreva[1,2]

Received: 6 May 2022 / Accepted: 11 August 2022 / Published online: 17 September 2022

Abstract
Despite their recent successes, generative adversarial networks (GANs) for semantic image synthesis still suffer from poor image quality when trained with only adversarial supervision. Previously, additionally employing the VGG-based perceptual loss has helped to overcome this issue, significantly improving the synthesis quality, but at the same time limited the progress of GAN models for semantic image synthesis. In this work, we propose a novel, simplified GAN model, which needs only adversarial supervision to achieve high quality results. We re-design the discriminator as a semantic segmentation network, directly using the given semantic label maps as the ground truth for training. By providing stronger supervision to the discriminator as well as to the generator through spatially- and semantically-aware discriminator feedback, we are able to synthesize images of higher fidelity and with a better alignment to their input label maps, making the use of the perceptual loss superfluous. Furthermore, we enable high-quality multi-modal image synthesis through global and local sampling of a 3D noise tensor injected into the generator, which allows complete or partial image editing. We show that images synthesized by our model are more diverse and follow the color and texture distributions of real images more closely. We achieve a strong improvement in image synthesis quality over prior state-of-the-art models across the commonly used ADE20K, Cityscapes, and COCO-Stuff datasets using only adversarial supervision. In addition, we investigate semantic image synthesis under severe class imbalance and sparse annotations, which are common aspects in practical applications but were overlooked in prior works. To this end, we evaluate our model on LVIS, a dataset originally introduced for long-tailed object recognition. We thereby demonstrate high performance of our model in the sparse and unbalanced data regimes, achieved by means of the proposed 3D noise and the ability of our discriminator to balance class contributions directly in the loss function. Our code and pretrained models are available at https://github.com/boschresearch/OASIS.

Keywords Semantic image synthesis · GAN · Semantic segmentation · Label-to-image translation · Image editing

1 Introduction

Conditional generative adversarial networks (GANs) (Mirza & Osindero, 2014) synthesize images conditioned on class labels (Brock et al., 2019; Casanova et al., 2021), text (Reed et al., 2016; Zhang et al., 2018a, 2021), other images (Isola et al., 2017; Huang et al., 2018; Park et al., 2020), or semantic label maps (Park et al., 2019b; Liu et al., 2019; Wang et al., 2021b). In this work, we focus on the latter, addressing semantic image synthesis. Taking pixel-level annotated semantic maps as input, semantic image synthesis enables the rendering of realistic images from user-specified layouts,

Communicated by Arun Mallya.

Vadim Sushko and Edgar Schönfeld have contributed equally to this work.

Vadim Sushko
vadim.sushko@bosch.com

Edgar Schönfeld
edgar.schoenfeld@bosch.com

Dan Zhang
dan.zhang2@bosch.com

Juergen Gall
gall@iai.uni-bonn.de

Bernt Schiele
schiele@mpi-inf.mpg.de

Anna Khoreva
anna.khoreva@bosch.com

[1] Bosch Center for Artificial Intelligence, Renningen, Germany

[2] University of Tübingen, Tübingen, Germany

[3] University of Bonn, Bonn, Germany

[4] Max Planck Institute for Informatics, Saarbrücken, Germany

Fig. 1 Existing semantic image synthesis models heavily rely on the VGG-based perceptual loss to improve the quality of generated images. In contrast, our model (OASIS) can synthesize diverse and high-quality images while only using an adversarial loss, without any external supervision

without the use of an intricate graphics engine. Therefore, its applications range widely from content creation and image editing to producing training data for downstream applications that adhere to specific semantic requirements (Park et al., 2019a; Ntavelis et al., 2020).

Despite the recent progress on stabilizing GANs (Miyato et al., 2018; Zhang & Khoreva, 2019; Karras et al., 2020a; Sauer et al., 2021) and developing their architectures (Karras et al., 2021, 2019, 2020b; Brock et al., 2019; Liu et al., 2021), state-of-the-art GAN-based semantic image synthesis models (Park et al., 2019b; Liu et al., 2019; Wang et al., 2021b) still greatly suffer from training instabilities and poor image quality when the generator is only trained to fool the discriminator in an adversarial fashion (see Fig. 1). An established practice to overcome this issue is to employ a perceptual loss (Wang et al., 2018) to train the generator, in addition to the discriminator loss. The perceptual loss aims to match intermediate features of synthetic and real images, that are estimated via an external perception network. A popular choice for such a network is VGG (Simonyan & Zisserman, 2015), pre-trained on ImageNet (Deng et al., 2009). Although the perceptual loss substantially improves the performance of previous methods, it comes with the computational overhead introduced by utilizing an extra network for training. Moreover, as we show in our experiments, it dominates over the adversarial loss during training, as the generator starts to learn mostly through minimizing the VGG loss, which has a negative impact on the diversity and quality of generated images. Therefore, in this work we propose a novel, simplified model that establishes new state-of-the-art results without requiring a perceptual loss.

To achieve semantic image synthesis of high quality, the training signal to the GAN generator should contain feedback on whether the generated images are well aligned to the input label maps. Thus, a fundamental question for GAN-based semantic image synthesis models is how to design the discriminator that would efficiently utilize information from given semantic label maps, in addition to judging the realism of given images. Conventional methods (Park et al., 2019b; Wang et al., 2018, 2021b; Liu et al., 2019; Isola et al., 2017; Ntavelis et al., 2020) adopt a multi-scale classification network, taking the label map as input along with the image, and making a global image-level real/fake decision. This discriminator has limited representation power, as it is not incentivized to learn high-fidelity pixel-level details of the images and their precise alignment with the input semantic label maps. For example, such a classification-based discriminator can base its decision solely on image realism, without the need of examining the alignment between the image and label map. To mitigate this issue, we propose an alternative architecture for the discriminator, re-designing it as an encoder-decoder semantic segmentation network (Ronneberger et al., 2015), and directly exploiting the given semantic label maps as ground truth via an $(N + 1)$-class cross-entropy loss. This new discriminator provides semantically-aware pixel-level feedback to the generator, partitioning the image into segments belonging to one of the N real semantic classes or the fake class. With this design, the network cannot ignore the provided label maps, as it has to predict a correct class label for each pixel of an image. Enabled by the discriminator per-pixel response, we further introduce a LabelMix regularization, which fosters

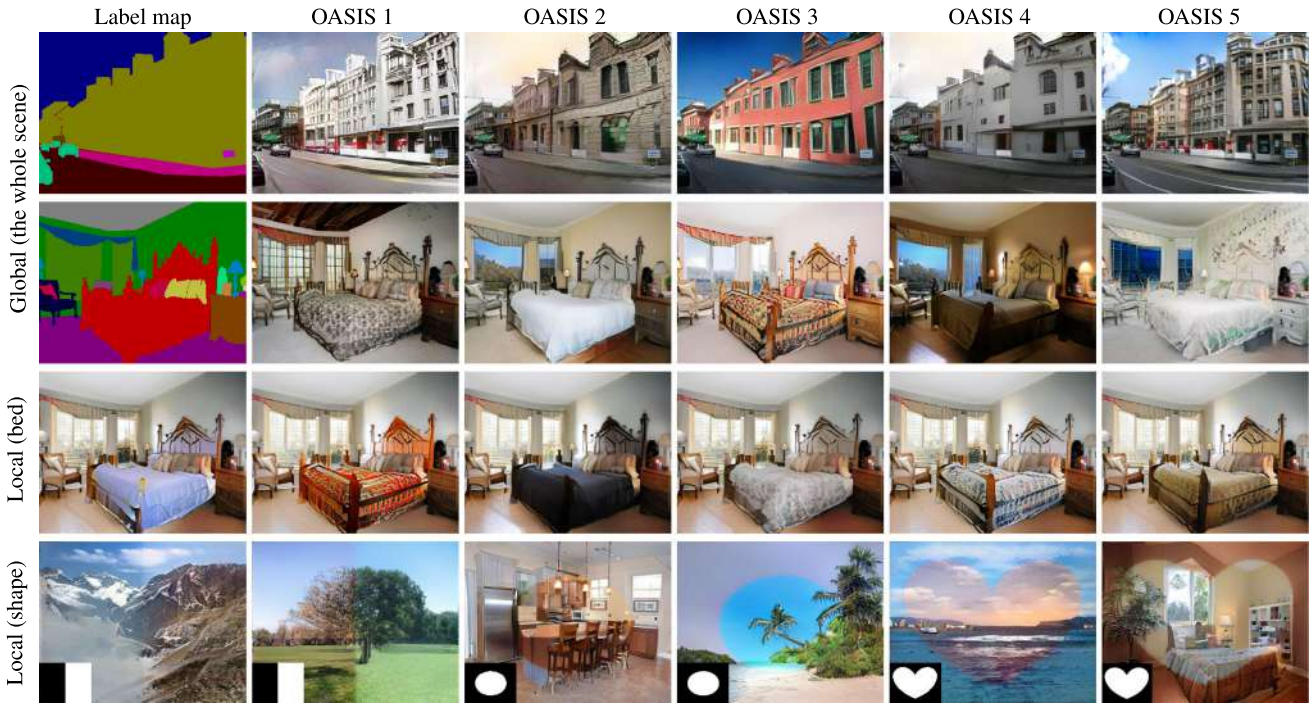

Fig. 2 OASIS multi-modal synthesis results. The 3D noise can be sampled globally (first 2 rows), changing the whole scene, or locally (last 2 rows), partially changing the image. For the latter, we sample different noise per region, like the bed segment (in red) or arbitrary areas defined by shapes (Color figure online)

the discriminator to focus more on the semantic and structural differences of real and synthetic images. The proposed changes lead to a much stronger discriminator, that maintains a powerful semantic representation of objects, giving more meaningful feedback to the generator, and thus making the perceptual loss supervision superfluous (see Fig. 1).

Semantic image synthesis is naturally a one-to-many mapping, where one label map can correspond to many possible real images. Thus, a desirable property of a generator is to generate a diverse set of images from a single label map, only by sampling noise. This property is known as multi-modality. Previously, only using a noise vector as input was not sufficient to achieve multi-modality, because the generator tended to mostly ignore the noise or synthesized images of poor quality (Isola et al., 2017; Wang et al., 2018). Thus, prior work (Wang et al., 2018; Park et al., 2019b) resorted to using an image encoder to produce multi-modal outputs. In this work, we enable multi-modal synthesis of the generator via a newly-introduced 3D noise sampling method, without requiring an image encoder and not relying on availability of a reference image to produce new image styles. Empowered by our stronger discriminator, the generator can now effectively synthesize different images by simply resampling a 3D noise tensor, which is used not only as the input, but is also combined with intermediate features via conditional normalization at every layer. This procedure makes the generator spatially sensitive to noise, so we can re-sample it both globally (channel-wise) and locally (pixel-wise), allowing to change not only the appearance of the whole scene, but also of specific semantic classes or any chosen area (see Fig. 2). As shown in our experiments, the proposed 3D noise injection scheme enables a significantly higher diversity of synthesis compared to previous methods.

With the proposed modifications in the discriminator and generator design, we outperform the prior state of the art in synthesis quality across the commonly used ADE20K (Zhou et al., 2017), COCO-Stuff (Caesar et al., 2018) and Cityscapes (Cordts et al., 2016) datasets. Omitting the necessity of the VGG perceptual loss, our model generates samples of higher quality and diversity, and follows the color and texture distributions of real images more closely.

A well known challenge for semantic segmentation applications is the problem of class imbalance. In practice, a dataset can contain underrepresented classes (representing a very small fraction of the dataset pixels), which can lead to suboptimal performance of models (Sudre et al., 2017). However, to the best of our knowledge, this problem has not been studied in the context of semantic image synthesis. For this reason, we propose to extend the evaluation setup used in previous works by using the highly imbalanced LVIS dataset (Gupta et al., 2019). Originally introduced as a dataset for long-tailed object recognition, LVIS contains

a large set of 1203 classes, the majority of which appear only in a few images. Moreover, to simplify dataset curation, label maps in LVIS were annotated sparsely, with large image areas being occupied with a generic background label. The above properties make LVIS a very challenging evaluation setting for previous semantic image synthesis models, as we demonstrate by the example of the state-of-the-art SPADE model (Park et al., 2019b). As the classification-based discriminator of SPADE makes a global real/fake decision for each image-label pair, the loss contribution originating from underrepresented classes can be dominated by the loss contribution of well represented classes. In contrast, our proposed discriminator mitigates this issue: with the $(N + 1)$-class cross-entropy loss computed for each image pixel, it becomes possible to assign higher weights for the pixels belonging to underrepresented classes. As shown in our experiments, our model successfully deals with both the extreme class imbalance and sparsity in label maps, outperforming SPADE on the LVIS dataset by a large margin.

To extend the evaluation of our model further, we test the efficacy of generated images when applied as synthetic data augmentation for the training of semantic segmentation networks. This way, the performance of semantic image synthesis is assessed through a task that holistically requires high image quality, diversity, and precise image alignment to the label maps. We demonstrate that the synthetic data produced by our model achieves high performance on this test, eliciting a notable increase in downstream segmentation performance. In doing so, our model outperforms a strong baseline SPADE (Park et al., 2019b), indicating its high potential to be applied in segmentation applications. In addition, we also demonstrate how our model for the first time enables the application of a GAN-based semantic image synthesis model to unlabelled images, without requiring external segmentation networks. Thanks to a good segmentation performance of our trained discriminator, we can infer the label map of an image and generate many alternative versions of the same scene by varying the 3D noise. We find these results promising for future utilization of our model in applications.

We call our model OASIS, as it needs **o**nly **a**dversarial supervision for **s**emantic **i**mage **s**ynthesis. In summary, our main contributions include:

- We propose a novel segmentation-based discriminator architecture, that gives more powerful feedback to the generator and eliminates the necessity of the perceptual loss supervision.
- We present a simple 3D noise sampling scheme, notably increasing the diversity of multi-modal synthesis and enabling both complete or partial resampling of a generated image.
- With the OASIS model, we achieve high-quality results on the ADE20K, Cityscapes and COCO-Stuff datasets, outperforming previous state-of-the-art models while relying only on adversarial supervision. We show that images synthesized by OASIS exhibit much higher diversity and more closely follow the color and texture distributions of real images.
- We propose to use the LVIS dataset (Gupta et al., 2019) to assess image generation in the regime with many underrepresented semantic classes, leading to a severe class imbalance. We show how the OASIS design directly addresses these issues and thereby outperforms the strong baseline SPADE (Park et al., 2019b) by a large margin.
- We test the efficacy of generated images for synthetic data augmentation, as a unified measure that simultaneously depends on image quality, diversity, and label map alignment. The images generated by OASIS elicit a stronger increase in downstream segmentation performance compared to SPADE, suggesting a higher potential of our model for future utilization in applications.

This paper is an extended version of our previous work (Schönfeld et al., 2021). Compared to the prior conference version, we provide a a significantly extended experimental evaluation and a more in-depth discussion of the proposed contributions. In particular, the conventional evaluation setup is extended to the extremely imbalanced data regime on the LVIS dataset (see Sect. 4.3). We further extend the evaluation by testing the efficacy of synthetic images as data augmentation for the task of semantic segmentation (see Sect. 4.5). We add new results on the synthesis of diverse images from unlabelled data (see Sect. 4.4 and Fig. 13). These new results highlight specific benefits of our approach compared to other models. Finally, we offer a new detailed ablation study of the method (see Tables 7, 10, 11, 12a) and extend the discussion of our model by analysing its independence on the perceptual loss (Sect. 3.4).

2 Related Work

Semantic image synthesis. The task of semantic image synthesis is to solve the inverse problem of semantic image segmentation: generate photorealistic and diverse images from provided semantic label maps. Currently, the most prominent approaches for this task are based on conditional GANs (Mirza & Osindero, 2014), as first proposed by the Pix2pix model (Isola et al., 2017). Pix2pix generates images with an encoder-decoder generator that takes label maps as input, and employs a PatchGAN discriminator which is induced to distinguish between real and fake image-label pairs. Lately, various GAN models with modified generator and discriminator architectures have been introduced (Wang et al., 2018; Park et al., 2019b; Liu et al., 2019; Tang et al., 2020c, b; Ntavelis et al., 2020; Wang et al., 2021b; Richard-

son et al., 2021; Li et al., 2021) to improve the quality and diversity of image synthesis. Besides GANs, Chen and Koltun (2017) proposed to use a cascaded refinement network (CRN) for high-resolution semantic image synthesis, and SIMS (Qi et al., 2018) extended it with a non-parametric component, serving as a memory bank of source material to assist the synthesis. Further, Li et al. (2019) employed implicit maximum likelihood estimation (Li & Malik, 2018) to increase the synthesis diversity of the CRN model. However, these approaches still underperform in comparison to state-of-the-art GAN models. Therefore, we next focus on the recent GAN architectures for semantic image synthesis.

Discriminator architectures. To provide a powerful guiding signal to the generator, a GAN discriminator for semantic image synthesis should evaluate both the image realism and its alignment to the provided semantic label map. Thus, a fundamental question is to find the most efficient way for the discriminator to utilize the given semantic label maps. To this end, Pix2pix (Isola et al., 2017), Pix2pixHD (Wang et al., 2018) and SPADE (Park et al., 2019b) rely on concatenating the label maps directly to the input image, which is fed to a multi-scale PatchGAN discriminator. Alternatively, SESAME (Ntavelis et al., 2020) employed a projection-based discriminator (Miyato & Koyama, 2018), applying an additional branch to process semantic label maps separately from images, and merging the two streams before the last convolutional layer via a pixel-wise multiplication. CC-FPSE (Liu et al., 2019) proposed a feature-pyramid discriminator, embedding both images and label maps into a joint feature map, and then consecutively upsampling it in order to classify it as real/fake at multiple scales. LGGAN (Tang et al., 2020c) introduced a classification-based feature learning module to learn more discriminative and class-specific features. In this work, we propose to use a simple pixel-wise semantic segmentation network as a discriminator instead of multi-scale image classifiers as in the above approaches, and to directly exploit the semantic label maps for its supervision. Segmentation-based discriminators have been shown to improve semantic segmentation (Souly et al., 2017) and unconditional image synthesis (Schönfeld et al., 2020), but to the best of our knowledge have not been explored for semantic image synthesis and our work is the first to apply an adversarial semantic segmentation loss for this task.

Generator architectures. To enforce the alignment between the generated images and the conditioning label maps, previous methods explored different ways to incorporate the label maps into the generator training. In many conventional approaches (Isola et al., 2017; Wang et al., 2018; Tang et al., 2020b, c; Ntavelis et al., 2020; Richardson et al., 2021), label maps are provided to the generator via an additional encoder network. However, this solution has been shown to be suboptimal at preserving the semantic information until the later stages of image generation. Therefore, SPADE introduced a spatially-adaptive normalization layer that directly modulates the label map onto the generator's hidden layer outputs at various scales. Alternatively, CC-FPSE proposed to use spatially-varying convolution kernels conditioned on the label map. Most recently, SC-GAN (Wang et al., 2021b) utilized label maps as input to generate class-specific semantic vectors at different scales, which are used as conditioning at different layers of the image rendering network; and CollageGAN (Li et al., 2021) proposed to extract a label map representation via feature pyramid encoder and inject it as spatial style tensor to a StyleGAN2 generator.

While improving the quality of generated images, the above models struggled to achieve multi-modality through sampling the input noise, as the generator tended to become insensitive to noise or achieved only poor quality, as first observed by (Isola et al., 2017). Thus, the above approaches resorted to having an image encoder in the generator design to enable multi-modal synthesis. The generator then combines the extracted image style with the label map to reconstruct the original image. By alternating the style vector, one can generate multiple outputs conditioned on the same label map. However, using an image encoder is a resource-demanding solution. In this work, we enable multi-modal synthesis directly through sampling of a 3D noise tensor which is injected at every layer of the network. Different from the structured noise injection of Alharbi and Wonka (2020) and class-specific latent codes of Zhu et al. (2020), we inject the 3D noise along with label maps and adjust it to image resolution, also enabling re-sampling of selected semantic segments (see Fig. 2).

Perceptual losses. Gatys et al. (2015, 2016); Johnson et al. (2016) and Bruna et al. (2016) were pioneers at exploiting perceptual losses to produce high-quality images for super-resolution and style transfer using convolutional networks. Such a loss extracts deep features from real and generated images by an external classification network, and minimizes their L1-distance to bring fake images closer to the real data. For semantic image synthesis, the VGG-based perceptual loss was first introduced by CRN (Chen & Koltun, 2017), and later adopted by Pix2pixHD (Isola et al., 2017). Since then, it has become a default for training the generator (Park et al., 2019b; Liu et al., 2019; Tan et al., 2020; Tang et al., 2020a; Richardson et al., 2021; Wang et al., 2021b; Li et al., 2021). As the perceptual loss is based on a VGG network pre-trained on ImageNet (Deng et al., 2009), methods relying on it are constrained by the ImageNet domain and the representational power of VGG. With the recent progress on GAN training, e.g., by architecture designs and regularization techniques, the actual necessity of the perceptual loss requires a reassessment. We experimentally show that such loss imposes unnecessary constraints on the generator, significantly limiting the diversity among samples. Trained without

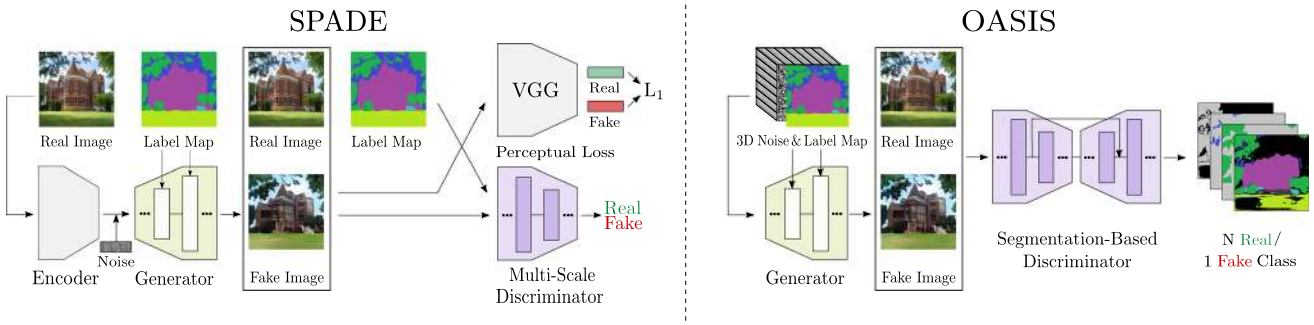

Fig. 3 SPADE (left) versus OASIS (right). OASIS outperforms SPADE, while being simpler and lighter: it uses only an adversarial loss as supervision and a single segmentation-based discriminator, without relying on heavy external networks. Furthermore, OASIS learns to synthesize multi-modal outputs by directly re-sampling the 3D noise tensor, instead of using an image encoder as in SPADE

the VGG loss, our model achieves improved diversity, at the same time not compromising the quality of generated images.

3 The OASIS Model

In this section, we present our OASIS model, which, in contrast to other semantic image synthesis methods, needs only adversarial supervision for training. Using SPADE as a starting point (Sect. 3.1), we first propose to re-design the discriminator as a semantic segmentation network, directly using the given semantic label maps as ground truth (Sect. 3.2). Empowered by spatially- and semantically-aware feedback of the new discriminator, we next re-design the SPADE generator, enabling its effective multi-modal synthesis via 3D noise sampling (Sect. 3.3). Lastly, we illustrate the superfluity of the VGG loss for our model (Sect. 3.4).

3.1 The SPADE Baseline

We choose SPADE as our baseline as it is a state-of-the-art model and a relatively simple representative of conventional semantic image synthesis models. As depicted in Fig. 3, the discriminator of SPADE largely follows the PatchGAN multi-scale discriminator (Isola et al., 2017), adopting two image classification networks operating at different resolutions. Both of them take the channel-wise concatenation of the semantic label map and the real/fake image as input, and produce real/fake classification scores. On the generator side, SPADE adopts spatially-adaptive normalization layers to effectively integrate the semantic label map into the synthesis process from low to high scales. Additionally, the image encoder is used to extract the style vector from the reference image, which is then combined with a 1D noise vector for multi-modal synthesis. The training loss of SPADE consists of three terms, namely, an adversarial loss, a feature matching loss and the VGG-based perceptual loss:

$$\mathcal{L} = \max_G \min_D \mathcal{L}_{\text{adv}} + \lambda_{\text{fm}} \mathcal{L}_{\text{fm}} + \lambda_{\text{vgg}} \mathcal{L}_{\text{vgg}}. \quad (1)$$

Overall, SPADE is a resource-demanding model at both training and test time, i.e., with two PatchGAN discriminators, an image encoder in addition to the generator, and the VGG loss. In the following, we revisit its architecture and introduce a simpler and more efficient solution that offers better performance and reduces the model complexity.

3.2 The OASIS Discriminator

To train the generator to synthesize high-quality images that are well aligned with the input semantic label maps, we need a powerful discriminator that coherently captures discriminative semantic features at different image scales. While classification-based discriminators, such as PatchGAN, take label maps as input concatenated to images, they can afford to ignore them and make the decision solely on image patch realism. Thus, we propose to cast the discriminator task as a multi-class semantic segmentation problem to directly utilize label maps for supervision, and accordingly alter its architecture to an encoder-decoder segmentation network (see Fig. 3). Encoder-decoder networks have proven to be effective for semantic segmentation (Badrinarayanan et al., 2016; Chen et al., 2018). Thus, we build our discriminator architecture upon U-Net (Ronneberger et al., 2015), which consists of the encoder and decoder connected by skip connections. This discriminator architecture is multi-scale through its design, integrating information over up- and down-sampling pathways as well as through the encoder-decoder skip connections. The segmentation task of the discriminator is formulated to predict the per-pixel class label of the real images, using the given semantic label maps as ground truth. In addition to the N semantic classes from the label maps, all pixels of fake images are categorized as one extra class. As the formulated semantic segmentation problem has $N + 1$ classes, we propose to use an $(N + 1)$-class cross-entropy loss for training.

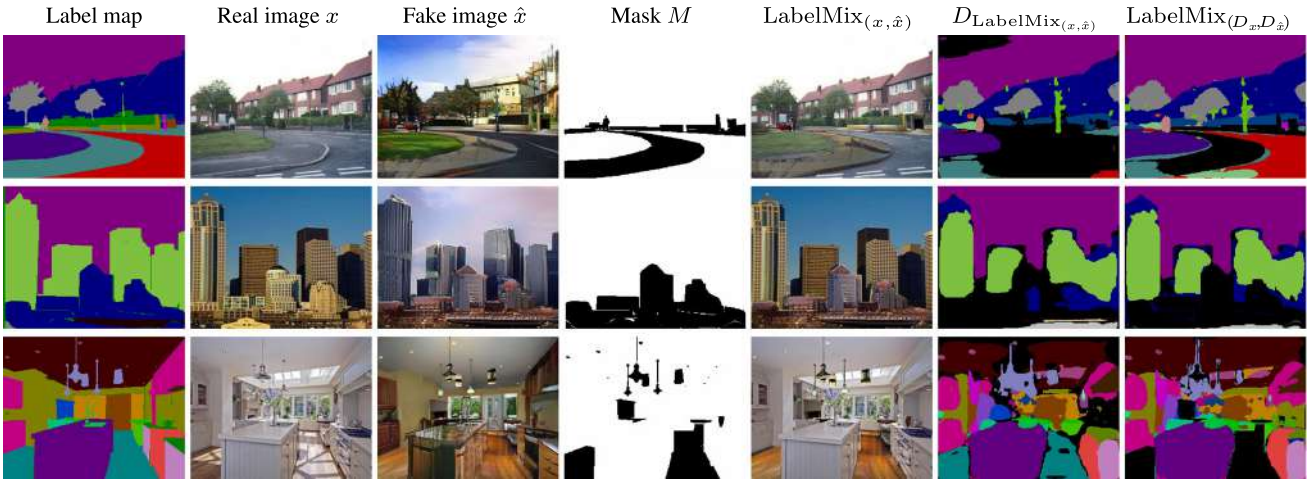

Fig. 4 LabelMix regularization. Real x and fake $\hat{x}$ images are mixed using a binary mask M, sampled based on the label map, resulting in LabelMix$_{(x,\hat{x})}$. The consistency regularization minimizes the L2 distance between the logits of $D_{\text{LabelMix}_{(x,\hat{x})}}$ and LabelMix$_{(D_x,D_{\hat{x}})}$. In this visualization, **black** corresponds to the fake class in the $N+1$ segmentation output

In practice, the N semantic classes are often imbalanced, as some of the classes represent significantly less pixels of the dataset compared to others. The loss contribution for such underrepresented classes can be dominated by well represented classes, which can lead to suboptimal performance. To mitigate this issue, empowered by the pixel-level loss computation of our discriminator, we propose to weight each class by its inverse pixel-wise frequency in a batch, thus giving underrepresented semantic classes more weight. In doing so, the loss contributions of each class are equally balanced, and, thus, the generator is also encouraged to pay more attention to underrepresented classes. Mathematically, the new discriminator loss is expressed as:

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}_{(x,t)}\left[\sum_{c=1}^{N}\alpha_c \sum_{i,j}^{H\times W} t_{i,j,c}\log D(x)_{i,j,c}\right] \\ &- \mathbb{E}_{(z,t)}\left[\sum_{i,j}^{H\times W}\log D(G(z,t))_{i,j,c=N+1}\right],\end{aligned} \tag{2}$$

where x denotes the real image; (z, t) is the noise-label map pair used by the generator G to synthesize a fake image; and the discriminator D maps the real or fake image into a per-pixel $(N + 1)$-class prediction probability. The ground truth label map t has three dimensions, where the first two correspond to the spatial position $(i, j) \in H \times W$, and the third one is a one-hot vector encoding the class $c \in \{1, .., N+1\}$. The class balancing weight α_c is the inverse pixel-wise frequency of a class c per batch:

$$\alpha_c = \frac{H \times W}{\sum_{i,j}^{H\times W} E_t\left[\mathbb{1}[t_{i,j,c}=1]\right]}. \tag{3}$$

In effect, improving the synthesis of underrepresented and well represented classes is equally necessary to minimize the loss. As we show in Sect. 4.3, this step helps to improve the synthesis quality of underrepresented classes.

LabelMix regularization. In order to encourage our discriminator to focus on differences in content and structure between the fake and real classes, we propose a LabelMix regularization. Based on the semantic layout, we generate a binary mask M to mix a pair $(x, \hat{x})$ of real and fake images conditioned on the same label map: $\text{LabelMix}(x, \hat{x}, M) = M \odot x + (1 - M) \odot \hat{x}$, as visualized in Fig. 4. Given the mixed image, we further train the discriminator to be equivariant under the LabelMix operation. This is achieved by adding a consistency loss term $\mathcal{L}_{cons}$ to Eq. 2:

$$\begin{aligned}\mathcal{L}_{cons} = &\Big\| D_{\text{logits}}\Big(\text{LabelMix}(x, \hat{x}, M)\Big) \\ &- \text{LabelMix}\Big(D_{\text{logits}}(x), D_{\text{logits}}(\hat{x}), M\Big)\Big\|^2,\end{aligned} \tag{4}$$

where D_{logits} are the logits attained before the last softmax activation layer, and $\|\cdot\|$ is the L_2 norm. This consistency loss compares the output of the discriminator on the LabelMix image with the LabelMix of its outputs, penalizing the discriminator for inconsistent predictions. LabelMix is different to CutMix (Yun et al., 2019), which randomly samples the binary mask M. A random mask will introduce inconsistency between the pixel-level labels and the scene layout provided by the label map. For an object with the class label c, it will contain pixels from both real and fake images, resulting in two labels, i.e. c and $N+1$. To avoid such inconsistency, the mask of LabelMix is generated according to the label map, providing natural borders between semantic regions, see Mask M in Fig. 4. Under LabelMix regular-

ization, the generator is encouraged to respect the natural semantic boundaries, improving pixel-level realism while also considering the class segment shapes.

Alternative ways to encode label maps. Besides the proposed $(N+1)$-class cross entropy loss, there are other ways to incorporate a label map into the training of a segmentation-based discriminator. One can concatenate the label map to the input image, analogous to SPADE. Another option is to use projection, by taking the inner product between the last linear layer output and the embedded label map, analogous to class-label conditional GANs (Miyato & Koyama, 2018). For both alternatives, the training loss is the pixel-level real/fake binary cross-entropy (Schönfeld et al., 2020). As in these two variants the label maps are used as input to the discriminator (concatenated to the input image or fed to the last linear layer), they are propagated *forward* through the network. In contrast, the (N+1)-setting uses label maps only as targets for the loss computation, so they are propagated *backward* through the network via the gradients updates. Backward propagation ensures that the discriminator learns semantic-aware features, in contrast to forward propagation, where the alignment of a generated image to the input label map can be ignored. The comparison between the above label map encodings is shown in Table 9.

3.3 The OASIS Generator

To stay in line with the OASIS discriminator design, the training loss for the generator is changed to

$$\mathcal{L}_G = -\mathbb{E}_{(z,t)}\left[\sum_{c=1}^{N}\alpha_c \sum_{i,j}^{H\times W} t_{i,j,c}\log D(G(z,t))_{i,j,c}\right], \quad (5)$$

which is a direct outcome of the non-saturation trick (Goodfellow et al., 2014) to Eq. 2. We next re-design the generator to enable multi-modal synthesis through noise sampling. SPADE is deterministic in its default setup, but can be trained with an extra image encoder to generate multi-modal outputs. We introduce a simpler version, that enables synthesis of diverse outputs directly from input noise. For this, we construct a noise tensor of size $M\times H\times W$, matching the spatial dimensions of the label map of size $N\times H\times W$, where N is the number of semantic labels and $H\times W$ corresponds to the height and width of the image. Note that for simplicity during training we sample the 3D noise tensor globally, i.e. per-channel, replicating each channel value spatially along the height and width of the tensor. In other words, a M-dimensional latent vector is sampled and then broadcasted to each pixel of an image. We analyze alternative ways of sampling 3D noise during training in the ablation section (see Sect. 4.6). After sampling, the noise and the label map are concatenated along the channel dimensions to form a combined noise-label 3D tensor of size $(M+N)\times H\times W$. This combined tensor serves as input to the first generator layer, but also as input to the spatially-adaptive normalization layers in every generator block. This way, all intermediate feature maps are conditioned on both the semantic labels and the noise (see Fig. 3), making the noise hard to ignore. As the 3D noise is channel- and pixel-wise sensitive, at test time, one can sample the noise globally, per-channel, and locally, per-segment or per-pixel, for controlled synthesis of the whole scene or of specific semantic objects. For example, when generating a scene of a bedroom, one can re-sample the noise locally and change the appearance of the bed alone (see Fig. 2).

Note that using image styles via an encoder, as in SPADE, is also possible in our setting, as the 3D noise can be simply concatenated to the encoder style features. Lastly, to further reduce the complexity, we remove the first residual block in the generator, reducing the number of parameters from 96M to 72M without a noticeable performance loss (see Table 7).

3.4 Superfluity of the Perceptual Loss for OASIS

In contrast to SPADE, which strongly relies on the perceptual loss during training (see Fig. 1), the OASIS generator is trained only with the adversarial loss from the segmentation-based discriminator, according to Eq. 5. To illustrate the insignificance of the VGG loss for OASIS, in Fig. 5 we compare the curves of the VGG and generator adversarial loss functions of SPADE and OASIS, for comparison additionally trained with the perceptual loss. We see that SPADE focuses on minimizing the VGG loss during training, but keeps the adversarial generator loss constant. Without a rich training signal from its Patch-GAN discriminator, the generator of SPADE resorts to learning mostly from the VGG loss. In contrast, with the stronger discriminator supervision provided by the semantic label maps and the multi-scale U-Net architecture, OASIS achieves a better adversarial balance. Hence, the generator is forced to learn semantically meaningful features that the segmentation-based discriminator judges as real, and the generator loss does not stay constant (see Fig. 5).

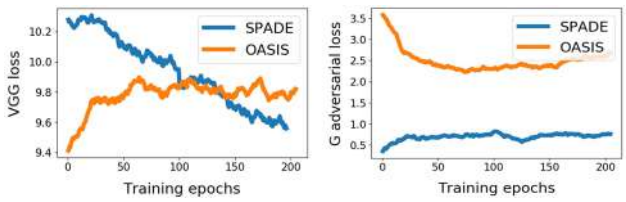

Fig. 5 VGG and adversarial generator loss functions for SPADE and OASIS trained with VGG loss on ADE20k dataset. The adversarial loss scales are different due to different objectives (binary or (N+1)-class cross entropy loss)

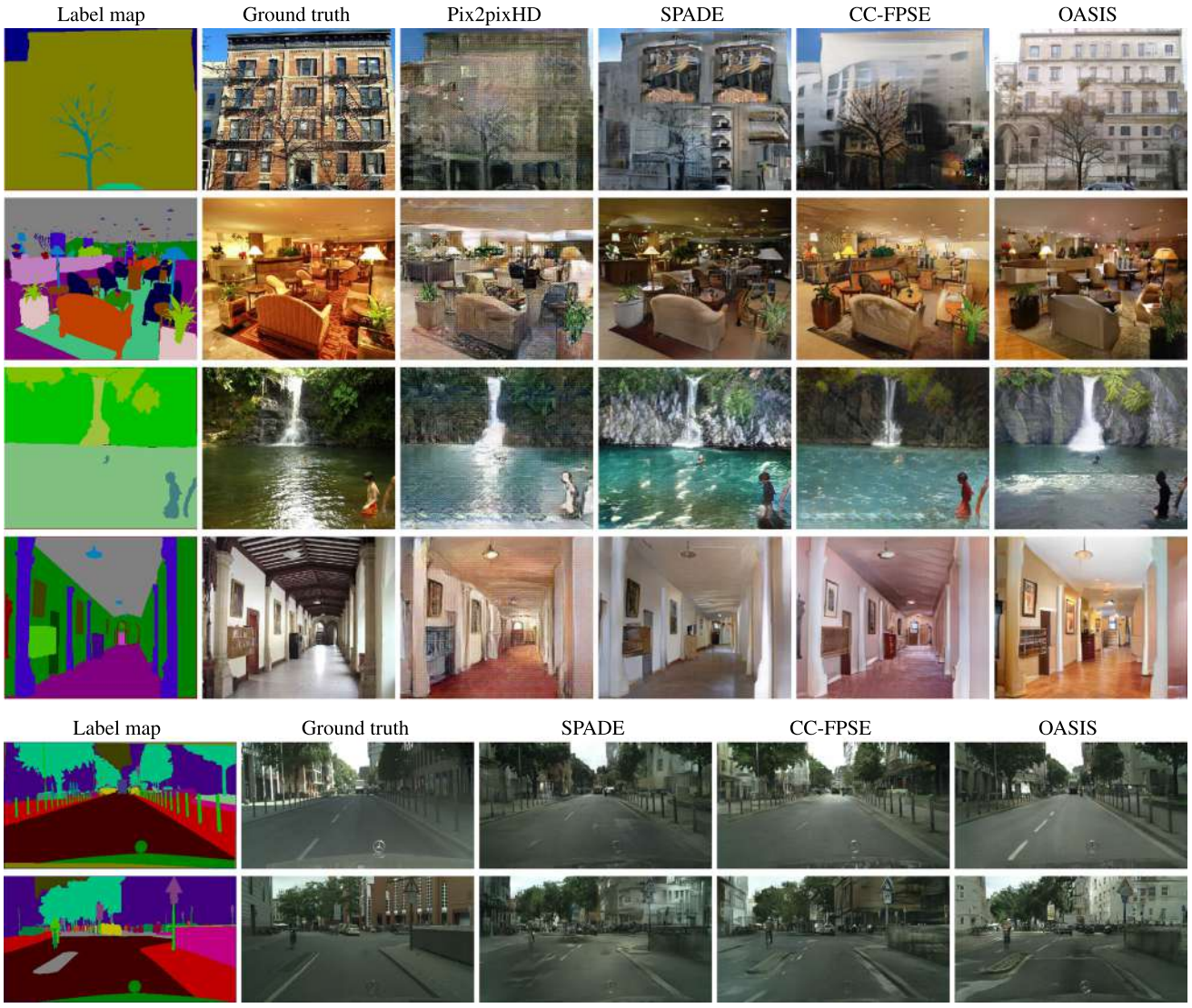

Fig. 6 Qualitative comparison of OASIS with other methods on ADE20K and Cityscapes. Trained with only adversarial supervision, our model generates images with better perceptual quality and structure

The advantage of training the generator only with the adversarial loss is three-fold. Firstly, the perceptual loss can bias the training signal with the color and texture statistics encoded in the VGG features extracted from ImageNet. As shown in Sect. 4.2, the strong adversarial supervision from the OASIS discriminator, without the VGG loss, allows to generate images with color and texture distributions closer to the provided real data. Secondly, the perceptual loss can induce unnecessary constraints on the generator and thus significantly limit the diversity of multi-modal image syntesis. This effect is further demonstrated in Table 2. Lastly, removing the perceptual loss eliminates the computational overhead which was introduced by an additional VGG network during training.

4 Experiments

We provide an extensive experimental evaluation of our contributions, using the official implementation of SPADE[1] as our baseline. The setup of our experiments is described in detail in Sect. 4.1. Firstly, we compare OASIS with prior methods on common semantic image synthesis benchmark datasets, comparing their performance in terms of both image quality and diversity (Sect. 4.2). To further highlight the advantages of OASIS over the SPADE baseline, we provide additional discussions on different aspects of the semantic image synthesis. In particular, Sect. 4.3 is devoted to the performance analysis on the underrepresented classes, extending the comparison of the models to the LVIS

[1] https://github.com/NVlabs/SPADE.

dataset (Gupta et al., 2019). Section 4.4 demonstrates new semantic image editing techniques enabled by OASIS. Section 4.5 explores the application of generated images as synthetic data augmentation for the training of semantic segmentation networks. Lastly, we provide an extensive ablation study to verify the effectiveness of the proposed contributions (Sect. 4.6).

4.1 Experimental Setup

Datasets. We conduct experiments on several challenging datasets. Firstly, to compare OASIS with prior models, we use the ADE20K (Zhou et al., 2017), COCO-Stuff (Caesar et al., 2018) and Cityscapes (Cordts et al., 2016), which are the three benchmark datasets commonly used in the semantic image synthesis literature (see Sect. 4.2). The image resolution is set to 256x256 for ADE20K and COCO-Stuff, and 256x512 for experiments on Cityscapes. Following Qi et al. (2018), we also evaluate OASIS on ADE20K-outdoors, the subset of ADE20K containing only outdoor scenes.

Secondly, to test the capability of models to learn underrepresented classes, we conduct additional evaluations on the ADE20K and LVIS dataset (Gupta et al., 2019) (see Sect. 4.3). We select ADE20K among conventional datasets for its notable class imbalance, as among its 150 classes, more than 86% of the image pixels belong only to the 30 best represented ones (see Table 3). In addition, to test the networks under more extreme class imbalance, we propose to use LVIS, the dataset that has been originally introduced for the task of long-tailed instance segmentation. LVIS employs the same set of training images as COCO-Stuff, but its annotations are different in two important ways. Firstly, LVIS provides a significantly larger set of 1203 annotated classes, following a long-tailed distribution in which some classes are present only in one or a few training samples (see Fig. 7). Secondly, due to a fixed labelling budget, different background types were not considered for annotation in LVIS. Consequently, the images in LVIS dataset contain large areas belonging to the background class, which sometimes covers more than 90% of the pixels in an image (see grey areas in Fig. 10). For the above two reasons, the structure of LVIS poses a new challenge for semantic image synthesis, as models need to account for a much more extreme class imbalance. We conduct experiments on LVIS at the image resolution of 128x128.

Training. We follow the experimental setting of Park et al. (2019b). The Adam (Kingma & Ba, 2015) optimizer was used with momenta $\beta = (0, 0.999)$ and constant learning rates $(0.0001, 0.0004)$ for G and D. We did not use the GAN feature matching loss for OASIS, as we did not observe any improvement with it, and used the VGG loss only for ablations with $\lambda_{\text{VGG}} = 10$. The parameter for LabelMix λ_{LM} was set to 5 for ADE20k and Cityscapes, and to 10 for COCO-Stuff and LVIS. The latent dimension M was set to 64. We did not experience any training instabilities and, thus, did not employ any extra stabilization techniques. All our models use an exponential moving average (EMA) of the generator weights with 0.9999 decay. All the experiments were run on 4 Tesla V100 GPUs, with a batch size of 20 for Cityscapes and 32 for the other datasets. The training epochs are 200 on ADE20K and Cityscapes, and 100 for the larger COCO-Stuff and LVIS datasets. On average, a complete forward-backward pass with batch size 32 on ADE20k takes around 0.95ms per training image.

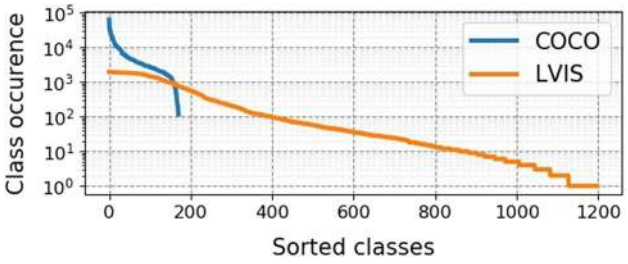

Fig. 7 Comparison of class distributions of the COCO and LVIS datasets. LVIS has a much larger vocabulary of 1203 classes with a long tail of underrepresented classes

Evaluation metrics. Following prior work (Park et al., 2019b; Liu et al., 2019), we evaluate the *quality* of semantic image synthesis by computing the FID (Heusel et al., 2017) and evaluate the *alignment* of the generated images with their semantic label maps via mIoU (mean intersection-over-union) or mAP (mean average precision) on the test set (see Sect. 4.2). mIoU evaluates the alignment of generated images with their ground truth label maps, as measured by an external pre-trained semantic segmentation network. We use UperNet101 (Xiao et al., 2018) for ADE20K, multi-scale DRN-D-105 (Yu et al., 2017) for Cityscapes, and DeepLabV2 (Chen et al., 2015) for COCO-Stuff. Differently, for the LVIS dataset, the alignment of generated images to ground truth label maps is measured using mAP instead of mIoU, following the official guidelines for evaluating instance segmentation models on this dataset (see Sect. 4.3). We compute mAP using a state-of-the-art instance segmentation model from Wang et al. (2021a), pre-trained on LVIS.

In addition, to better understand how the perceptual loss influences the synthesis performance, we propose to compare the *color and texture statistics* of generated and real images. For this, we compute color histograms in the LAB space and measure the earth mover's distance between the real and generated image sets (Rubner et al., 2000). We also measure the texture similarity to the real data as the χ^2-distance between Local Binary Patterns histograms (Ojala et al., 1996). As different semantic classes have different color and texture distributions, we aggregate the histogram distances separately per class and compute their average.

Table 1 Comparison with other methods across datasets

Method	# param	VGG	ADE20K		ADE-outd.		Cityscapes		COCO-stuff	
			FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
CRN	84M	✓	73.3	22.4	99.0	16.5	104.7	52.4	70.4	23.7
SIMS	56M	✓	n/a	n/a	67.7	13.1	49.7	47.2	n/a	n/a
Pix2pixHD	183M	✓	81.8	20.3	97.8	17.4	95.0	58.3	111.5	14.6
LGGAN	n/a	✓	31.6	41.6	n/a	n/a	57.7	68.4	n/a	n/a
CC-FPSE	131M	✓	31.7	43.7	n/a	n/a	54.3	65.5	19.2	41.6
SC-GAN	66M	✓	29.3	45.2	n/a	n/a	49.5	66.9	18.1	42.0
SESAME	104M	✓	31.9	**49.0**	n/a	n/a	54.2	66.0	n/a	n/a
SPADE	102M	✓	33.9	38.5	63.3	30.8	71.8	62.3	22.6	37.4
SPADE+	102M	✓	32.9	42.5	51.1	32.1	47.8	64.0	21.7	38.8
		✗	60.7	21.0	65.4	22.7	61.4	47.6	99.1	16.1
OASIS	94M	✗	**28.3**	48.8	**48.6**	**40.4**	**47.7**	**69.3**	**17.0**	**44.1**

Bold denotes the best performance

To measure the *diversity* among synthesized samples in the multi-modal image generation regime, we evaluate MS-SSIM (Wang et al., 2003) and LPIPS (Zhang et al., 2018b) between the images generated from the same label map. For each label map in the test set, we generate 20 images and compute the mean pairwise scores. For the final numbers, the scores are averaged over all label maps.

Lastly, we propose to test the efficacy of generated images when applied as *synthetic data augmentation* for the task of semantic segmentation (see Sect. 4.5). For this, we take a DeepLab-V3 segmentation network with a ResNeSt-50 backbone (Zhang et al., 2020) and train it on ADE20K and Cityscapes. At each training step of DeepLab-V3, we add for each training image its synthetic counterpart to the batch, generated from the same label map. The efficacy of synthetic images is therefore measured by its effect on the downstream mIoU performance of DeepLab-V3.

4.2 Evaluation of the Synthesis Quality and Diversity

In this section, we compare OASIS to previous state-of-the-art methods. For a fair comparison to the baseline SPADE, we additionally train this model without the feature matching loss and using EMA (Yaz et al., 2018) at the test phase. We refer to this improved baseline as SPADE+.

Synthesis quality. Table 1 compares the image synthesis quality achieved by OASIS and previous methods. In this table, we report the results of our evaluation for OASIS and SPADE+, and the officially reported numbers for all the other models. As seen from Table 1, OASIS outperforms prior state-of-the-art models in FID on all benchmark datasets. Our model also achieves the highest mIoU scores on three out of four datasets, being almost on par with the highest score on ADE20K achieved by SESAME (Ntavelis et al., 2020) Importantly, OASIS achieves the improvement using only adversarial supervision from its segmentation-based discriminator. On the contrary, in the absence of the VGG loss, the baseline SPADE+ does not produce images of high visual quality (see Fig. 1), with two-digit drops in FID scores observed for all the datasets in Table 1. The strong adversarial supervision also allows OASIS to produce images with color and texture distributions closer to the real data. Such improvement over SPADE+ on the ADE20K dataset is shown in Fig. 8, where OASIS achieves the lowest color and texture distances to the target distribution. In contrast, SPADE+ needs to compensate a weaker discriminator signal with the VGG loss, struggling to learn the color and texture distribution of real images without it (see Fig. 8).

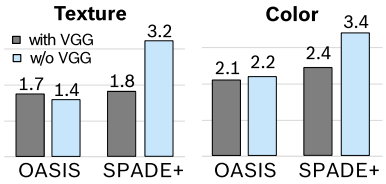

Fig. 8 Histogram distances to real data on the ADE20K validation set. While SPADE+ relies on the VGG loss to learn colors and textures, OASIS achieves low scores without it

Figure 6 shows a qualitative comparison of our results to previous models. Our approach noticeably improves image quality, synthesizing finer textures and more natural colors. While the previous methods occasionally produce areas with unnatural checkerboard artifacts, OASIS generates large objects and surfaces with higher photorealism. Notably, the improvement over previous models is especially remarkable for the semantic classes that occupy large areas, e.g, wall (rows 1,4 in Fig. 6), road (rows 5,6) or water (row 3).

Table 2 Multi-modal synthesis evaluation on ADE20K

Method	Multi-mod.	VGG	MS-SSIM↓	LPIPS↑	FID↓	mIoU↑
SPADE+	Encoder	✓	0.85	0.16	33.4	40.2
SPADE+	3D noise	✗	**0.35**	**0.50**	*58.4*	*18.7*
		✓	0.53	0.36	34.4	36.2
OASIS	3D noise	✗	0.65	0.35	**28.3**	48.8
		✓	*0.88*	*0.15*	31.6	**50.8**

Bold and italic denote the best and the worst performance

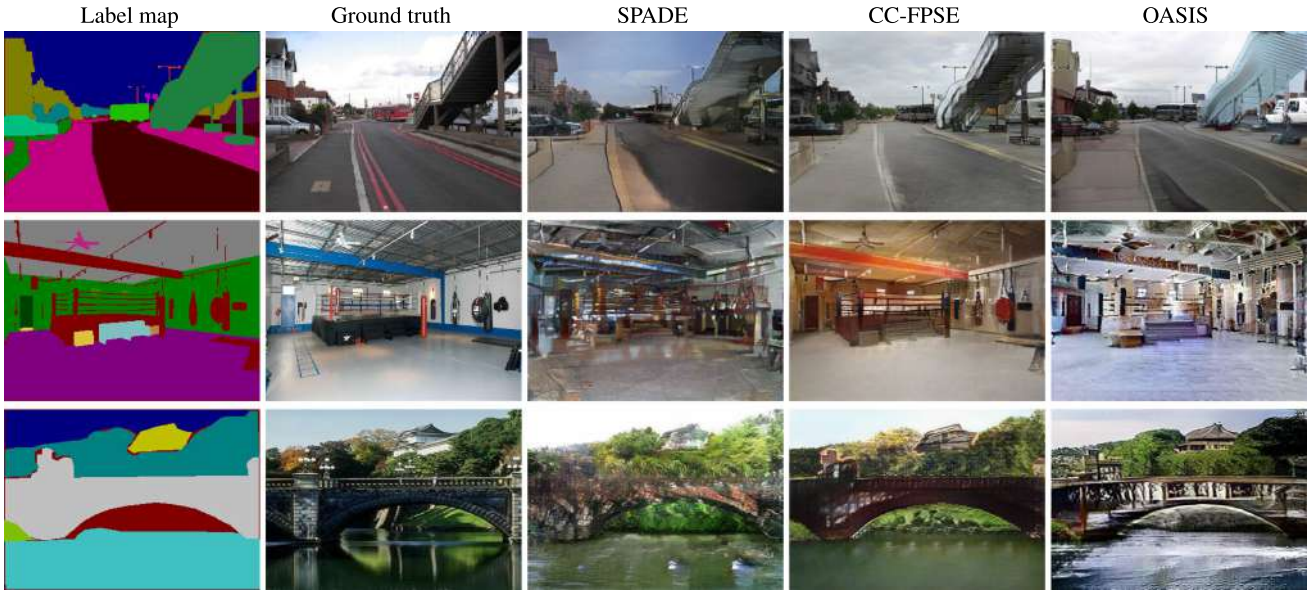

Fig. 9 Failure mode of OASIS. Without the VGG loss, OASIS has less constraints on the diversity in colors and textures. This helps to achieve higher diversity among the generated samples, but sometimes leads to synthesis of objects with outlier colors and textures which may look less realistic compared to Park et al. (2019b) and Liu et al. (2019)

Synthesis diversity. By resampling the input 3D noise, OASIS can produce diverse images given the same label map (see Fig. 2). To measure the diversity of such multi-modal synthesis, we evaluate MS-SSIM (Wang et al., 2003) and LPIPS (Zhang et al., 2018b). The lower the MS-SSIM and the higher the LPIPS scores, the more diverse the generated images are. As seen from Table 2, OASIS outperforms SPADE+ in both diversity metrics, improving the MS-SSIM scores from 0.85 to 0.65 and LPIPS from 0.16 to 0.35. To assess the effect of the perceptual loss and the noise sampling on diversity, we train SPADE+ with 3D noise or the image encoder, and with or without the perceptual loss. Table 2 shows that OASIS, without the perceptual VGG loss, improves over SPADE+ with the image encoder, both in terms of image diversity (MS-SSIM, LPIPS) and quality (mean FID, mIoU across 20 realizations). Using 3D noise further increases diversity for SPADE+. However, a strong quality-diversity trade-off exists for SPADE+: 3D noise improves diversity at the cost of quality, and the perceptual loss improves quality at the cost of diversity. We conclude that our 3D noise injection strongly improves the synthesis diversity, while the VGG loss decreases it.

While the increased diversity is a big advantage, it can also lead to failures in rare cases: for some samples the colors and textures of objects may lie further from the real distribution and seem unnatural to the human eye (see Fig. 9).

4.3 Synthesis Performance on Underrepresented Classes

Class imbalance is a well-known challenge in semantic segmentation applications (Sudre et al., 2017). Similarly to semantic segmentation, to ensure good performance in real-life test scenarios, semantic image synthesis models should account for a possible dataset class imbalance, especially considering that GANs are notorious for dropping modes of training data (Arjovsky & Bottou, 2017). However, to the best of our knowledge, this issue was not addressed in prior works. Thus, in what follows, we evaluate the performance of OASIS and SPADE+ on the ADE20K and LVIS datasets,

Table 3 Per-class IoU scores on ADE20k, grouped by pixel-wise frequency (the fraction of all pixels in the datasets belonging to one class)

Classes IDs	Pixel-wise frequency (%)	mIoU		
		SPADE+	OASIS (w/o α_c)	OASIS (w. α_c)
0–29	86.4	63.7	**69.1**	68.8
30–59	7.2	47.4	52.4	**56.6**
60–89	3.5	45.3	47.0	**51.5**
90–119	1.8	29.3	36.2	**41.5**
120–149	1.0	26.2	31.2	**39.7**
0–149 (all classes)	100	42.4	47.2	**51.6**

Bold denotes the best performance. Training with per-class loss balancing is denoted by α_c

Table 4 Comparison of SPADE+ and OASIS on the LVIS dataset with 1203 classes and a long tail of underrepresented classes

Method	FID ↓	mAP, % ↑	Classes with AP > 0 ↑
SPADE+	26.8	4.56	439
OASIS	**15.3**	**5.38**	**510**
real data	0	6.70	624

Bold denotes the best performance. Last row shows the scores for the LVIS validation set

considering their class imbalances. While the class imbalance in ADE20K is notable (e.g., 86.4% of all image pixels belongs to the 30 best represented classes), this issue is much more amplified in LVIS, which has a long tail of underrepresented classes (see Fig. 7).

Evaluation on ADE20K. OASIS significantly outperforms the SPADE+ baseline in the alignment between generated images and label maps, as measured by mIoU (see Table 1). As shown in Table 3, the improvement in mIoU on ADE20K comes mainly from the better IoU scores achieved for underrepresented semantic classes.

To illustrate this, the semantic classes are sorted by their pixel-wise frequency in the training images, obtained by dividing the number of pixels a class occupies in the dataset by the total number of pixels of all images (2nd column in Table 3). Table 3 highlights that the relative gain in mIoU is especially high for the groups of underrepresented semantic classes, that cover less than 3% of all pixels in the dataset. For these classes, the relative gain over the SPADE+ baseline exceeds 40%. Remarkably, the gain for this group mainly comes from the per-class balancing applied in the OASIS loss function (columns "w/o α_c" and "w. α_c"), which draws the attention of the discriminator to underrepresented semantic classes, thus allowing a higher quality of their generation. This class balancing computes a weight α_c for the losses of each class c on a per-batch basis, for which the total number of pixels in a given batch is divided by the number of pixels belonging to the class (see Eqs. 2 and 3). We note that the possibility to introduce the pixel-wise frequency based balancing requires the loss to be computed separately for each image pixel. This is a unique property of the OASIS discriminator, in contrast to conventional classification-based discriminators, which have to evaluate realism with a single score for images containing both well- and underrepresented classes together.

Evaluation on LVIS. A quantitative comparison between the models on the LVIS dataset is shown in Table 4. In this more extremely imbalanced data regime, the gain of our model is pronounced: OASIS outperforms SPADE+ by a large margin, lowering the FID by 43% (from 26.8 to 15.3). Figure 10 shows a qualitative comparison between the models. OASIS produces images of higher visual quality with more natural colors and textures. In Table 4 we report the mean Average Precision (mAP) of the instance segmentation network evaluated on the set of generated images. OASIS outperforms SPADE+ in mAP by a notable margin (5.38 vs 4.56), thus producing objects with a more realistic appearance and largely reducing the gap to real data (mAP of 6.70). To evaluate the ability of the models to generate underrepresented classes at the tail of the LVIS data distribution, we count the number of classes for which a non-zero AP score is achieved. Table 4 shows that OASIS can model more semantic classes: OASIS achieves a positive AP for 510 semantic classes compared to 439 for SPADE+, thus exhibiting a better capability to synthesize underrepresented classes.

In addition to better handling the class imbalance, OASIS also visually outperforms SPADE+ on the LVIS label maps with a very large proportion of the background class. As seen in Fig. 10 (four rightmost columns), from such label maps, SPADE+ fails to produce plausible images and suffers from mode collapse. In contrast, OASIS successfully deals with such kinds of inputs, producing diverse and visually plausible images even for the least annotated label maps, with the highest proportion of the background class.

In conclusion, we consider long-tailed datasets, such as LVIS, an interesting direction for future work, as the improved synthesis of multiple tail classes under severe imbalance can significantly boost the applicability of semantic image synthesis to real-world applications.

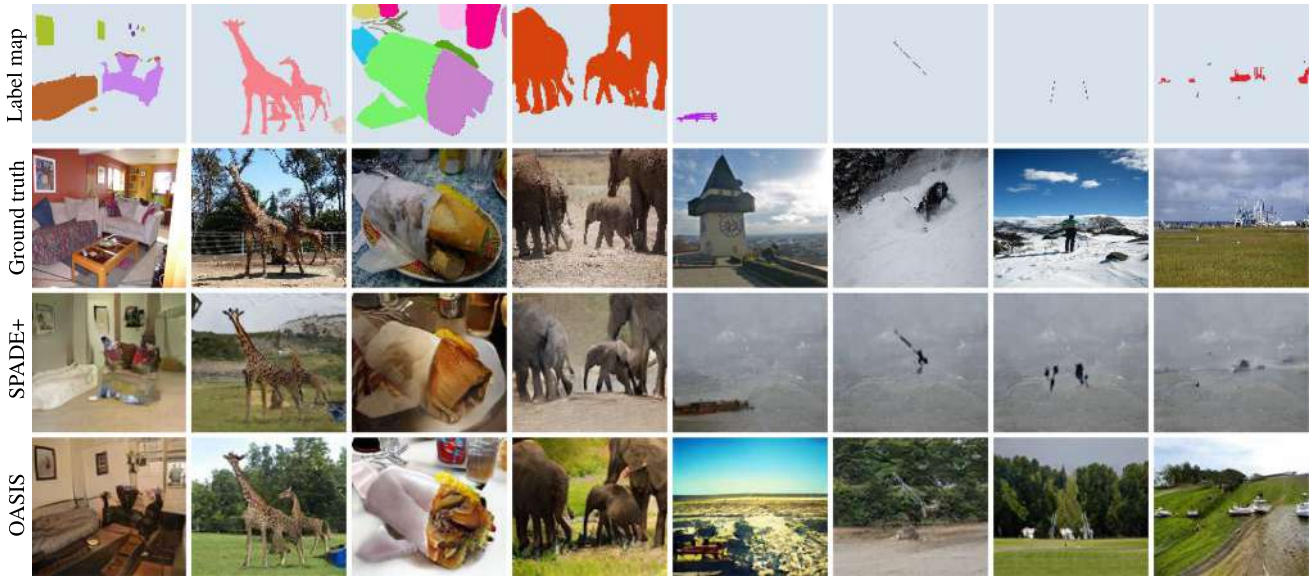

Fig. 10 Qualitative comparison between OASIS and SPADE+ on the long-tailed LVIS dataset with 1203 classes. OASIS generates higher-quality images with more natural colors and textures. For label maps covered mostly by the background class (four right columns), OASIS hallucinates plausible and diverse images, while SPADE+ suffers from mode collapse

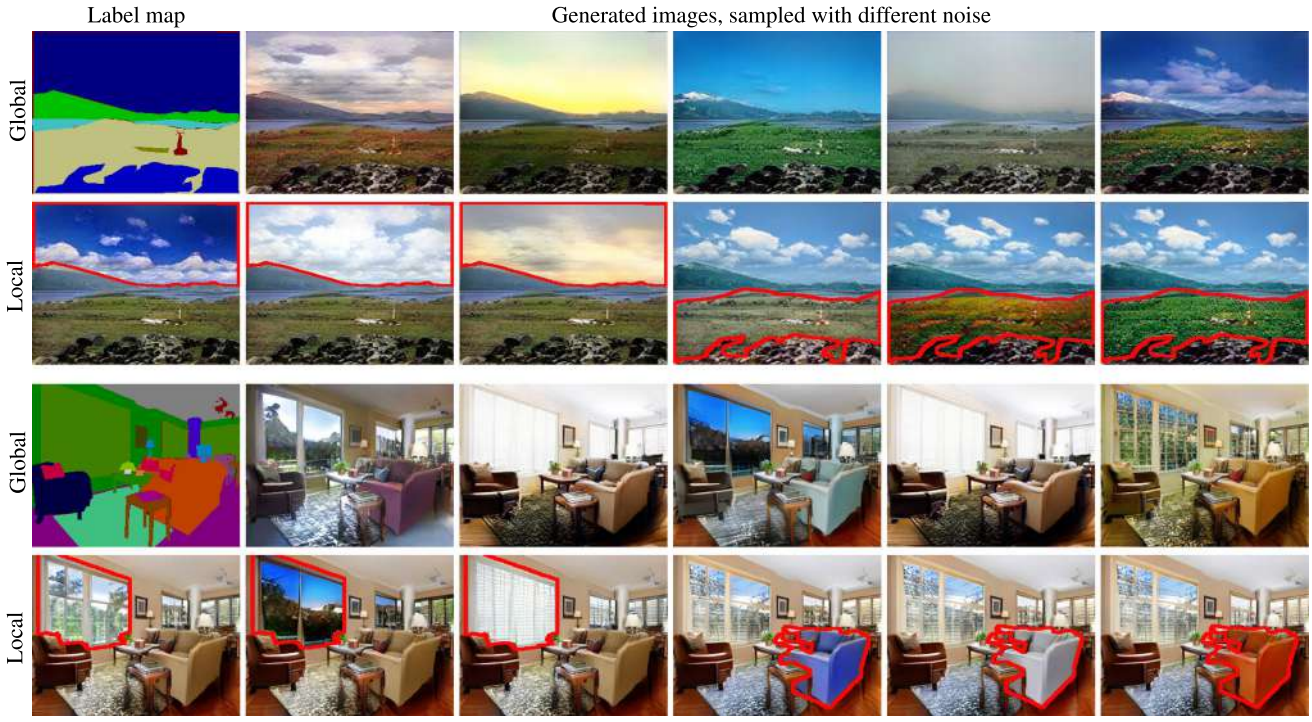

Fig. 11 Images generated by OASIS on ADE20K with 256 × 256 resolution using different 3D noise inputs. For both input label maps, the noise is re-sampled globally (first row) or locally in the areas marked in red (second row)

4.4 Image Editing with OASIS

OASIS can generate many different-looking images for a single label map by directly resampling input 3D noise. In the following, we present qualitative multi-modal results and discuss two unique semantic image editing techniques enabled by our model: local resampling of selected semantic classes and diverse resampling of unlabelled images.

Global and local resampling of the 3D noise. The 3D noise of OASIS modulates the activations directly at every

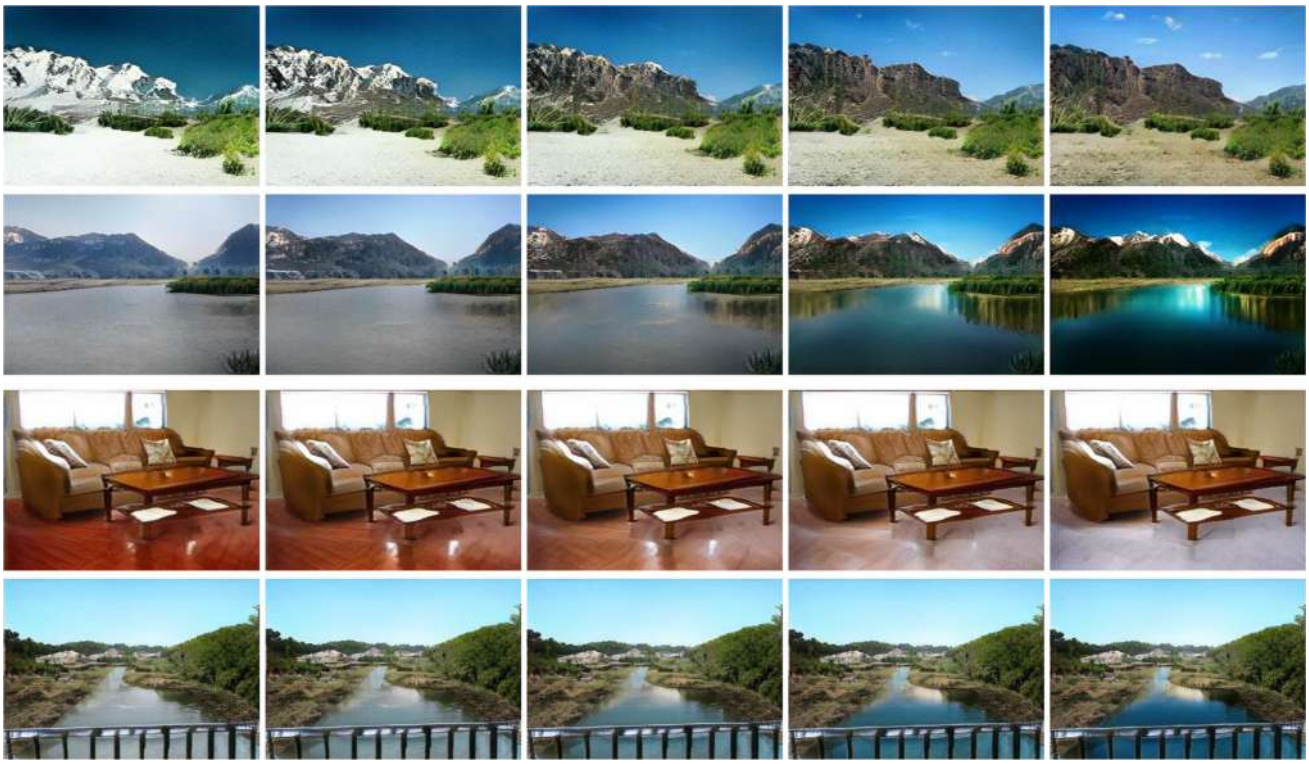

Fig. 12 Latent space interpolations between images generated by OASIS for the ADE20K dataset at resolution 256 × 256. The first two rows display *global* interpolations. The second two rows show *local* interpolations of the floor or water only

generator layer, matching the spatial resolution of features at different generation scales. Therefore, such modulation affects both global and local characteristics of a generated image. At test time, this allows different strategies for noise sampling. For example, the noise can be sampled globally for all pixels, varying the whole image (see Fig. 11, first and third rows). Alternatively, a noise vector can be re-sampled only for specified image regions, resulting in local image editing while preserving the rest of the scene. For example, the local strategy allows to re-sample only the sky area in a landscape scenery, or only the window in a scene of a bedroom (see Fig. 11, second and fourth rows). Spatial sensitivity of OASIS to 3D noise is further demonstrated in Fig. 12, showing interpolations in the latent space. The learned latent space captures well the semantic meaning of objects and allows smooth interpolations not only globally, but also locally for selected objects (see Fig. 12, two last rows).

Creating diverse images from unlabelled data. In contrast to previous semantic image synthesis methods, the OASIS discriminator can be reused as a stand-alone image segmenter. To obtain a segmentation prediction for a given image, a user just needs to feed it to our pre-trained discriminator and select the highest activation among real classes in its $(N + 1)$-channel output for each pixel. When tested as an image segmenter on the validation set of ADE20K, the OASIS discriminator reaches a mIoU of 40.0. For comparison, the state-of-the-art model DeepLab-V3 with a ResNeST backbone (Zhang et al., 2020) achieves an mIoU of 46.91. The good segmentation performance allows OASIS to be applied to unlabelled images: given an unseen image without the ground truth annotation, OASIS can predict a label map via the discriminator. Subsequently feeding this prediction to the generator allows to synthesize a scene with the same layout but different style (see Fig. 13). The recreated scenes closely follow the ground truth label map of the original image and vary considerably, due to the high sensitivity of OASIS to the 3D noise. We note that OASIS uniquely reaches this ability using only adversarial training, without the need for an external segmentation network or additional loss functions. We believe that the ability to create multiple versions of one image while retaining the layout, but not requiring the ground truth label map, may provide useful data augmentation for various applications in future research.

4.5 Synthetic Data Augmentation

As an additional evaluation method, we test the efficacy of generated images when applied as synthetic data augmentation for the task of semantic segmentation. Synthetic data augmentation is a task that benefits from both image quality and diversity, as well as the ability to generate semantic classes that are underrepresented in the original data (see

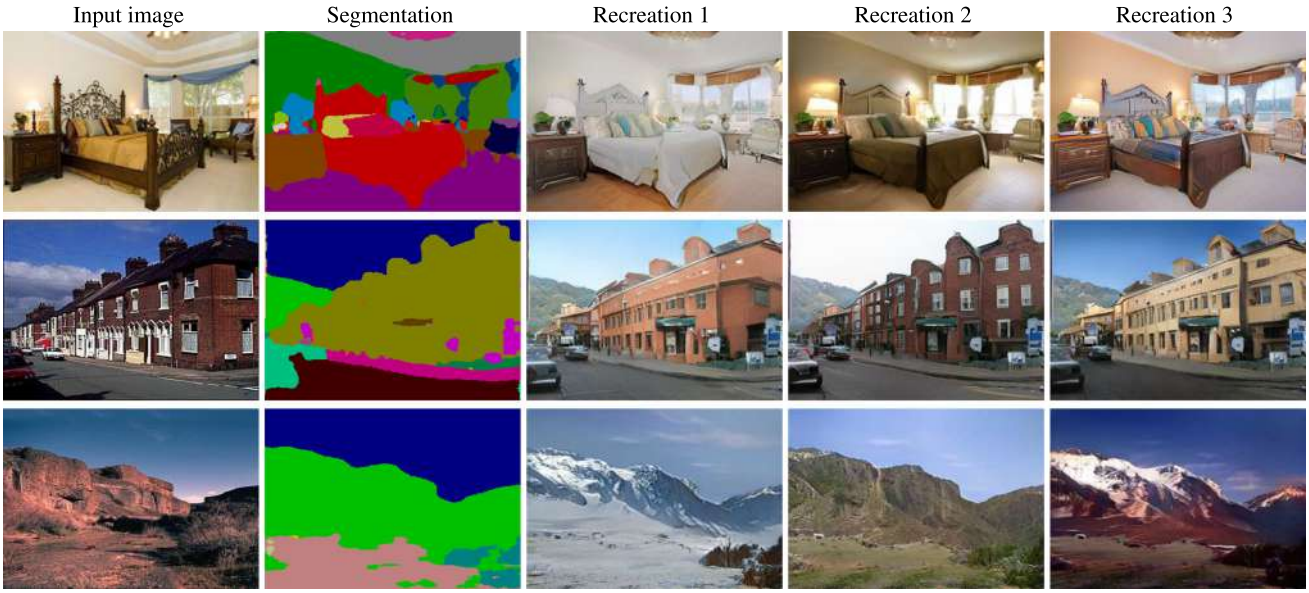

Fig. 13 After training, the OASIS discriminator can be used to segment images. The first two columns show the real image and the segmentation of the discriminator. Using the predicted label map, the generator can produce multiple versions of the original image by resampling noise (Recreations 1–3). Note that no ground truth maps are required

Table 5 Semantic segmentation performance of ResNeSt-50 with and without synthetic data augmentation (DA)

Data augmentation	Cityscapes mIoU↑	ADE20K mIoU↑
No synthetic DA	62.7	41.0
With SPADE	62.6	41.6
With OASIS	**64.7**	**41.8**

Bold denotes the best performance

Table 3). Therefore, the effect of synthetic data augmentation on downstream performance can constitute a more holistic evaluation of semantic image synthesis models. To test the efficiency of OASIS, we train a DeepLab-V3 segmentation network on ADE20K and Cityscapes, at each step augmenting each training image with its synthetic augmentation, produced by OASIS from the same label map.

We compare OASIS against the strong baseline SPADE in Table 5. Between the two methods, OASIS elicits a stronger increase in segmentation performance with an improvement of 2.0 mIoU on Cityscapes and 0.8 mIoU on ADE20K, compared to DeepLab-V3 trained without synthetic augmentation. The higher performance improvement of OASIS compared to SPADE is explained by all the previously observed gains in image quality, diversity, and the alignment to input label maps (see Fig. 8, Tables 1 and 2). In addition to that, the segmentation performance is also improved due to the fact that OASIS tends to synthesize underrepresented classes better than SPADE, which is evident from Table 6. This table compares the IoU performance of DeepLab-V3 on the well represented and underrepresented classes of Cityscapes, as measured by the pixel-wise frequency of the semantic class in the dataset. Examples of well represented classes are road and building (see the 1st row of Table 6), while classes like bicycle or traffic light are the least represented in the dataset (see 4th row in Table 6). Note that the IoU comparison in Table 6 is different from Table 3, where the IoU was measured directly on synthetic data using a pretrained segmenter. It can be seen that the improvement in IoU through OASIS can be mostly attributed to better performance on underrepresented classes, as the gap in performance between OASIS and SPADE becomes larger for the classes which are less represented. Lastly, since the OASIS generator was trained to fool an image segmenter (the OASIS discriminator), it may synthesize harder examples for semantic segmentation than SPADE, thus having higher potential to improve the generalization of segmentation networks to challenging corner cases. We find the above results promising for future utilization of OASIS in various downstream applications. Moreover, for future research, we find it interesting to explore synthetic data augmentation in combination with other data augmentation techniques, e.g., RandAugment (Cubuk et al., 2020), which has the potential to provide further performance gains for downstream applications.

Table 6 Per-class IoU scores on Cityscapes, obtained without (None) and with synthetic data augmentation using SPADE or OASIS

Sorted classes	Pixel-wise frequency (%)	None	SPADE		OASIS	
			abs	rel	abs	rel
0–4	82.7	90.6	90.6	+0.0	**90.9**	**+ 0.3**
5–8	12.5	66.2	66.2	+0.0	**67.4**	**+ 1.2**
9–12	3.3	50.2	49.1	− 1.1	**52.2**	**+ 2.0**
13–18	1.6	51.9	52.3	+0.4	**55.4**	**+3.5**
All classes	100	62.7	62.6	− 0.1	**64.7**	**+ 2.0**

The classes are sorted and grouped by class pixel-wise frequency, as measured by the total fraction of pixels in the dataset belonging to one class. Bold denotes the best performance. The absolute (abs) and relative (rel) mIoU gain via data augmentation is shown

Table 7 Main ablation on ADE20K. The OASIS generator is a lighter version of the SPADE+ generator (72M vs 96M parameters)

G	*D*	VGG	LabelMix	FID↓	mIoU↑
SPADE+	SPADE+	✗	✗	60.7	21.0
SPADE+	OASIS	✗	✗	29.0	**52.1**
OASIS	OASIS	✗	✗	29.3	51.6
		✗	✓	28.4	50.6
OASIS +3D noise	OASIS	✗	✓	**28.3**	48.8
		✓	✓	31.6	50.8

Bold denotes the best performance

Table 8 Ablation on the *D* architecture

D architecture	w/o VGG		with VGG	
	FID↓	mIoU↑	FID↓	mIoU↑
MS-PatchGAN (2x)	60.7	21.0	32.9	42.5
PatchGAN	*197*	*0.62*	34.2	42.2
ResNet-PatchGAN	*147*	*0.42*	32.4	45.1
OASIS	**29.3**	**51.6**	**29.2**	**51.1**

Bold denotes the best performance, italics shows collapsed runs

Table 9 Ablation on the label map encoding runs

Label encoding	w/o VGG		with VGG	
	FID↓	mIoU↑	FID↓	mIoU↑
Input concatenation	*280*	*0.02*	30.0	43.9
Projection	32.4	44.9	**28.0**	46.9
N+1 loss	**28.3**	47.2	28.6	49.8
Balanced N+1 loss	29.3	**51.6**	29.2	**51.1**

Bold denotes the best performance, italics shows collapsed runs

4.6 Ablations

We conduct all our ablations on the ADE20K dataset. We choose this dataset as it more challenging (with 150 classes) than Cityscapes (35 classes) and ADE20K-Outdoors (110 classes), and has more reasonable training time (5 days) compared to COCO-Stuff and LVIS (4 weeks). Our main ablation shows the impact of the main technical components of OASIS, including the new discriminator, lighter generator, LabelMix and the 3D noise. Further ablations are concerned with the architecture changes in the discriminator, the label map encoding in the discriminator, different noise sampling strategies, LabelMix and the GAN feature matching loss.

Main ablation. Table 7 shows that SPADE+ achieves low performance on the image quality metrics without the perceptual loss. Replacing the SPADE+ discriminator with the OASIS discriminator, while keeping the generator fixed, improves FID and mIoU by more than 30 points. Changing the SPADE+ generator to the lighter OASIS generator leads to a negligible degradation of 0.3 in FID and 0.5 in mIoU, but reduces the number of parameters from 96M to 72M. With LabelMix FID improves further by about 1 point. Adding 3D noise improves FID but degrades mIoU, as diversity complicates the task of the pre-trained semantic segmentation network used to compute the mIoU score. For OASIS the perceptual loss deteriorates FID by more than 2 points, but improves mIoU. Overall, without the VGG loss the new discriminator is the key to the performance boost over SPADE+.

Ablation on the discriminator architecture. We train the OASIS generator with three alternative discriminators: the original multi-scale PatchGAN consisting of two networks, a single-scale PatchGAN, and a ResNet-based discriminator, corresponding to the encoder of the U-Net shaped OASIS discriminator. Table 8 shows that the alternative discriminators only perform well with perceptual supervision, while the OASIS discriminator achieves superior performance independent of it. The single-scale discriminators even collapse without the perceptual loss (italic in Table 8).

Table 10 Different 3D noise sampling strategies during training. Bold denotes the best performance

Sampling	Cityscapes			ADE20K		
	FID↓	mIoU↑	MS-SSIM↓	FID↓	mIoU↑	MS-SSIM↓
Image-level	47.7	69.3	0.64	**28.3**	**48.8**	0.65
Region-level	48.1	69.7	**0.62**	28.8	48.1	**0.58**
Pixel-level	50.9	65.5	0.84	28.6	34.0	0.68
Mix	**46.4**	**70.9**	0.68	28.5	47.6	0.66

Ablation on the discriminator label map encoding. We study four different ways to use label maps in the discriminator: the first encoding is input concatenation, as in SPADE. The second option is a pixel-wise projection-based GAN loss (Miyato & Koyama, 2018). Unlike Miyato and Koyama (2018), we condition the GAN loss on the label map instead of a single label. The third and fourth option is to employ the label maps as ground truth for the $N + 1$ segmentation loss, or for the class-balanced $N + 1$ loss (see Sect. 3.2). For a fair comparison we use neither 3D noise nor LabelMix. As shown in Table 9, input concatenation is not sufficient without additional perceptual loss supervision, leading to training collapse. Without the perceptual loss, the $N + 1$ loss outperforms the input concatenation and the projection in both the FID and mIoU metrics. Finally, the class balancing enables enhanced supervision for underrepresented semantic classes, which noticeably improves mIoU scores. On the other hand, we observed that the FID metric is more sensitive to the synthesis of well represented classes and not underrepresented classes, which explains the negative effect of the class balancing on FID.

Ablation on noise sampling strategies for training. Our 3D noise can contain the same sampled vector for each pixel, or different vectors for different regions. This allows for different sampling strategies during training. Table 10 shows the effect of using different methods of sampling 3D noise for different locations during training: *Image-level* sampling creates one global 1D noise vector and replicates it along the height and width of the label map to create a 3D noise tensor. *Region-level* sampling relies on generating one 1D noise vector per semantic class, and stacking them in 3D to match the height and width of the semantic label map. *Pixel-level* sampling creates different noise for every spatial position, with no replication taking place. *Mix* switches between image-level and region-level sampling via a coin flip decision at every training step. With no obvious winner in performance, we choose the simplest scheme (image-level) for our experiments. We find a further investigation with more advanced strategies an interesting direction for future work.

Ablation on LabelMix. Consistency regularization for the segmentation output of the discriminator requires a method of generating binary masks. Therefore, we compare the effectiveness of CutMix (Yun et al., 2019) and our proposed LabelMix. Both methods produce binary masks, but only LabelMix respects the boundaries between semantic classes in the label map. Table 11 compares the FID and mIoU scores of OASIS trained with both methods on the Cityscapes dataset. As seen from the table, LabelMix improves both FID (51.5 vs. 47.7) and mIoU (66.3 vs. 69.3), in comparison to OASIS without consistency regularization. CutMix-based consistency regularization only improves the mIoU (66.3 vs. 67.4), but not as much as LabelMix (69.3). We suspect that since the images are already partitioned through the label map, an additional partition through CutMix results in a dense patchwork of areas that differ by semantic class and real/fake class identity. This may introduce additional label noise during training for the discriminator. To avoid such inconsistency between semantic classes and real/fake identity, the mask of LabelMix is generated according to the label map, providing natural borders between semantic regions, so that the real and fake objects are placed side-by-side without interfering with each other. Under LabelMix regularization, the generator is encouraged to respect the natural semantic class boundaries, improving pixel-level realism while also considering the class segment shapes.

Table 11 Ablation study on the impact of LabelMix and CutMix for consistency regularization (CR) in OASIS on Cityscapes

Transformation	FID↓	mIoU ↑
No CR	51.5	66.3
CutMix	52.1	67.4
LabelMix	**47.7**	**69.3**

Bold denotes the best performance

Ablation on the feature matching loss. We measure the effect of the discriminator feature matching loss (FM) in the absence and presence of the perceptual loss (VGG). The discriminator feature matching loss is used by default in SPADE. Table 12 presents the results for OASIS and SPADE+ on Cityscapes. For SPADE+, we observe that the feature matching loss affects the metrics notably only when no perceptual loss is used. In this case, the FM loss improves mIoU by 8.2 points. In contrast, the effect of the FM loss on the mIoU is small when the perceptual loss is used (0.4 points). Hence, the role of the FM loss in the training of SPADE+ is to improve performance by stabilizing the training, similar to the perceptual loss. This observation is in line with the

Table 12 The effect of the discriminator feature matching loss (FM) in the absence or presence of the perceptual loss (VGG)

VGG	FM	FID↓	mIoU↑
(a) OASIS on Cityscapes			
✗	✗	47.7	69.3
✗	✓	48.5	69.1
✓	✗	**46.1**	**72.0**
✓	✓	46.5	70.9
(b) SPADE+ on Cityscapes			
✗	✗	61.4	47.6
✗	✓	57.3	55.8
✓	✗	**47.8**	64.0
✓	✓	48.1	**64.4**

Bold denotes the best performance

general observation that SPADE and other semantic image synthesis models require the help of additional loss functions because the adversarial supervision through the discriminator is not strong enough. In comparison, we did not observe any training collapses in OASIS, despite not using any extra loss functions. For OASIS, the feature matching loss results in a worse FID (by 0.8 points) in the absence of the perceptual loss. We also observe a degradation of 1.1 mIoU points through the FM loss, in the case where the perceptual supervision is present. This indicates that the FM loss negatively affects the strong supervision from the semantic segmentation adversarial loss of OASIS.

5 Conclusion

This work studies semantic image synthesis, the task of generating diverse and photorealistic images from semantic label maps. Conventionally, semantic image synthesis GAN models employed a perceptual VGG loss to overcome training instabilities and improve the synthesis quality. In our experiments we demonstrated that the VGG-based perceptual loss imposes unnecessary constraints on the feature space of the generator, significantly limiting its ability to produce diverse samples from input noise, as well as the ability to produce images with colors and textures closely matching the distribution of real images. Therefore, in this work we propose OASIS, a semantic image synthesis model that needs only adversarial supervision to achieve high-quality results.

The improvement over the prior work in image synthesis quality is achieved via the detailed spatial and semantic-aware supervision from our novel segmentation-based discriminator, which uses semantic label maps as ground truth for training. With this powerful discriminator, OASIS can easily generate diverse outputs from the same semantic label map by resampling 3D noise, eliminating the need for additional image encoders to achieve multi-modality. The proposed 3D noise injection scheme can work both in a global and local regime, allowing to change the appearance of the whole scene and of individual objects. With the proposed modifications, OASIS significantly improves over previous state-of-the-art models in terms of image synthesis quality.

Furthermore, we proposed to use the LVIS dataset to evaluate semantic image synthesis under severe class imbalance and sparse label annotations. Thanks to the class balancing mechanism enabled by its segmentation-based discriminator, OASIS achieves more realistic synthesis of underrepresented classes, achieving pronounced gains on the extremely unbalanced LVIS dataset. Lastly, the design of OASIS can be better suited for image editing applications compared to the SPADE baseline, enabling diverse resampling of scenes from unlabelled images, as well as for synthetic data augmentation, improving the performance of a downstream segmentation network by a larger margin.

Acknowledgements Juergen Gall has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2070 -390732324 and the ERC Consolidator Grant FORHUE (101044724).

References

Alharbi, Y., & Wonka, P. (2020). Disentangled image generation through structured noise injection. In *Conference on computer vision and pattern recognition (CVPR)*.

Arjovsky, M., & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations (ICLR)*.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2016). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *Transactions on Pattern Analysis and Machine Intelligence, 39*, 2481–2495.

Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *International conference on learning representations (ICLR)*.

Bruna, J., Sprechmann, P., & LeCun, Y. (2016). Super-resolution with deep convolutional sufficient statistics. In *International conference on learning representations (ICLR)*.

Caesar, H., Uijlings, J., & Ferrari, V. (2018). Coco-stuff: Thing and stuff classes in context. In *Conference on computer vision and pattern recognition (CVPR)*.

Casanova, A., Careil, M., Verbeek, J., Drozdzal, M., & Romero Soriano, A. (2021). Instance-conditioned gan. In *Advances in neural information processing systems (NeurIPS)*.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2015). Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International conference on learning representations (ICLR)*.

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision (ECCV)*.

Chen, Q., & Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. In *International conference on computer vision (ICCV)*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Conference on computer vision and pattern recognition (CVPR)*.

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in neural information processing systems (NeurIPS)*.

Deng, J., Dong, W., Socher, R., Li, LJ., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Conference on computer vision and pattern recognition (CVPR)*.

Gatys, L., Ecker, A. S., Bethge, M. (2015). Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems (NeurIPs)*.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems (NeurIPS)*.

Gupta, A., Dollar, P., & Girshick, R. (2019). LVIS: A dataset for large vocabulary instance segmentation. In *Conference on computer vision and pattern recognition (CVPR)*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in neural information processing systems (NeurIPS)*.

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *European conference on computer vision (ECCV)*.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020a). Training generative adversarial networks with limited data. In *Advances in neural information processing systems (NeurIPS)*.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020b). Analyzing and improving the image quality of stylegan. In *Conference on computer vision and pattern recognition (CVPR)*.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. In *Advances in neural information processing systems (NeurIPS)*.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations (ICLR)*.

Li, K., & Malik, J. (2018). Implicit maximum likelihood estimation. arXiv:1809.09087.

Li, K., Zhang, T., & Malik, J. (2019). Diverse image synthesis from semantic layouts via conditional imle. In *International conference on computer vision (ICCV)*.

Li, Y., Li, Y., Lu, J., Shechtman, E., Lee, Y. J., & Singh, K. K. (2021). Collaging class-specific gans for semantic image synthesis. In *International conference on computer vision (ICCV)*.

Liu, B., Zhu, Y., Song, K., & Elgammal, A. (2021). Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International conference on learning representations (ICLR)*.

Liu, X., Yin, G., Shao, J., Wang, X., & Li, H. (2019). Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in neural information processing systems (NeurIPS)*.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv:1411.1784

Miyato, T., & Koyama, M. (2018). cGANs with projection discriminator. In *International conference on learning representations (ICLR)*.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International conference on learning representations (ICLR)*.

Ntavelis, E., Romero, A., Kastanis, I., Van Gool, L., & Timofte, R. (2020). Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In *European conference on computer vision (ECCV)*.

Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition, 29*, 51–59.

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019a). Gaugan: Semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH*.

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019b). Semantic image synthesis with spatially-adaptive normalization. In *Conference on computer vision and pattern recognition (CVPR)*.

Park, T., Efros, A. A., Zhang, R., & Zhu, J. Y. (2020). Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision (ECCV)*.

Qi, X., Chen, Q., Jia, J., & Koltun, V. (2018). Semi-parametric image synthesis. In *Conference on computer vision and pattern recognition (CVPR)*.

Reed, S. E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In *International conference on machine learning (ICML)*.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D. (2021). Encoding in style: A stylegan encoder for image-to-image translation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision (IJCV), 40*, 99–121.

Sauer, A., Chitta, K., Müller, J., & Geiger, A. (2021). Projected gans converge faster. In *Advances in neural information processing systems (NeurIPS)*.

Schönfeld, E., Schiele, B., & Khoreva, A. (2020). A u-net based discriminator for generative adversarial networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Schönfeld, E., Sushko, V., Zhang, D., Gall, J., Schiele, B., & Khoreva, A. (2021). You only need adversarial supervision for semantic image synthesis. In *International conference on learning representations (ICLR)*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations (ICLR)*.

Souly, N., Spampinato, C., & Shah, M. (2017). Semi supervised semantic segmentation using generative adversarial network. In *International conference on computer vision (ICCV)*.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*.

Tan, Z., Chen, D., Chu, Q., Chai, M., Liao, J., He, M., Yuan, L., & Yu, N. (2020). Rethinking spatially-adaptive normalization. arXiv:2004.02867.

Tang, H., Bai, S., & Sebe, N. (2020a). Dual attention gans for semantic image synthesis. In *ACM international conference on multimedia*.

Tang, H., Qi, X., Xu, D., Torr, P. H., & Sebe, N. (2020b). Edge guided gans with semantic preserving for semantic image synthesis. arXiv:2003.13898.

Tang, H., Xu, D., Yan, Y., Torr, PH. ., & Sebe, N. (2020c). Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Conference on computer vision and pattern recognition (CVPR)*.

Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C. C., & Lin, D. (2021a). Seesaw loss for long-tailed instance segmentation. In *Conference on computer vision and pattern recognition (CVPR)*.

Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional GANs. In *Conference on computer vision and pattern recognition (CVPR)*.

Wang, Y., Qi, L., Chen, Y. C., Zhang, X., & Jia, J. (2021b). Image synthesis via semantic composition. In *ICCV*.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *Asilomar conference on signals, systems & computers*.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*.

Yaz, Y., Foo, C. S., Winkler, S., Yap, K. H., Piliouras, G., & Chandrasekhar, V. (2018). The unusual effectiveness of averaging in gan training. In *International conference on learning representations (ICLR)*.

Yu, F., Koltun, V., & Funkhouser, T. (2017) Dilated residual networks. In *Conference on computer vision and pattern recognition (CVPR)*.

Yun, S,, Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International conference on computer vision (ICCV)*.

Zhang, D., & Khoreva, A. (2019). PA-GAN: Improving GAN training by progressive augmentation. In *Advances in neural information processing systems (NeurIPS)*.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2018a). StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 41,* 1947–1962.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., & Smola, A. (2020). Resnest: Split-attention networks. arXiv:2004.08955.

Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., & Yang, Y. (2021). Cross-modal contrastive learning for text-to-image generation. In *Conference on computer vision and pattern recognition (CVPR)*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018b). The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on computer vision and pattern recognition (CVPR)*.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Conference on computer vision and pattern recognition (CVPR)*.

Zhu, Z., Xu, Z., You, A., & Bai, X. (2020). Semantically multi-modal image synthesis. In *Conference on computer vision and pattern recognition (CVPR)*.

C One-Shot GAN: Learning to Generate Samples from Single Images and Videos

In this appendix, we provide the conference publication that Chapter 5 of the thesis is based on:

- **One-Shot GAN: Learning to Generate Samples from Single Images and Videos**
  Vadim Sushko, Juergen Gall, Anna Khoreva
  IEEE Computer Vision and Pattern Recognition Conference (CVPR) Workshops, 2021.
  DOI: 10.1109/CVPRW53098.2021.00293

One-Shot GAN: Learning to Generate Samples from Single Images and Videos

Vadim Sushko
Bosch Center for Artificial Intelligence
vadim.sushko@bosch.com

Jürgen Gall
University of Bonn
gall@iai.uni-bonn.de

Anna Khoreva
Bosch Center for Artificial Intelligence
anna.khoreva@bosch.com

Figure 1: Our proposed One-Shot GAN needs only one video (first two rows) or one image (last two rows) for training. At inference phase, it generates novel scene compositions with varying content and layouts. E.g., from a single video with a car on a road, One-Shot GAN can generate the scene without the car or with two cars, and for a single air balloon image, it produces layouts with different number and placement of the balloons. (Original samples are shown in grey or red frames.)

Abstract

Training GANs in low-data regimes remains a challenge, as overfitting often leads to memorization or training divergence. In this work, we introduce One-Shot GAN that can learn to generate samples from a training set as little as one image or one video. We propose a two-branch discriminator, with content and layout branches designed to judge the internal content separately from the scene layout realism. This allows synthesis of visually plausible, novel compositions of a scene, with varying content and layout, while preserving the context of the original sample. Compared to previous single-image GAN models, One-Shot GAN achieves higher diversity and quality of synthesis. It is also not restricted to the single image setting, successfully learning in the introduced setting of a single video.

1. Introduction

Without sufficient training data, GANs are prone to overfitting, which often leads to mode collapse and training instabilities [9, 4]. This dependency on availability of training data limits the applicability of GANs in domains where collecting a large dataset is not feasible. In some real-world applications, collection even of a small dataset remains challenging. It may happen that rare objects or events are present only in one image or in one video, and it is difficult to obtain a second one. This, for example, includes pictures of exclusive artworks or videos of traffic accidents recorded in extreme conditions. Enabling learning of GANs in such *one-shot* scenarios has thus a potential to improve their utilization in practice. Previous work [9, 4] studied one-shot image generation in the context of learning from a *single image*. In this work, we introduce a novel setup of learn-

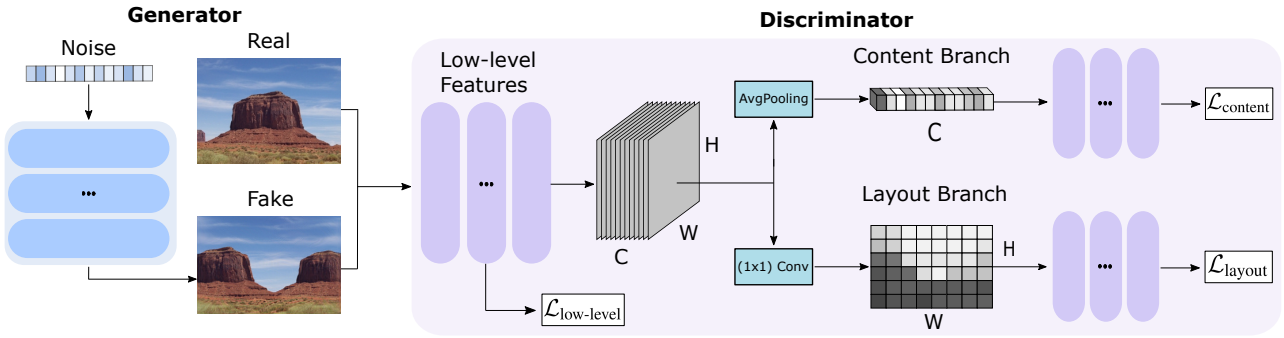

Figure 2: One-Shot GAN. The two-branch discriminator judges the content distribution separately from the scene layout realism and thus enables the generator to produce images with varying content and global layouts. See Sec. 2 for details.

ing to generate new images from frames of a *single video*. In practice, recoding a video lasting for several seconds can take almost as little effort as collecting one image. However, a video contains much more information about the scene and the objects of interest (e.g., different poses and locations of objects, various camera views). Learning from a video can enable generation of images of higher quality and diversity, while still operating in a one-shot mode, and therefore can improve its usability for applications.

To mitigate overfitting in the one-shot mode, recent single-image GAN models [9, 4] proposed to learn an image patch-based distribution at different image scales. Though these models overcome memorization, producing different versions of a training image, they cannot learn high-level semantic properties of the scene. They thus often suffer from incoherent shuffling of image patches, distorting objects and producing unrealistic layouts (see Fig. 3 and 4). Other low-data GAN models, such as FastGAN [6], have problems with memorization when applied in the one-shot setting (see Sec. 3). In this work, we go beyond patch-based learning, seeking to generate novel plausible compositions of objects in the scene, while preserving the original image context. Thus, we aim to keep novel compositions visually plausible, with objects preserving their appearance, and the scene layout looking realistic to a human eye.

To this end, we introduce One-Shot GAN, an unconditional single-stage GAN model, which can generate images that are significantly different from the original training sample while preserving its context. This is achieved by two key ingredients: the novel design of the discriminator and the proposed diversity regularization technique for the generator. The new One-Shot GAN discriminator has two branches, responsible for judging the content distribution and the scene layout realism of images separately from each other. Disentangling the discriminator decision about the content and layout helps to prevent overfitting and provides more informative signal to the generator. To achieve high diversity of generated samples, we also extend the regularization technique of [13, 2] to one-shot unconditional image synthesis. As we show in Sec. 3, The proposed One-Shot GAN generates high-quality images that are significantly different from training data. One-Shot GAN is the first model that succeeds in learning from both single images and videos, improving over prior work [9, 6] in image quality and diversity in these one-shot regimes.

2. One-Shot GAN

Content-Layout Discriminator. We introduce a solution to overcome the memorization effect but still to generate images of high quality in the one-shot setting. Building on the assumption that to produce realistic and diverse images the generator should learn the appearance of objects and combine them in a globally-coherent way in an image, we propose a discriminator judging the *content* distribution of an image separately from its *layout* realism. To achieve the disentanglement, we design a two-branch discriminator architecture, with separate content and layout branches (see Fig.2). Our discriminator D consists of the low-level feature extractor $D_{low\text{-}level}$, the content branch $D_{content}$, and the layout branch D_{layout}. Note that the branching happens after an intermediate layer in order to learn a relevant representation. $D_{content}$ judges the content of this representation irrespective from its spatial layout, while D_{layout}, inspects only the spatial information. Inspired by the attention modules of [7, 12], we extract the *content* from intermediate representations by aggregating spatial information via global average pooling, and obtain *layout* by aggregating channels via a simple (1×1) convolution. This way, the content branch judges the fidelity of objects composing the image independent of their spatial location, while the layout branch is sensitive only to the realism of the global scene layout. Note that $D_{content}$ and D_{layout} receive only limited information from previous layers, which prevents overfitting. This helps to overcome the memorization of training data and to produce different images.

Diversity regularization. To improve variability of generated samples, we propose to add diversity regularization (DR) loss term $\mathcal{L}_{DR}$ to the objective. Previously proposed

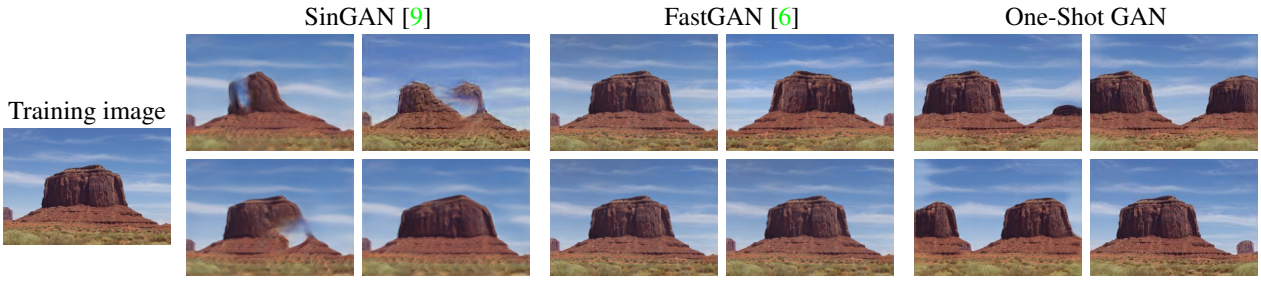

Figure 3: Comparison with other methods in the single image setting. Single-image GAN of [9] is prone to incoherently shuffle image patches (e.g. sky textures appear on the ground), and the few-shot FastGAN model [6] collapses to producing the original image or its flipped version. In contrast, One-Shot GAN produces diverse images with realistic *global* layouts.

Method	Single Image				Single Video			
	SIFID↓	LPIPS↑	MS-SSIM ↓	Dist. to train	SIFID↓	LPIPS↑	MS-SSIM ↓	Dist. to train
SinGAN [9]	0.13	0.26	0.69	0.30	2.47	0.32	0.65	0.51
FastGAN [6]	0.13	0.18	0.77	0.11	0.79	**0.43**	0.55	0.13
One-Shot GAN	**0.08**	**0.33**	**0.63**	0.37	**0.55**	**0.43**	**0.54**	0.34

Table 1: Comparison in the Single Image and Single Video settings on DAVIS-YFCC100M [8, 10] dataset.

regularization terms [13, 2] aimed to encourage the generator to produce different outputs depending on the input latent code, in such a way that the generated samples with closer latent codes should look more similar to each other, and vice versa. In contrast, in the one-shot image synthesis setting, the perceptual distance of the generated images should not be dependent on the distance between their latent codes. As we operate in one semantic domain, the generator should produce images that are in-domain but more or less equally different from each other and substantially different from the original sample. Thus, we propose to encourage the generator to produce perceptually different image samples independent of their distance in the latent space. $\mathcal{L}_{DR}$ is expressed as follows:

$$\mathcal{L}_{DR}(G) = \mathbb{E}_{z_1,z_2}\big[\frac{1}{L}\sum_{l=1}^{L}\|G^l(z_1) - G^l(z_2)\|\big], \quad (1)$$

where $\|\cdot\|$ denotes the $L1$ norm, $G^l(z)$ indicates features extracted from the l-th block of the generator G given the latent code z. Contrary to prior work, we compute the distance between samples in the feature space, enabling more meaningful diversity of the generated images, as different generator layers capture various image semantics, inducing both high- and low-level diversity.

Final objective. We compute adversarial loss for each discriminator part: $D_{low\text{-}level}$, $D_{content}$, and D_{layout}. This way, the discriminator decision is based on low-level details of images, such as textures, and high-scale properties, such as content and layout. The overall adversarial loss is

$$\mathcal{L}_{adv}(G, D) = \mathcal{L}_{D_{content}} + \mathcal{L}_{D_{layout}} + 2\mathcal{L}_{D_{low\text{-}level}}, \quad (2)$$

where $\mathcal{L}_{D_*}$ is the binary cross-entropy $\mathbb{E}_x[\log D_*(x)] + \mathbb{E}_z[\log(1 - D_*(G(z)))]$ for real image x and generated image $G(z)$. As the two branches of the discriminator operate at high-level image features, contrary to only one $D_{low\text{-}level}$ operating at low-level features, we double the weighting for the $\mathcal{L}_{D_{low\text{-}level}}$ loss term. This is done in order to properly balance the contributions of different feature scales and encourage the generation of images with good low-level details, coherent contents and layouts.

The overall One-Shot GAN objective can be written as:

$$\min_G \max_D \; \mathcal{L}_{adv}(G, D) - \lambda\mathcal{L}_{DR}(G), \quad (3)$$

where λ controls the strength of the diversity regularization and $\mathcal{L}_{adv}$ is the adversarial loss in Eq. 2.

Implementation. The One-Shot GAN discriminator uses ResNet blocks, following [1]. We use three ResNet blocks before branching and four blocks for the content and layout branches. We employ standard image augmentation strategies for the discriminator training, following [5]. λ for $\mathcal{L}_{DR}$ in Eq. 3 is set to 0.15. We use the ADAM optimizer with $(\beta_1, \beta_2) = (0.5, 0.999)$, a batch size of 5 and a learning rate of 0.0002 for both G and D.

3. Experiments

Evaluation settings. We evaluate One-Shot GAN on two different one-shot settings: training on a single image and a single video. We select 15 videos from DAVIS [8] and YFCC100M [10] datasets. In the Single Video setting, we use all frames of a video as training images, while for the Single Image setup we use only one middle frame. The chosen videos last for 2-10 seconds and consist of 60-100 frames. To assess the quality of generated images, we measure single FID (SIFID) [9]. Image diversity is assessed by the average LPIPS [3] and MS-SSIM [11] across pairs of

Figure 4: Comparison with other methods in the single video setting. While other models fall into reproducing the training frames or fail to learn textures, One-Shot GAN generates high-quality images significantly different from the original video.

generated images. To verify that the models do not simply reproduce the training set, we report average LPIPS to the nearest image in the training set, augmented in the same way as during training (Dist. to train). We compare our model with a single image method SinGAN [9] and with a recent model on few-shot image synthesis, FastGAN [6].

Main results. Table 1 presents quantitative comparison between the models in the Single Image and Video settings, while the respective visual results are shown in Fig. 3 and 4. As seen from Table 1, One-Shot GAN notably outperforms other models in both quality and diversity metrics. Importantly, our model is the only one which successfully learns from both single images and single videos.

As seen from Fig. 1 and 3, in the Single Image setting, One-Shot GAN produces diverse samples of high visual quality. For example, our model can change the number of rocks on the background or change their shapes. Note that such changes keep appearance of objects, preserving original content, and maintain scene layout realism. In contrast, single-image method SinGAN disrespects layouts (e.g. sky textures may appear below horizon), and is prone to modest diversity, especially around image corners. This is reflected in higher SIFID and lower diversity in Table 1. The few-shot FastGAN suffers from memorization, only reproducing the training image or its flipped version. In Table 1 this is reflected in low diversity and small Dist. to train (in red).

In the proposed Single Video setting, there is much more information to learn from, so generative models can learn more interesting combinations of objects and scenes. Fig. 1 and 4 show images generated by the models in this setting. One-Shot GAN produces plausible images that are substantially different from the training frames, adding/removing objects and changing scene geometry. For example, having seen a bus following a road, One-Shot GAN varies the length of a bus and placement of trees. For the video with an equestrian competition, our model can remove a horse from the scene and change the jumping obstacle configuration. In contrast, SinGAN, which is tuned to learn from a single image, does not generalize to this setting, producing "mean in class" textures and failing to learn appearance of objects (low diversity and very high SIFID). FastGAN, on the other hand, learns high-scale scene properties, but fails to augment the training set with non-trivial changes, having a very low distance to the training data (0.13 in Table 1).

Table 1 confirms that the proposed two-branch discriminator in combination with diversity regularization manages to overcome the memorization effect, achieving high distance to training data in both settings (0.37 and 0.34). This means that One-Shot GAN augments the training set with structural transformations that are orthogonal to standard data augmentation techniques, such as horizontal flipping or color jittering. To achieve this, the model requires as little data as one image or one short video clip. We believe, such ability can be especially useful to generate samples for augmentation of limited data, for example by creating new versions of rare examples.

4. Conclusion

We propose One-Shot GAN, a new unconditional generative model operating at different one-shot settings, such as learning from a single image or a single video. At such low-data regimes, our model mitigates the memorization problem and generates diverse images that are structurally different from the training set. Particularly, our model is capable of synthesizing images with novel views and different positions of objects, preserving their visual appearance. We believe, such structural diversity provides a useful tool for image editing applications, as well as for data augmentation in domains, where data collection remains challenging.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. 2, 3

[3] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPs)*, 2016. 3

[4] Tobias Hinz, Matthew Fisher, Oliver Wang, and Stefan Wermter. Improved techniques for training single-image gans. *arXiv preprint arXiv:2003.11512*, 2020. 1, 2

[5] Tero Karras, Miika Aittala, Janne Hellsten, S. Laine, J. Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[6] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized {gan} training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2021. 2, 3, 4

[7] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 2

[8] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 3

[9] Tamar Rott Shaham, Tali Dekel, and T. Michaeli. Singan: Learning a generative model from a single natural image. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 4

[10] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3

[11] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, 2003. 3

[12] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[13] Dingdong Yang, Seunghoon Hong, Y. Jang, T. Zhao, and H. Lee. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019. 2, 3

D One-Shot Synthesis of Images and Segmentation Masks

In this appendix, we provide the conference publication that Chapter 6 of the thesis is based on:

- **One-Shot Synthesis of Images and Segmentation Masks**
  Vadim Sushko, Dan Zhang, Juergen Gall, Anna Khoreva
  IEEE Winter Conference on Applications of Computer Vision (WACV), 2023.
  DOI: 10.1109/WACV56688.2023.00622

One-Shot Synthesis of Images and Segmentation Masks

Vadim Sushko[1] Dan Zhang[1,2] Juergen Gall[3] Anna Khoreva[1,2]
[1]Bosch Center for Artificial Intelligence [2]University of Tübingen [3]University of Bonn
{vadim.sushko,dan.zhang2,anna.khoreva}@bosch.com, gall@iai.uni-bonn.de

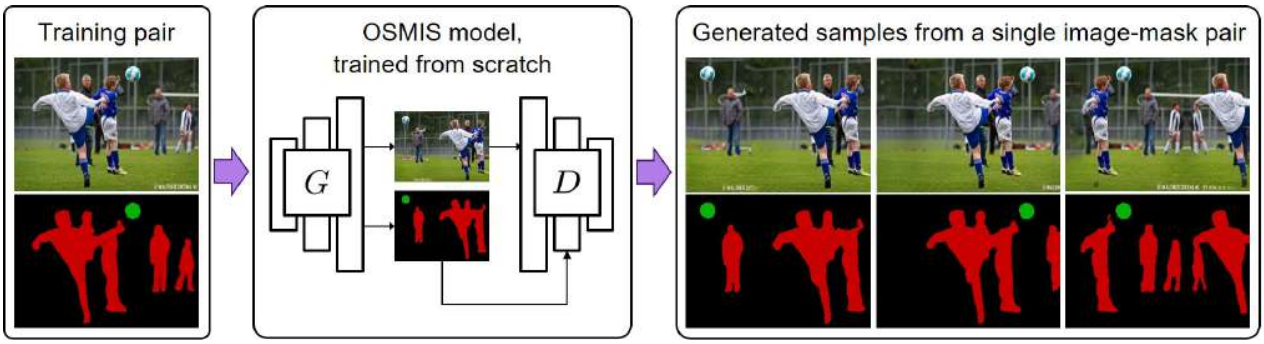

Figure 1. We introduce a new task of generating new images and their segmentation masks from a single training pair, without access to any pre-training data. Under this challenging regime, our proposed GAN model (OSMIS) achieves a synthesis of a high structural diversity, preserving the photorealism of original images and a precise alignment of produced segmentation masks to the generated content.

Abstract

Joint synthesis of images and segmentation masks with generative adversarial networks (GANs) is promising to reduce the effort needed for collecting image data with pixel-wise annotations. However, to learn high-fidelity image-mask synthesis, existing GAN approaches first need a pre-training phase requiring large amounts of image data, which limits their utilization in restricted image domains. In this work, we take a step to reduce this limitation, introducing the task of one-shot image-mask synthesis. We aim to generate diverse images and their segmentation masks given only a single labelled example, and assuming, contrary to previous models, no access to any pre-training data. To this end, inspired by the recent architectural developments of single-image GANs, we introduce our OSMIS model which enables the synthesis of segmentation masks that are precisely aligned to the generated images in the one-shot regime. Besides achieving the high fidelity of generated masks, OSMIS outperforms state-of-the-art single-image GAN models in image synthesis quality and diversity. In addition, despite not using any additional data, OSMIS demonstrates an impressive ability to serve as a source of useful data augmentation for one-shot segmentation applications, providing performance gains that are complementary to standard data augmentation techniques. Code is available at `https://github.com/boschresearch/one-shot-synthesis`.

1. Introduction

Deep neural networks have been shown powerful at solving various segmentation problems in computer vision [8, 10, 14, 23, 21, 32]. The success of these segmentation models strongly relies on the availability of a large-scale collection of labelled data for training. Nevertheless, annotation of a large dataset is not always feasible in practice due to a very high cost of manual labelling of segmentation masks [7]. For example, accurately labelling a single image with many objects can take more than 30 minutes [35]. Therefore, diminishing the human effort required for obtaining diverse and precisely aligned image-mask data is an important problem for many practical applications.

Recently, several works [30, 35, 15, 26] proposed to tackle this issue by jointly generating images and segmentation masks with generative adversarial networks (GANs). Utilizing a few provided pixel-level annotations in addition to an image dataset for training, such GAN models become a source of labelled data that can be used to train neural networks in various practical applications. Despite achieving impressive synthesis of segmentation masks based on limited annotated examples, existing image-mask GAN models still require large pre-training image datasets to learn high-fidelity image synthesis. This naturally restricts their application only to the data domains where such datasets are available (e.g., images of faces or cars). However, in some practical scenarios such a dataset can be difficult to find, for

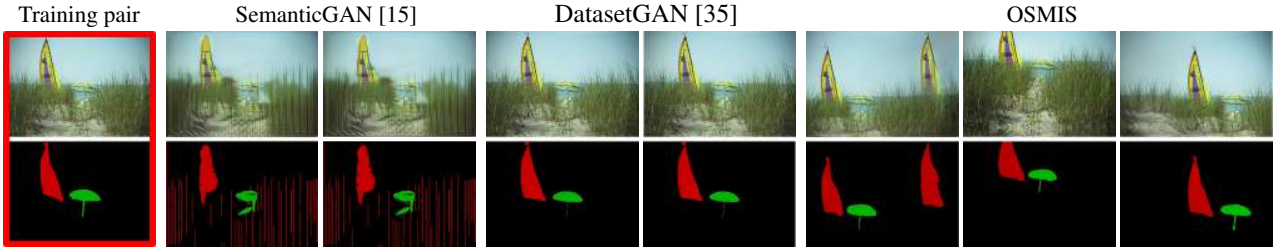

Figure 2. A comparison to SemanticGAN [15], trained on a single image-mask pair (in red), and DatasetGAN [35], pre-trained on a single image and trained on a single manual mask annotation. Both models suffer from memorization, while SemanticGAN also has poor quality due to training instabilities. In contrast, OSMIS avoids mode collapse and generates diverse high-quality samples. This is achieved by means of a discriminator that judges the realism of different objects separately, which prevents memorization of the whole given image.

example in one-shot segmentation applications [1], where the object types can be rare. Therefore, in this work we aim to learn a high-fidelity joint mask and image synthesis having as little limitations on the data domain as possible. To this end, we propose a novel GAN training setup, in which we assume availability only of a single training image and its segmentation mask, not relying on any image dataset for pre-training (see Fig. 1). After training, we aim to generate diverse new image samples and supplement them with accurate segmentation masks. To the best of our knowledge, we are the first to consider such a training scenario for GANs.

Training a GAN from a single training sample is well known to be challenging due to the problem of memorization [20], as in many cases the generator converges to reproducing the exact copies of training data. For example, as shown in our experiments, this issue occurs in the prior image-mask GAN models from [15, 35] (see Fig. 2). Recently, the issue of memorization has been mitigated in the line of works on single-image GANs, which enabled diverse image synthesis from a single training image [27, 12, 28]. Inspired by these models, we aim to extend this ability to a joint synthesis of images and segmentation masks. To this end, we propose a new model, introducing two modifications to conventional GAN architectures. Firstly, we introduce a mask synthesis branch for the generator, enabling the synthesis of segmentation masks in addition to images. Secondly, to ensure that the produced segmentation masks are precisely aligned to the generated image content, we propose a masked content attention module for the discriminator, allowing it to judge the realism of different objects separately from each other. This way, to fool the discriminator, the generator is induced to label synthesized images accurately. In effect, our proposed model enables a structurally diverse, high-quality **o**ne-**s**hot joint **m**ask and **i**mage **s**ynthesis (see Fig. 1), and we thus name it **OSMIS**. As we show in our experiments, compared to prior single-image GANs [27, 12, 28], OSMIS not only offers an additional ability to generate accurate segmentation masks, but also achieves higher quality and diversity of generated images.

Despite using only a single image-mask pair for training, OSMIS can generate a set of labelled samples of a high structural diversity, which sometimes cannot be achieved with standard data augmentation techniques (e.g., flipping, zooming, or rotation). For example, for a given scene, OSMIS can change the relative locations of foreground objects or edit the layout of backgrounds (see Fig. 1, 4, 5). Moreover, in contrast to [15, 35], OSMIS can successfully handle masks of different types, e.g., having class-wise (see Fig. 1) or instance-wise (see Fig. 4) annotations. This suggests a good potential of our model to serve as a source of additional labelled data augmentation for practical applications. We demonstrate this potential in Sec. 4.2, where we apply OSMIS at the test phase of one-shot video object segmentation [23] and one-shot semantic image segmentation [1]. The results indicate that the data generated by OSMIS helps to improve the performance of state-of-the-art networks: OSVOS [6], STM [22], and RePRI [5], providing complementary gains to standard data augmentation. We find these results promising for utilization of one-shot image-mask synthesis in future research.

2. Related Work

GANs generating segmentation masks. Recently, it was observed that a GAN generator, trained on a large dataset, implicitly learns discriminative pixel-wise features of the generated scene objects [30]. Thus, several works proposed to collect feature activations from different generator layers and transform them into a segmentation mask using a small decoder. RepurposeGAN [30] and DatasetGAN [35] proposed to train the decoder using a handful of manually annotated generated images. LinearGAN [33] replaced manual annotations by the predictions of an external segmentation network. Alternatively, SemanticGAN [15] and EditGAN [18] enforced the alignment between generated images and masks with the loss from an additional discriminator, which takes both images and masks as inputs.

Although the above models require only a few masks to achieve high-quality image-mask synthesis, they are not successful when the number of training images is not sufficient. For example, DatasetGAN and SemanticGAN suffer from instabilities and memorization issues when trained on a single image-mask pair (see Fig. 2 and A in the sup-

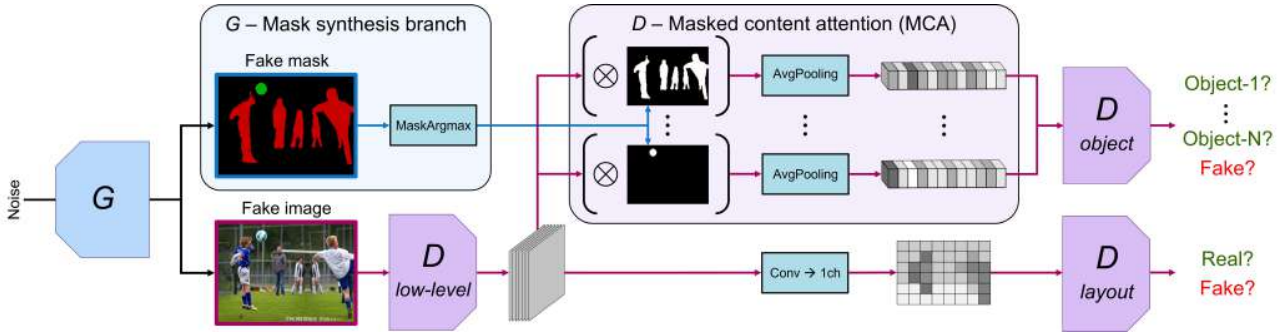

Figure 3. OSMIS model. A simple mask synthesis branch in the generator *G* allows the generation of segmentation masks of objects together with images. The precise alignment between the masks and the generated image content is enforced by a masked content attention (MCA) module in the discriminator *D*, designed to evaluate the realism of different objects separately from each other.

plementary material.). In contrast, our model learns in this regime successfully, as it does not rely on large-scale pre-training data. As shown in experiments, this makes our model better suited for the scenarios dealing with restricted data domains, such as one-shot segmentation applications. Furthermore, our model is trained in a purely adversarial fashion without any additional overhead, e.g., not requiring manual annotations of generated images, external segmentation networks, or additional discriminators.

Single Image GANs. A line of works investigated unconditional GAN training using only a single image. Under such critically low-data regime, the models are susceptible to training instabilities, as the discriminator can simply memorize the training sample and provide uninformative gradients to the generator [13]. SinGAN [27] proposed to mitigate this issue using a cascade of GANs, where each GAN stage is restricted to learn only the patch distribution at a certain image scale. ConSinGAN [12] improved the performance and efficiency of SinGAN by rebalancing the training of different GAN stages and by training several stages concurrently. Since then, numerous further variations of multi-stage GAN training have been proposed [2, 9, 4, 11]. More recently, One-Shot GAN [28] proposed a two-branch content-layout discriminator, trained as a single stage, enabling the synthesis of images with content and layouts significantly differing from the original sample. Our paper has a similar motivation to the above works, since we also aim to train a GAN model on a single data instance. However, we extend the single image setup with the synthesis of segmentation masks, which no prior work has considered, to the best of our knowledge.

3. Method

Given a single image with its pixel-level segmentation mask and assuming no access to any pre-training data, we aim to generate a diverse set of new image-mask pairs. In this section, we present OSMIS, our one-shot image-mask synthesis model. Adopting One-Shot GAN [28] as a state-of-the-art image synthesis baseline (Sec. 3.1), we propose modifications to the generator and discriminator architecture, enabling one-shot synthesis of segmentation masks that are precisely aligned with generated images (Sec. 3.2).

3.1. One-Shot GAN baseline

As the baseline network architecture, we select the state-of-the-art model One-Shot GAN [28], as it achieves the highest quality and diversity of one-shot image synthesis among previous works. One-Shot GAN proposed a two-branch discriminator, in which an input image x is first transformed into a feature representation $F(x)$ by a low-level discriminator $\mathcal{D}_{low-level}$. Next, two separate discriminators assess different aspects of $F(x)$. The content discriminator $\mathcal{D}_{content}$ judges the realism of objects regardless of their spatial location by averaging out the spatial information contained in $F(x)$ via global average pooling. On the other hand, the layout discriminator $\mathcal{D}_{layout}$ evaluates the realism only of the spatial scene layouts by squeezing $F(x)$ with a one-channel convolution. In addition, the discriminator applies feature augmentation in the content and layout representations of $F(x)$ to further increase the high-level diversity among generated samples. The adversarial loss of the One-Shot GAN model consists of three terms:

$$\mathcal{L}_{adv}(G, D) = \mathcal{L}_{D_{content}} + \mathcal{L}_{D_{layout}} + 2\mathcal{L}_{D_{low\text{-}level}}, \quad (1)$$

where each term is the mean of binary cross entropies obtained at different layers of respective discriminator parts.

3.2. OSMIS model

In contrast to one-shot image synthesis, we assume that the single training image is provided with its pixel-level mask of objects, not assuming any fixed annotation type (e.g., class-wise or instance-wise). To incorporate it into the training process, we introduce two modifications to the architecture of the baseline model. Firstly, we propose to generate segmentation masks simultaneously with images via an additional generator's mask synthesis branch. Secondly, to enforce the precise mask alignment to the generated image content, we re-formulate the objective of the content discriminator $\mathcal{D}_{content}$, designing it to judge the fidelity of different objects separately from each other. This

is made possible by the introduced masked content attention module, which builds a separate content feature vector for each object considering the provided segmentation mask. The overview of our model architecture is shown in Fig. 3. Next, we describe the proposed modifications in detail.

Mask synthesis branch in the generator. In line with [30, 35], we hypothesize that during training the generator should be able to learn discriminative features that completely describe the appearance of generated objects. Thus, while synthesizing an image, we collect feature activations of the generator layers and use them as input for the mask synthesis branch. In contrast to [30, 35], we use only the activations after the last generator block, as this simplest solution already performs well in our experiments. Using a simple convolution followed by a softmax activation, we transform these features into an N-channel soft probability map, where each channel corresponds to one of $N-1$ objects of interest in the segmentation mask or to the background. To obtain the final discrete mask prediction, an argmax operation T along the channel dimension is applied.

To enable the training of the mask synthesis branch with the discriminator loss, the generated masks should allow back-propagation of gradients, similarly to generated images. In our experiments, feeding the discriminator the continuous segmentation probability maps obtained before the non-differentiable argmax operation T impaired the GAN training, as the discriminator learnt to detect the continuous-discrete discrepancy between fake and real inputs. Thus, inspired from [31, 3], we enable back-propagation through argmax by developing a straight-through gradient estimator:

$$\mathrm{MaskArgmax}(y) = y + T(y) - sg[y], \tag{2}$$

where sg denotes a stop-gradient operation. This way, the discriminator is provided with the generated masks in a discrete form $T(y)$, which enables its effective training, while the generator can be trained with the gradients passing through its probability map prediction y.

Yet, this solution can sometimes lead to degenerate solutions, e.g., when all the pixels are predicted as the background channel. This cannot be corrected during training, as in this case the gradient flow through all the other mask channels is blocked. We found that it can be mitigated by softening the argmax operation T at the beginning of training. For this, during the first P_0 epochs we regard each mask pixel as a random variable following Bernoulli distribution:

$$T(y) = \begin{cases} \sim \mathrm{Bernoulli}(y) & \text{epoch} < P_0, \\ \mathrm{argmax}(y) & \text{epoch} \geq P_0. \end{cases} \tag{3}$$

Masked content attention in the discriminator. To provide a training signal to the generator's mask synthesis branch, we propose to incorporate the learning of the image-mask alignment to the objective of the content discriminator $\mathcal{D}_{content}$. In [28], $\mathcal{D}_{content}$ was designed to judge the content distribution of the whole given image. Considering the provided segmentation mask, we can now select the image areas belonging to different objects, and require the discriminator to learn their appearance separately from each other. With this objective, as the discriminator can compare the appearance of the area belonging to the same object in real and fake images, it encourages the generator not only to synthesize realistic objects, but also to label them correctly.

To this end, we introduce a masked content attention (MCA) module. As shown in Fig. 3, MCA receives a downsampled segmentation mask y along with an intermediate feature representation $F(x) = \mathcal{D}_{low-level}(x)$ of an input image x, and thereout produces a set of N content vectors, corresponding to the masked content representations of each of the $N-1$ objects of interest and the background:

$$\mathrm{MCA}(x, y) = \{\mathrm{AvgPool}\left(F(x) \times \mathbb{1}_{y=i}\right)\}_{i=1}^{N}. \tag{4}$$

Accordingly, we re-design the objective of the content discriminator (further denoted $\mathcal{D}_{object}$). For each of the obtained object representations, our proposed $\mathcal{D}_{object}$ is induced to predict a correct identity of each object or background of a real image, while all the identities of fake images should be categorized as an additional fake class:

$$\begin{aligned} \mathcal{L}_{D_{object}} = &- \mathbb{E}_{(x,y)} \left[\sum_{i=1}^{N} \alpha_i \log \mathcal{D}^i_{object}(\mathrm{MCA}^i(x, y)) \right] \\ &- \mathbb{E}_z \left[\sum_{i=1}^{N} \log(1 - \mathcal{D}^{fake}_{object}(\mathrm{MCA}^i(G(z))) \right], \end{aligned} \tag{5}$$

where z is the noise vector used by the generator G to synthesize a fake image-mask pair $G(z) = \{G_x(z), G_y(z)\}$, (x, y) denotes the real image-mask pair, and $\mathcal{D}^i(*)$ is the discriminator logit for the object i. Considering that different objects or background can occupy different areas, we introduce a class balancing weight α_i, which is the inverse of the per-pixel class frequency in the segmentation mask y:

$$\alpha_i = \frac{(\mathrm{sum}(\mathbb{1}_{y=i}))^{-1}}{\sum_{j=1}^{N}(\mathrm{sum}(\mathbb{1}_{y=j}))^{-1}}. \tag{6}$$

Note that the balancing is applied only for real images, as in Eq. 5 all fake objects are considered as the same class.

Our $\mathcal{D}_{object}$ learns the content distribution of each object separately. The advantage of such a training scheme is two-fold. Firstly, a generator now needs to synthesize correct segmentation masks in order to fool the discriminator. The precise image-mask alignment is thus enforced directly by the adversarial loss, without the need for using additional networks or manual annotation. Secondly, as MCA provides representations only of separate objects, $\mathcal{D}_{object}$ has restricted access to the content distribution of the whole image. In effect, the discriminator memorization of the whole training sample becomes more difficult, which enables more diverse image synthesis (see Table 3).

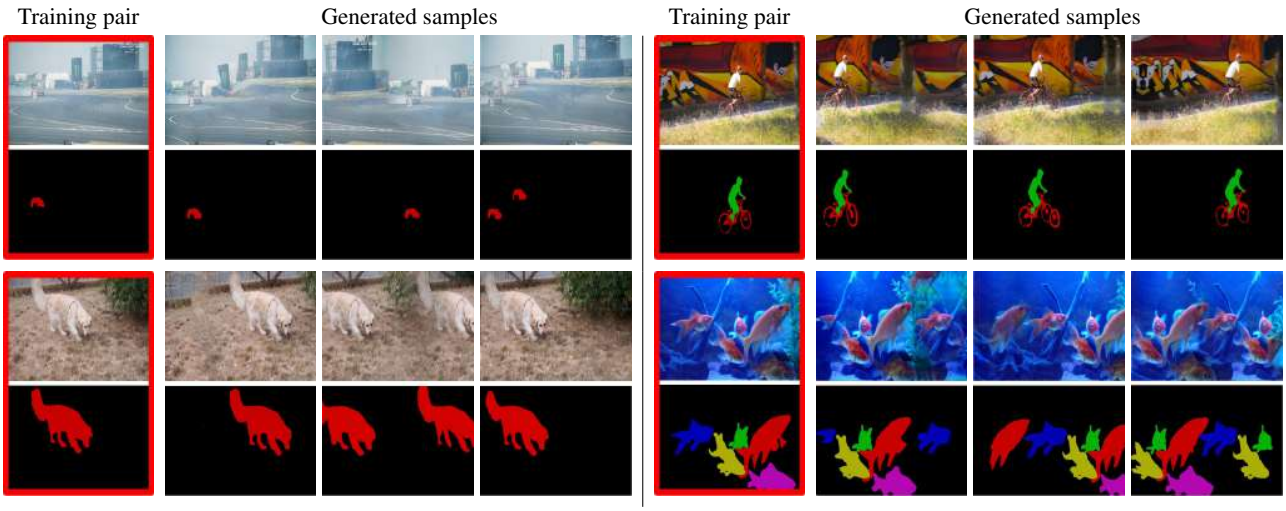

Figure 4. Qualitative results of OSMIS on DAVIS [23]. Given a single image-mask pair for training, our model achieves high-fidelity image synthesis with a high structural diversity, changing the positions of objects or editing the layout of backgrounds. For each synthesized image, it produces segmentation masks that accurately annotate the generated content. Training pairs are shown in red frames.

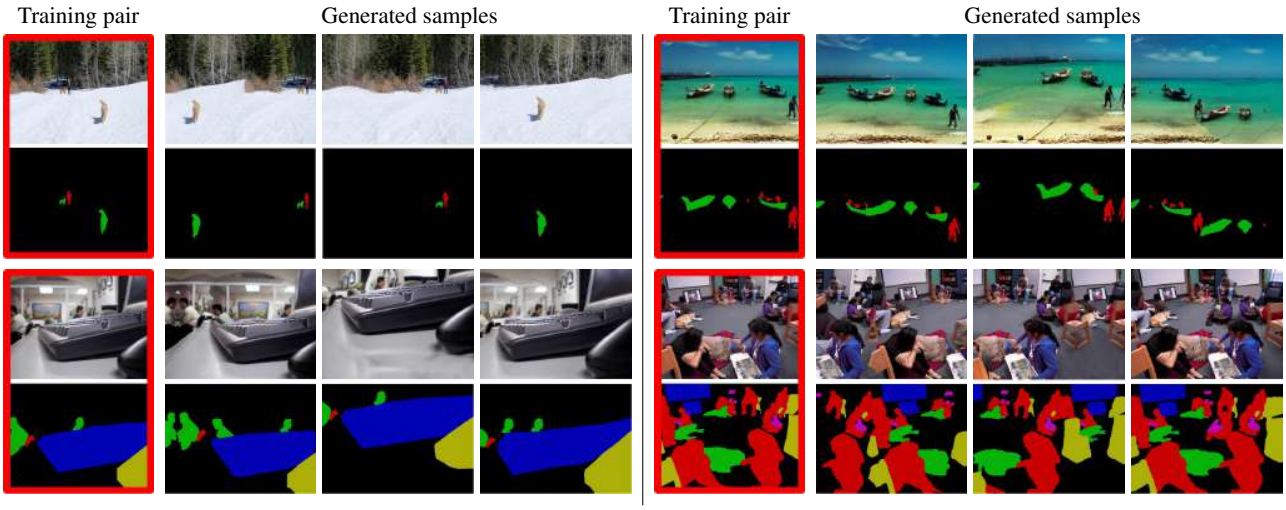

Figure 5. Qualitative results of OSMIS on COCO [17]. OSMIS successfully deals with different scene types and annotation styles. For example, it achieves high quality and diversity for both indoor and outdoor scenes, or sparse and dense annotations of foreground objects.

4. Experiments

We evaluate our model as follows. Firstly, we provide the qualitative and quantitative assessment of the achieved one-shot image-mask synthesis, evaluating the quality and diversity of generated images, as well as their alignment to the produced segmentation masks (Sec. 4.1). Secondly, we apply OSMIS to two one-shot segmentation applications, demonstrating the potential of the generated image-mask pairs to be used as data augmentation (Sec. 4.2).

4.1. Evaluation of one-shot image-mask synthesis

Training details. We train our model with the loss from Eq. (5) for the object discriminator $\mathcal{D}_{object}$, setting P_0=15000. We employ differentiable augmentation (DA) of input images and masks while training the discriminator, using the whole set of transformations as proposed in [13]. We use an exponential moving average of the generator weights with a decay of 0.9999, and follow [28] in setting all the other hyperparameters. More training details are shown in the supplementary material.

Datasets. To evaluate the synthesis, we use the DAVIS dataset [23], originally introduced for video object segmentation. For each video from the DAVIS-17 validation split, we take the first frame and its segmentation mask of objects, which results in 30 image-mask pairs on which we train separate models. The resolution is set to 640x384. For additional visual results, we use samples from COCO [17], trying to closely fit their resolution. Note that the datasets have

Method	SIFID↓	LPIPS↑
SinGAN [27]	0.131	0.267
ConSinGAN [12]	0.103	0.296
One-Shot GAN [29]	0.091	0.347
OSMIS (ours)	**0.073**	**0.387**

Table 1. Comparison of image quality and diversity to single-image GANs on DAVIS-17. Bold denotes the best performance.

Method	SIFID↓	LPIPS↑	mIoU
DatasetGAN [35]	0.118	0.007	91.1*
SemanticGAN [15]	0.211	0.012	65.8
OSMIS (ours)	**0.073**	**0.387**	**86.6**

Table 2. Comparison to prior image-mask GANs on DAVIS-17. Bold denotes the best performance. Red indicates mode collapse. * Indicates manual annotation of masks for DatasetGAN training.

different annotation types (class-wise and instance-wise).

Metrics. To mind a possible quality-diversity trade-off in our one-shot regime [24, 16], we assess the quality and diversity of generated images separately. For this, we report the average SIFID [27] as the measure of image quality, while the average LPIPS [34] between the pairs of generated images is used to assess the diversity of synthesis.

On the other hand, evaluating the quality of generated masks is challenging, because generated images do not have ground truth segmentation annotations. To bypass this issue, we propose to evaluate the alignment between generated masks and synthetic images using an external segmentation network. For this, we take a UNet [25] and train it on the generated image-mask pairs for 500 epochs. After training, we compute its mIoU performance on the original real image, augmented with standard geometric transformations. Intuitively, a good performance on this test reveals that synthetic masks describe well the objects from the real data, indicating precise alignment between the generated images and their masks.

Qualitative results. Fig. 4 and 5 show image-mask pairs generated by OSMIS trained on samples from DAVIS and COCO. Given only a single image-mask pair, our model learns to generate new image-mask pairs, demonstrating a remarkable structural diversity among samples, photorealism of synthesized images, and a high quality of generated annotations. For example, OSMIS can re-synthesize the provided scene with a different number of foreground objects, e.g., more dogs (3$^{\text{rd}}$ example in Fig. 4), less people (2$^{\text{nd}}$ example in Fig. 5), or edit layouts of backgrounds (1$^{\text{st}}$ examples in Fig. 4-5), in all cases providing accurate segmentation masks for the re-synthesized scenes. We note that reaching such structural differences to training data simultaneously with photorealism is extremely difficult from a single sample. For example, it could not be achieved with DatasetGAN or SemanticGAN due to memorization issues and training instabilities (see Fig. 2). Lastly, we remark that

Mask supervision	SIFID↓	LPIPS↑	mIoU
None	0.071	0.368	-
Projection [19]	**0.071**	0.362	72.1
Input concat.	0.079	0.328	82.4
SemanticGAN D_m [15]	0.074	0.351	83.3
MCA (ours)	0.073	**0.387**	**86.6**

Table 3. Comparison of MCA to other mask synthesis supervision mechanisms on DAVIS-17. Red indicates decreased diversity compared to the baseline. Bold denotes the best performance.

OSMIS successfully deals with very different scene types (e.g., both indoor and outdoor scenes), supports masks with both sparse and dense object annotations (e.g., foreground objects occupying small or large image areas), and can handle masks with many objects or even separate instances of the same semantic class (e.g., fish in 4$^{\text{th}}$ example in Fig. 4).

Quantitative results. We compare the quality and diversity of generated *images* to the single-image GAN models SinGAN [27], ConSinGAN [12] and One-Shot GAN [28]. The image-mask synthesis is compared to the previous methods DatasetGAN [35] and SemanticGAN [15]. We use the official repositories provided by the authors.

The quantitative comparison of the image synthesis to single-image GAN models on DAVIS-17 is presented in Table 1. Compared to these models, OSMIS not only offers an additional ability to generate segmentation masks, but also achieves higher image quality and diversity. As seen in Table 1, despite a potential trade-off between SIFID and LPIPS, our model outperforms previously published baselines in both metrics by a notable margin. Further, Table 2 demonstrates that prior image-mask methods, DatasetGAN and SemanticGAN, suffer from instabilities and fail to achieve diverse synthesis, scoring very low in LPIPS.

Ablations. In Table 3 we compare the proposed masked content attention module (MCA) with three alternative discriminator mechanisms to provide supervision for the generator's mask synthesis branch. The simplest baseline is to concatenate the input masks to images, requiring the discriminator to judge their realism jointly. Another method is to use projection [19], by taking the inner product between the last linear layer output of $D_{\text{low-level}}$ and the pixel-wise linear projection of the input mask. Finally, we compare to the approach of SemanticGAN [15], adding a separate discriminator network $\mathcal{D}_m$ which takes both segmentation masks and images, and propagate its gradients only to the generator's mask synthesis branch. While training these baselines, we preserve all the OSMIS hyperparameters, but remove the MCA and use the original $\mathcal{D}_{content}$ as in [28]. As seen from mIoU in Table 3, MCA enables the generation of segmentation masks with the best alignment to the generated image content, as measured by an external segmentation network. Notably, while all the alternative methods negatively affect diversity, MCA improves it (0.387 vs 0.368 LPIPS), highlighting its regularization effect which

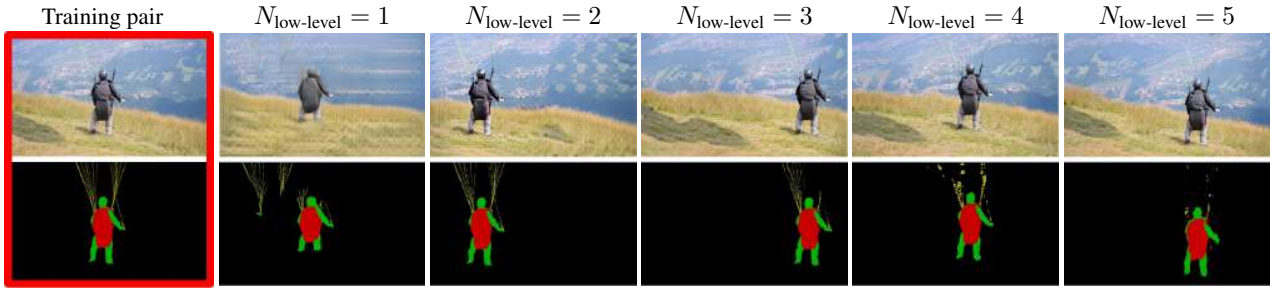

Figure 6. Trade-off between the image and mask quality when varying the number of $\mathcal{D}_{\text{low-level}}$ discriminator blocks. Increased number improves image quality, but harms the ability of masks to capture fine-grained object details due to stronger downsampling during training.

$N_{\text{low-level}}$	SIFID↓	LPIPS↑	mIoU
1	0.262	0.395	82.4
2	0.165	**0.404**	**87.1**
3	0.102	0.394	86.9
4	0.073	0.387	86.6
5	**0.070**	0.321	83.9

Table 4. Ablation on the number of $\mathcal{D}_{\text{low-level}}$ discriminator blocks on DAVIS-17. Bold denotes the best performance.

prevents the discriminator memorization of training data.

While enabling on average higher image diversity and mask quality, we found that MCA can struggle if the training sample contains annotations of fine-grained object details, due to downsampling of input masks. This is illustrated in Fig. 6 and Table 4, for which we train OSMIS with different numbers of low-level discriminator blocks $N_{\text{low-level}}$, corresponding to different degrees of mask downsampling. We observe a trade-off between the quality of images and masks: decreasing $N_{\text{low-level}}$ improves the image diversity and pixel-level mask fidelity, but harms image quality. We selected $N_{\text{low-level}} = 4$ as a compromise between the metrics in Table 4, even though this configuration sometimes fails to annotate small object details (as in Fig. 6). Note that despite this limitation, MCA still outperforms alternative methods that do not use downsampling on DAVIS-17 (see Table 3), and leads to image-mask pairs that are more useful as data augmentation, as discussed next.

4.2. Application to one-shot segmentation tasks

After training, OSMIS can augment the provided image-mask pair with novel diverse samples. As such diversity (edited backgrounds, objects changing relative locations) is difficult to achieve by means of standard data augmentation, we foresee a potential usage of our model as a source of labelled data augmentation. Thus, in what follows, we test the efficacy of OSMIS generations when applied at test phase of two one-shot segmentation applications.

One-shot video object segmentation. We apply our model to the semi-supervised one-shot video segmentation benchmark DAVIS [23]. At test phase, this task provides a video and the segmentation mask of objects only in the first frame, while a model is required to segment all the remaining video frames. We select two popular models from the literature: OSVOS [6], which fine-tunes the network weights on the first video frame and segments other frames independently, and STM [22], which propagates the segmentation prediction sequentially using a space-time memory module. We conduct experiments on two DAVIS splits: *DAVIS-16*, having 20 videos with a single annotated object; and its extension *DAVIS-17*, having 30 videos with multi-instance annotations. To evaluate the video segmentation, we compute the average of the mean mIoU region similarity ($\mathcal{J}$) and the mean contour accuracy ($\mathcal{F}$) across all videos, which is a popular metric for this task [23].

Network	Augmentation: Standard	Augmentation: **Ours**	DAVIS-16	DAVIS-17
OSVOS [6]	✗	✗	76.9	51.3
	✓	✗	78.5 (80.2)	52.9 (52.8)
	✗	✓	78.2	52.6
	✓	✓	**79.8**	**54.2**
STM [22]	✗	✗	89.7 (89.4)	72.4 (72.2)
	✓	✗	89.9	72.4
	✗	✓	90.1	72.6
	✓	✓	**90.2**	**72.7**

Table 5. Effect of data augmentation on the mean of mIoU and contour accuracy ($\mathcal{J}\&\mathcal{F}$) of one-shot video object segmentation. Bold denotes the best performance. Round brackets show the results reported in [6, 22]. Reproduced and reported numbers for OSVOS differ as its official code lacks some model components.

One-shot semantic image segmentation. The second setup is the one-shot image segmentation benchmark COCO-20^i [17]. In this task, a segmentation model is first trained on a large dataset. At test phase, the model is given a single image-mask pair (support set) with an object of a previously unseen test class, and is then required to segment another sample (query image) containing instances of the same class. We conduct experiments with the state-of-the-art RePRI network [5]. COCO-20^i contains 80 classes, which are divided into 4 folds, with 60 base and 20 test classes in each fold. To test OSMIS, we randomly selected 5 support samples for each test class, resulting in 100 image-

Network	Augmentation: Standard	Augmentation: **Ours**	$COCO^0$	$COCO^1$	$COCO^2$	$COCO^3$
RePRI [5]	✗	✗	31.2 (31.2)	38.3 (38.1)	32.9 (33.3)	33.2 (33.0)
	✓	✗	31.8	38.5	33.4	33.8
	✗	✓	32.4	38.7	33.7	34.3
	✓	✓	**32.8**	**39.0**	**34.1**	**34.6**

Table 6. Effect of synthesized data augmentation on mIoU of one-shot image segmentation. In each data split, support examples were sampled from a subset of 100 image-mask pairs, for which our model was trained. Bold denotes the best performance. The round brackets contain the numbers reported in [5].

mask pairs in each of the folds, and trained OSMIS on all of them separately. The performance of this task is evaluated separately for each fold, using the average mIoU across many different support-query examples.

Experimental setup. For both applications, we train OSMIS on the single given image-mask pair (the first video frame or support sample). We try to closely fit the resolution of each image from COCO, and set a fixed resolution of 640x384 for images from the DAVIS benchmark. After training, we generate a pool of synthetic image-mask pairs consisting of $n = 100$ samples. As OSMIS can occasionally fail and synthesize noisy examples, we compute the SIFID metric [27] for each generated image as a measure of its quality. Ranking the images by the average of SIFID ranks at different InceptionV3 layers, we exclude bad-quality samples by filtering out 15% lowest-ranked images. Finally, we add the remaining synthetic samples to the original image-mask pair as data augmentation. See more setup details in Sec. B of the supplementary material.

Among the used segmentation models, only OSVOS [6] applies data augmentation at test phase (random combinations of image-mask flipping, zooming, and rotation). Thus, in experiments we compare our synthetic data augmentation to this pipeline (referred to as *standard* augmentation).

Results. The performance of segmentation networks using different data augmentation is shown in Tables 5 and 6. To account for the variance between runs, all the results are averaged across 5 runs with different seeds for augmentation. We generally managed to reproduce the official reported numbers closely, with the exception of OSVOS, for which the official codebase[1] does not implement the model in full configuration. As seen in Tables 5 and 6, the synthetic data augmentation produced by OSMIS yields a notable increase in segmentation performance, on average improving the metrics of OSVOS and STM by 1.3 and 0.3 $\mathcal{J}\&\mathcal{F}$ points, and RePRI by 0.9 mIoU points compared to the models using no data augmentation. Despite a possible mismatch between OSMIS training resolution and target image size (e.g., 640x384 vs 854x480 for DAVIS) and the need for image resizing, our synthetic data augmentation consistently outperforms standard data augmentation for STM and RePRI, and is almost on par for OSVOS, which was originally tuned for training with standard data augmentation. These results validate the ability of OSMIS to generate structurally diverse data augmentation of sufficient quality in the one-shot regime. Finally, we note that the effect of OSMIS generations is complementary to standard data augmentation, as the best results for all models are observed when the two pipelines are used in combination.

[1] https://github.com/kmaninis/OSVOS-PyTorch

Synthesis method	OSVOS, DAVIS-16 $\mathcal{J}\&\mathcal{F}$	RePRI, $COCO^0$ mIoU
Reference w/o synth. augm.	78.5	31.8
SemanticGAN [15]	73.1	29.4
DatasetGAN [35]	77.8	30.9
Projection [19]	78.4	30.9
Input concat.	79.3	31.9
SemanticGAN D_m [15]	79.5	32.3
MCA (ours)	**79.8**	**32.8**

Table 7. Impact on the performance of synthesized data produced with different models and mask supervision methods. The reference performance is obtained using standard data augmentation. Bold denotes the best performance.

Table 7 demonstrates the efficiency of synthetic data augmentation obtained with different GAN models. The previous image-mask models DatasetGAN and SemanticGAN both show poor applicability in the scenario of one-shot applications due to poor synthesis performance. Further, among the comparison methods for mask synthesis supervision, the strongest increase in performance is achieved with our proposed MCA module. This indicates that the high synthesis diversity and precise image-mask alignment (see Table 3) are the keys to achieve useful data augmentation.

5. Conclusion

We presented OSMIS, an unconditional GAN model that can learn to generate new high-quality image-mask pairs from a single training pair, not relying on any pre-training data. In such a low-data regime, our model generates photorealistic scenes that structurally differ from the original samples, while the produced masks are precisely aligned to the generated image content. Although the synthesis of OSMIS is inherently constrained by the appearance of objects in the original sample, it can serve as a source of useful data augmentation for one-shot segmentation applications, providing complementary gains to standard image augmentation. Thus, we find using one-shot image-mask synthesis in practical applications promising for future research.

Aknowledgement. Juergen Gall was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 -390732324 and the ERC Consolidator Grant FORHUE (101044724).

References

[1] Zhen Liu Irfan Essa Amirreza Shaban, Shray Bansal and Byron Boots. One-shot learning for semantic segmentation. In *British Machine Vision Conference (BMVC)*, 2017.

[2] Rajat Arora and Yong Jae Lee. SinGAN-GIF: Learning a generative video model from a single GIF. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv:1308.3432*, 2013.

[4] Raphael Bensadoun, Shir Gur, Tomer Galanti, and Lior Wolf. Meta internal learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[5] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision (ECCV)*, 2018.

[9] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical Patch VAE-GAN: Generating diverse videos from a single sample. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Xiaoyu He and Zhenyong Fu. Recurrent SinGAN: Towards scale-agnostic single image GANs. In *International Conference on Electronic Information Technology and Computer Engineering*, 2021.

[12] Tobias Hinz, Matthew Fisher, Oliver Wang, and Stefan Wermter. Improved techniques for training single-image GANs. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[13] Tero Karras, Miika Aittala, Janne Hellsten, S. Laine, J. Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] Yijun Li, Richard Zhang, Jingwan Cynthia Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[18] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[19] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations (ICLR)*, 2018.

[20] Vaishnavh Nagarajan, Colin Raffel, and Ian J Goodfellow. Theoretical insights into memorization in gans. In *Advances in Neural Information Processing Systems (NeurIPS) Workshops*, 2018.

[21] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[22] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *International Conference on Computer Vision (ICCV)*, 2019.

[23] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv:2010.11943*, 2021.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[26] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. GANORCON: Are generative models useful for few-shot segmentation? *arXiv:2112.00854*, 2021.

[27] Tamar Rott Shaham, Tali Dekel, and T. Michaeli. SinGAN: Learning a generative model from a single natural image. In *International Conference on Computer Vision (ICCV)*, 2019.

[28] Vadim Sushko, Juergen Gall, and Anna Khoreva. One-Shot GAN: Learning to generate samples from single images and videos. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[29] Vadim Sushko, Dan Zhang, Juergen Gall, and Anna Khoreva. Learning to generate novel scene compositions from single images and videos. *arXiv:2103.13389*, 2021.

[30] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing GANs for one-shot semantic part segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[31] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[32] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *International Conference on Computer Vision (ICCV)*, 2019.

[33] Jianjin Xu and Changxi Zheng. Linear semantics in generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[35] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. DatasetGAN: Efficient labeled data factory with minimal human effort. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

E Smoothness Similarity Regularization for Few-Shot GAN Adaptation

In this appendix, we provide the publication that Chapter 7 of the thesis is based on:

- **Smoothness Similarity Regularization for Few-Shot GAN Adaptation**
  Vadim Sushko, Ruyu Wang, Juergen Gall
  IEEE International Conference on Computer Vision (ICCV), 2023.
  DOI: 10.1109/ICCV51070.2023.00651

Smoothness Similarity Regularization for Few-Shot GAN Adaptation

Vadim Sushko[1,2] Ruyu Wang[1] Juergen Gall[2,3]
[1]Bosch Center for Artificial Intelligence [2]University of Bonn
[3]Lamarr Institute for Machine Learning and Artificial Intelligence
vad221@gmail.com ruyu.wang@de.bosch.com gall@iai.uni-bonn.de

Abstract

The task of few-shot GAN adaptation aims to adapt a pre-trained GAN model to a small dataset with very few training images. While existing methods perform well when the dataset for pre-training is structurally similar to the target dataset, the approaches suffer from training instabilities or memorization issues when the objects in the two domains have a very different structure. To mitigate this limitation, we propose a new smoothness similarity regularization that transfers the inherently learned smoothness of the pre-trained GAN to the few-shot target domain even if the two domains are very different. We evaluate our approach by adapting an unconditional and a class-conditional GAN to diverse few-shot target domains. Our proposed method significantly outperforms prior few-shot GAN adaptation methods in the challenging case of structurally dissimilar source-target domains, while performing on par with the state of the art for similar source-target domains.

1. Introduction

Generative adversarial networks (GANs) have been shown to be powerful at various image synthesis tasks [4, 28, 3, 13, 27, 26]. The success of these models is in large part enabled by the availability of large datasets for training, typically consisting of thousands of images. However, there are many applications and computer vision tasks such as one-shot or few-shot learning [1, 33], out-of-distribution detection [24], or long-tailed recognition tasks [8] where the number of available training images is very low.

Since training a GAN from scratch on very few samples does not perform well as shown in Fig. 1, a common strategy is to fine-tune a pre-trained GAN model on the few-shot dataset, typically employing additional regularization losses to penalize the degradation of the diversity [23, 37]. This approach, referred to as few-shot GAN adaptation, performs well when the target domain is structurally very similar to the dataset that has been used for pre-training, e.g., photographs vs. sketches of human faces. However, the performance drastically degrades in case of large dissimilarities between the source and target domain as shown in Fig. 1.

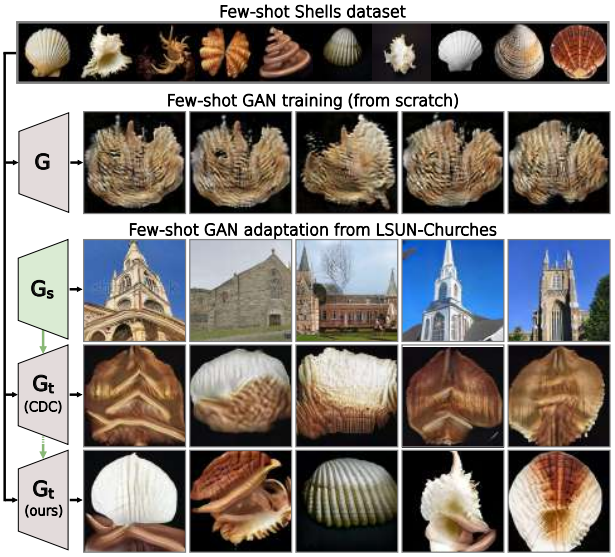

Figure 1. Training a GAN model G on a few-shot dataset (row 1) from scratch fails due to training instabilities (row 2). We thus aim to adapt a GAN G_s that has been pre-trained on a large dataset like LSUN-Church (row 3) to the target few-shot dataset (G_t). While fine-tuning [23] does not perform well either if source and target are dissimilar (row 4), our approach generates diverse and realistic images (row 5) by transferring the smoothness properties of G_s.

Such dissimilarities are a major bottleneck of using GANs in other disciplines like medicine, production, or crop science, where there is a lack of large datasets due to privacy, confidentiality, or simply lack of data. Motivated by this fact, we extend the protocol for few-shot GAN adaptation by investigating also pairs of datasets that are very different like churches and shells as shown in Fig. 1.

To improve few-shot GAN adaptation in the case of structurally dissimilar pairs, we propose a new GAN adaptation strategy. Firstly, we propose a new smoothness similarity regularization for the generator. Our key observation is that pre-trained GAN generators, regardless of the exact structure of objects in the pre-training dataset, learn well-structured and smooth latent spaces. For example, prior works demonstrated that various local shifts in the latent space can lead to interpretable and smooth transitions of

output images, such as translation of objects in the scene or changing their size [34, 9, 30]. As we show in our experiments, the proposed smoothness similarity regularization enables the transfer of this desirable property to other few-shot image domains without compromising the synthesis quality. Secondly, to overcome overfitting issues, we revisit the adversarial loss function of the discriminator and propose a simple yet efficient modification by computing the loss at different layers of the discriminator. This leads to the mitigation of overfitting and a more stabilized adaptation of the model to diverse target domains.

We evaluate our approach by adapting an unconditional [15] and a class-conditional GAN [2] to diverse few-shot target domains. Our model significantly outperforms previous state-of-the-art methods in image quality and diversity in the challenging case of dissimilar source and target domains, while performing on par with the state of the art on structurally similar dataset pairs. In summary, our contributions are as follows: (i) We extend the evaluation protocol for few-shot GAN adaptation by including new dataset pairs that are structurally much less similar than was considered in prior work. (ii) We propose a new smoothness similarity regularization, which enables diverse synthesis in the target domain by transferring the learned smoothness of a pre-trained GAN. (iii) We revisit the adversarial loss function of the discriminator to stabilize few-shot GAN adaptation across diverse target domains. (iv) Our proposed model enables high-quality synthesis in the challenging case of dissimilar source and target domains, significantly outperforming prior methods. In addition, we show that our method can be applied to different classes of GAN architectures, including unconditional and class-conditional GAN models.

2. Related Work

To address the image generation problem in the low data regime, existing works mainly follow three research lines – one-shot, low-shot, and few-shot learning. One-shot generation methods [29, 31] focus on leveraging the internal patch distribution within a single image, however, their extension to capture the distribution of a small collection of images is non-trivial. In low-shot learning [41], several works [41, 12] proposed to mitigate the limited-data-induced overfitting issue by adapting data augmentations to the generative networks. Others [18, 5] stabilized the training process and reduced overfitting by revising the network design. Despite the promising performance in many low data regimes (typically having 100+ images), these low-shot methods fail in the extremely few-shot setting (e.g., 10 images). Our work lies in the scope of few-shot learning.

Few-shot image synthesis. Conventional few-shot learning aims at learning a discriminative classifier under limited data scenarios. In the context of image synthesis with GANs, the goal instead is to produce diverse new images from the learned distribution while preventing overfitting to the few training samples. A straightforward way is to treat it as a domain adaptation problem and incorporate the commonly used transfer learning technique, i.e., fine-tuning, to ease the need for data. However, naive fine-tuning (TGAN) [36] often suffers from overfitting and results in poor performance. Researchers proposed remedies such as mining suitable parts of the latent space before fine-tuning [35] or restricting weight updates, for example, updating only the BatchNorm parameters of the generator [22], penalizing drastic changes in important weights [17], or freezing the earliest layers of the discriminator (FreezeD) [20]. More recent works focused on introducing different regularizations to preserve specific knowledge from the pre-trained model and prevent diversity degradation [42]. For example, CDC [23] proposed to preserve the pair-wise perceptual similarity between samples from the source domain and to transfer it to the target domain, while RSSA [37] designed a novel consistency term to align the structural information between source and target domains. Although the two aforementioned methods constitute the current state of the art in few-shot generative learning, their assumptions impose strong constraints on the structure of the few-shot target domain. As we show in experiments, they fail in the more challenging regime when the source and target domains are not restrictively similar. Most recently, [39] proposed to replace prior knowledge preservation criteria with adaptation-aware kernel modulation (AdAM), which relaxed the source-target proximity requirement of previous methods to some extent. In this work, we take a step further and introduce a new regularization term to preserve the generator's smoothness properties that are not limited to a specific domain, enabling successful adaptation between image domains of unprecedented structural dissimilarity.

Smoothness of image generators. Smooth transitions in the latent space are an important property for generative models, where it is believed to be a sign of a well-conditioned generator. Models trained on large datasets naturally possess this property with or without explicit regularization [2, 15]. For example, StyleGANv2 [15] introduced a regularization based on the perceptual path length measure (PPL) [14], which encourages that a fixed-size step in the latent space results in a fixed-magnitude change in the image space. However, achieving a smooth mapping of the generator is difficult for few-shot image synthesis since there are not enough training samples. Thus, MixDL [16] sought to alleviate the "staircase" latent space interpolations, i.e., jumps between training samples, by introducing a continuous coefficient vector and enforcing smooth interpolations between training images. Although the two above regularizers aim to encourage smoother interpolations between training samples and thus mitigate mode collapse, they are not designed to take advantage of the available pre-training

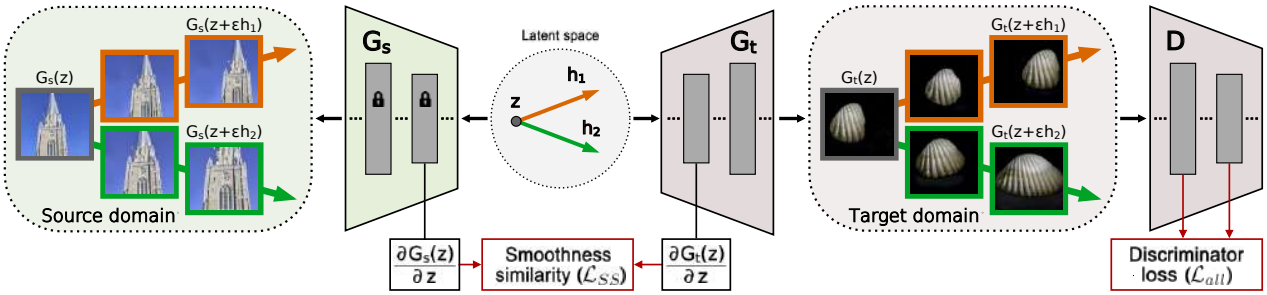

Figure 2. Given a pre-trained generator G_s, the proposed smoothness similarity regularization preserves the learned smoothness of G_s while adapting it to a target domain with very few images. To mitigate overfitting to the target domain, the discriminator loss utilizes features at various layers and automatically adjusts the impact of different semantic scales to the similarity of the source and target domain.

knowledge. In contrast, in this work we develop a new smoothness similarity regularization that leverages the well-structured latent space of a pre-trained GAN generator. In effect, our approach enables high-quality few-shot image synthesis by transferring smooth and realistic image transitions of pre-trained GANs to diverse few-shot domains.

3. Method

In the task of few-shot GAN adaptation, we are given a small target dataset T and a pre-trained GAN model, consisting of a discriminator D and a generator G_s, which produces an image $x = G_s(z)$ from a continuous input variable z, such as a random noise vector or a continuous class embedding. The goal is to adapt the generator to the target dataset such that it generates diverse and realistic images in the domain of the target dataset as shown in Fig. 1. We denote the adapted target generator by G_t.

To achieve few-shot synthesis with a high image quality and diversity, our model should adhere to the following two properties. Firstly, the generator G_t should not only memorize and generate the target images, which will be addressed by the smoothness similarity regularization (Sec. 3.1). Secondly, the discriminator D must avoid overfitting to the few target images in order to provide useful supervision for G_t (Sec. 3.2). The overview of our method is shown in Fig. 2.

3.1. Smoothness similarity regularization for G_t

In a low data regime like ours, G_t can easily overfit to the target dataset T and collapse to reproducing only the few modes represented in the training data. When walking in the latent space of such a generator, one would observe "staircase" patterns, where minor shifts in the latent space cause discontinuous transitions in the output image space (as shown in row 4 of Fig. 5). Naturally, to achieve a synthesis of high diversity, it is desirable for G_t to avoid such discontinuities, as having smoother image transitions allows to generate intermediate samples that can exhibit novel features. Therefore, in our model we aim to encourage G_t to produce smooth latent space interpolations, in which all the intermediate images are realistic.

Our approach is based on the observation that GANs trained on large datasets tend to have a well-structured latent space [34, 9, 30], in which different latent space directions can lead to smooth and interpretable image transitions. For example, in a generator pre-trained on a large dataset of churches, latent directions can emerge causing smooth zooming or translation of churches (see Fig. 2). Our observation is that the nature of such image transitions (e.g., zooming or translation) is remarkably general. Thus, we propose a regularizer that utilizes this smoothness property of the source generator G_s as a cue while adapting it to another image domain, which can be very different from the domain that was used for pre-training. For example, as shown in Fig. 2, the same latent directions of churches can cause similar zooming or translation effects on shells.

Mathematically, the smoothness of the generator can be represented via a Jacobian matrix $J_{G^l}(z) = ||\partial G^l(z)/\partial z||$, quantifying how the generator's intermediate features after the l-th block change under local shifts in the latent space. As we want the same latent shift to cause perceptually similar image transitions in the source and target domains, we design a regularization term that brings the Jacobian matrices of G_s^l and G_t^l closer together. As the computation of full Jacobian matrices is expensive, we use an unbiased estimator of their products with a Gaussian vector [6, 15], which can be computed with standard back-propagation:

$$J^T_{G^l}(z) \cdot y = \mathbb{E}_{(y)\sim N(0,1)} \nabla_z \langle G^l(z), y \rangle, \tag{1}$$

where y is a Gaussian tensor of the same shape as G^l. Our smoothness similarity regularization is then expressed as:

$$\mathcal{L}_{SS} = \lambda_{SS} \cdot \mathbb{E}_{(z,y)\sim N(0,1)} ||\nabla_z \langle G_s^l(z), y \rangle - \nabla_z \langle G_t^l(z), y \rangle||_2, \tag{2}$$

where λ_{SS} steers the impact of the regularizer. As shown in Fig. 2, the smoothness similarity regularization depends on both generators, but only G_t is updated. It is interesting to note that the Jacobian matrix is also used for the path length regularization [15], which forces $J_G(z)$ to be orthogonal up to a global scale at any z. While this alternative regularizer also induces some form of smoothness, it does not transfer the inherently learned smoothness

of a pre-trained GAN. We show in Sec. 4.1 that it struggles to enforce the realism of intermediate images. Furthermore, our approach shares the motivation with some prior regularization approaches that use noise perturbations to enforce diversity [23, 37]. In contrast to Eq. 2, these approaches incorporate non-gradient components, e.g., assuming similarity of images $G_s(z) \leftrightarrow G_t(z)$ or distributions $d(G_t(z_1), G_t(z_2)) \leftrightarrow d(G_s(z_1), G_s(z_2))$. As such assumptions are violated when source and target domains are dissimilar, they perform poorly compared to our smoothness similarity regularization $\mathcal{L}_{SS}$ as shown in the experiments.

3.2. Revisiting the D adversarial loss

To identify what kind of image transitions look realistic for the target domain, G_t requires strong supervision from the discriminator on image realism at different semantic scales. This includes the colors and textures of objects, as well as object shapes, especially if their distribution is different from the shapes of objects in the source domain. Learning the concept of image realism in low data regimes is, however, challenging due to the problem of overfitting.

Typically, a GAN discriminator consists of several consecutive blocks $\{D^i\}_{i=1}^N$ and computes for each given image x a real/fake logit after the last block $l = s^N \circ D^N(x)$, where s^N is a final processing layer such as a convolution. When adapting such a discriminator to a very small dataset, it is prone to memorizing the training set [32], leading to mode collapse and poor diversity of synthesized images [23]. A possible solution [23, 37] to overcome memorization is to use variants of the PatchGAN discriminator [11], discarding the latest discriminator layers: $l = s^k \circ D^k(x), k < N$. This solution allows to adapt colors and textures of generated images to the target domain while avoiding the memorization problem. However, it naturally has a limited capacity to learn more high-level semantic scene properties such as the shapes of objects, which we show in the experiments.

In order to avoid memorization, and yet to balance the adaptation of colors, textures, and shapes of generated objects to a new domain, we hypothesize that a more flexible attention to different levels of image realism is required by the discriminator. To this end, we perform a simple yet efficient modification to the loss function of the discriminator. Given a discriminator $\{D^i\}_{i=1}^N$ and its adversarial loss function $\mathcal{L}_{\mathcal{D}}(l)$ used for pre-training (e.g., cross-entropy or hinge loss), we design the discriminator to produce real/fake logits after *each* discriminator's block, and correspondingly compute the loss as the average across all blocks:

$$\mathcal{L}_{all}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathcal{D}}[l^i(x)], \quad l^i(x) = s^i \circ D^i(x). \quad (3)$$

With the new objective, D is given more freedom to utilize the features extracted at different scales to compute the loss. Our finding is that D dynamically adapts the magnitude of the loss at each scale to the target domain, without explicit supervision (see Fig. 6). Consequently, we observe a strong overall stabilization effect on the adaptation performance across diverse source-target dataset pairs.

4. Experiments

To demonstrate that our approach for few-shot GAN adaptation can be applied to unconditional and class-conditional GANs, we selected for each category a popular GAN architecture: unconditional StyleGANv2 [15] and class-conditional BigGAN [2]. For both models, we test our approach on a variety of source-target domain pairs. We focus on 10-shot target adaptation in the main paper, but we provide results for 1-shot and 5-shot adaptation in the supplementary material. For fair comparisons with prior works, most of our ablations and comparisons are conducted with StyleGANv2.

4.1. Adaptation of unconditional GAN

Datasets. In contrast to previous works that mostly considered pairs of similar datasets like *Face→Sketch* and *Face→Sunglasses*, we extend the protocol by including structurally dissimilar pairs of source and target domains, which is a more challenging task and is our primary interest. As source generators, we use StyleGANv2 checkpoints pretrained on FFHQ [14], LSUN-Church, and LSUN-Horse [38]. For the target datasets, we selected 10-shot subsets of various commonly used few-shot datasets, such as Anime-Face, Shells, or Pokemons [41, 18]. Results on more datasets are shown in the supplementary material.

Training details. We fine-tune StyleGANv2 using the $\mathcal{L}_{SS}$ and $\mathcal{L}_{all}$ loss terms as presented in Sec. 3. For the smoothness similarity regularization, we use the intermediate features G^l at resolution (32×32) and set $\lambda_{SS} = 5.0$. We follow [23] in choosing all the other hyperparameters, such as image resolution (256×256), learning rates, and batch size. Our experiments across all datasets use the same model configuration and set of hyperparameters.

Baselines. We compare our method to most recent few-shot GAN adaptation approaches: TGAN [36], FreezeD [20], CDC [23], RSSA [37], and AdAM [39]. In addition, we compare our proposed smoothness similarity regularizer $\mathcal{L}_{SS}$ to other regularization techniques: path length regularization (PPL) [15] and MixDL [16].

Evaluation. In low data regimes, it is necessary to judge results both in quality and diversity aspects, as there is a trade-off between them [25, 32]. We measure the quality with FID [10] between a held-out validation set and a generated set of the same size. Following [23], we evaluate diversity with the intra-LPIPS, clustering generated images according to their nearest training samples and computing

Figure 3. Visual comparison to prior methods on *Face→Anime* and *Church→Shells*, the source-target dataset pairs with a dissimilar structure (e.g., shapes of objects). In this challenging regime, we observe that prior methods suffer from training instabilities, memorization issues, or inability to adapt the shapes of objects to the new domain. In contrast, our method generates images that look realistic, flexibly combine features of different target images, and transfer the variation of images from the source domain to the target domain.

the average LPIPS [40] of all the clusters. We train all models for 30k epochs in case of dissimilar domain pairs and for 5k on closer domains, evaluating metrics every 1k epochs. Final checkpoints in all experiments correspond to best FID.

Results with dissimilar source-target domains. We first present our results on the source-target domain pairs with dissimilar structure: *Face→Anime*, *Church→Shells*, and *Horse→Pokemon* (see Fig. 3 and supplementary material). Our general observation from Fig. 3 is that in this challenging regime prior methods suffer either from training instabilities, memorization issues, or inability to adapt the shape of objects to the new domain. For example, for *Face→Anime*, despite an apparent correspondence between the two domains, none of the prior methods success-

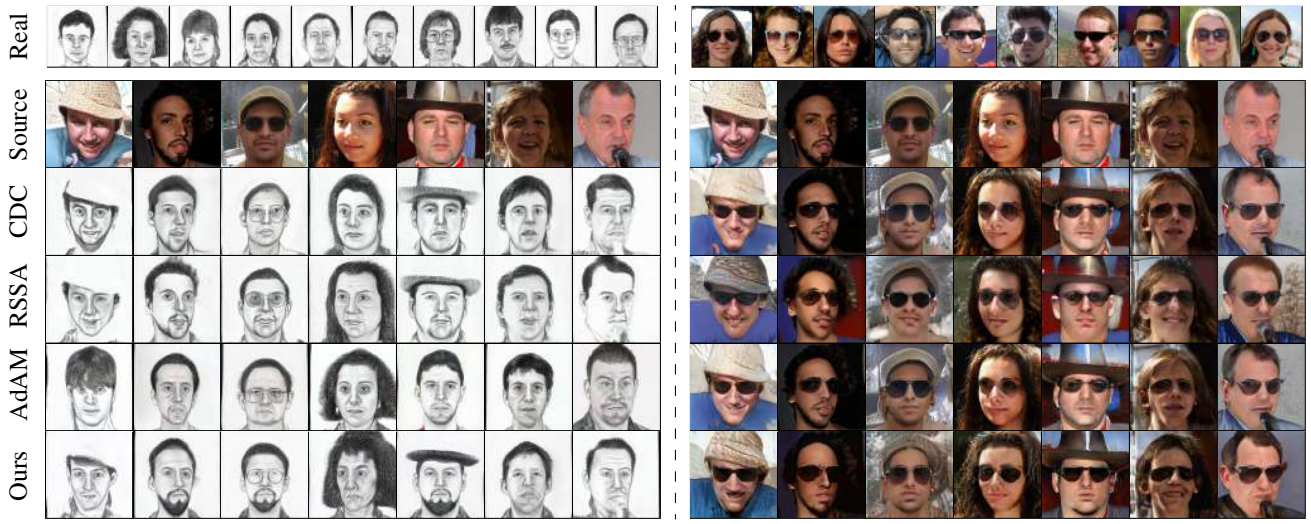

Figure 4. Visual comparison to most recent prior methods on *Face→Sketch* and *Church→Sunglasses*, the dataset pairs depicting similar image domains. In this regime, our method performs on par with previous state of the art. See Table 2 for a quantitative comparison.

Method	**Face→Anime**		**Church→Shells**		**Horse→Pokemons**	
	FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑
TGAN [36]	153.2	0.29	205.3	0.22	115.0	0.52
FreezeD [20]	112.4	0.22	180.8	0.27	123.3	0.49
CDC [23]	140.2	0.50	187.9	0.48	109.5	0.55
RSSA [37]	133.2	0.37	182.4	0.44	117.3	0.54
AdAM [39]	116.4	0.42	152.4	0.28	106.5	0.55
Ours	**97.3**	**0.57**	**140.5**	**0.53**	**84.1**	**0.61**

Table 1. Comparison of the adaptation performance in case of dissimilar source-target domains. Bold denotes best performance.

Method	**Face→Sketch**		**Face→Sunglasses**	
	FID↓	LPIPS↑	FID↓	LPIPS↑
TGAN [36]	54.2	0.38	36.8	0.56
FreezeD [20]	48.8	0.32	32.0	0.59
CDC [23]	54.2	0.40	30.5	0.59
RSSA [37]	61.4	**0.45**	36.3	0.58
AdAM [39]	56.3	0.37	31.1	**0.60**
Ours	**45.2**	0.44	**27.5**	**0.60**

Table 2. Comparison in case of structurally close source-target domains. Bold denotes best performance.

fully transfers the distribution of head poses to the anime style, e.g., overfitting too strongly to the 10 provided samples (FreezeD), failing to adapt the shape of faces to the style of anime (CDC), or not generating high-quality anime-faces due to instabilities (TGAN, RSSA, AdAM). Similarly, for *Church→Shells*, we observe that prior methods produce only copies of the example shells (FreezeD, AdAM), generate shells of unrealistic church-like shapes (CDC, RSSA), or suffer from instabilities (TGAN). In contrast, our method achieves high-quality synthesis, in which the generated images (i) look like realistic anime-faces and shells; (ii) flexibly combine features observed in different target images (e.g., anime hair color can be combined with various eye colors or background styles); and (iii) meaningfully transfer the variation of images from the source domain (e.g., generated shells adjust to the positions and shapes of churches).

The quantitative comparison in Table 1 confirms our analysis, where our method achieves the best quality and diversity scores across all datasets. We note a high average relative improvement of more than 18% and 11% in FID and LPIPS compared to the highest scores achieved by prior methods. Overall, we conclude that our method significantly improves over prior works on few-shot GAN adaptation with dissimilar source and target domains.

Results with close source-target domains. Next, we follow the evaluation of prior works and compare the models on similar source and target domains, such as adaptation of human faces to a different style. The visual results for *Face→Sketch* and *Face→Sunglasses* are shown in Fig. 4. Our method successfully performs the few-shot adaptation in this setting, adapting the colors and textures of faces to the gray-scale sketch domain, or adding a novel attribute (sunglasses). We note that our method is not explicitly designed to transfer all the details of a face from the source domain, thus changes in the generated images like facial hair are expected. Yet, we observe that our method generally does not lose distinctive features of faces in source images, performing on par with previous state-of-the-art methods. The quantitative comparison is provided in Table 2[1]: on both datasets our method achieves the best FID scores and performs on par with the best performer in LPIPS.

Ablations. We demonstrate the importance of our proposed loss terms in Fig. 5, which shows latent space interpolations of trained models and their similarity to the pre-

[1] FID evaluation differs from prior works (discussed in suppl. material).

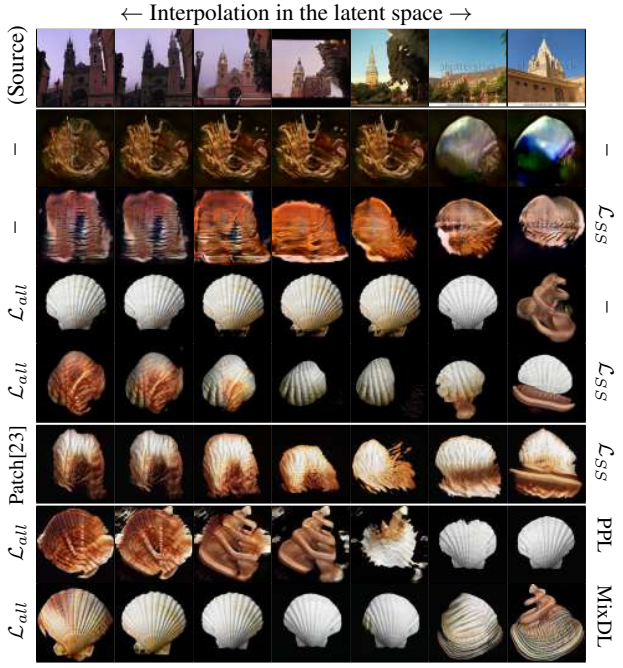

Figure 5. Latent space interpolations of the source generator and the ablation models from Tables 3-4. Leftmost and rightmost columns show the used D loss and G smoothness regularization.

D loss	Smooth reg. for G	**Face→Anime** FID↓	LPIPS↑	**Church→Shells** FID↓	LPIPS↑
StyleGANv2	-	178.0	0.21	243.8	0.17
StyleGANv2	SS (ours)	180.7	**0.61**	252.8	**0.62**
PatchGAN [23]	-	145.2	0.37	183.1	0.31
PatchGAN [23]	SS (ours)	132.2	0.55	184.2	0.56
$\mathcal{L}_{all}$ (ours)	-	116.4	0.36	175.4	0.43
$\mathcal{L}_{all}$ (ours)	SS (ours)	**97.3**	0.57	**140.5**	0.53

Table 3. Impact of $\mathcal{L}_{all}$ and $\mathcal{L}_{SS}$. Bold denotes best performance.

trained source model G_s (row 1). Firstly, we note that the plain StyleGANv2 model (row 2) suffers from instabilities in our low data regime, achieving poor image quality and diversity and having "staircase"-like latent space interpolations. Applying $\mathcal{L}_{SS}$ without $\mathcal{L}_{all}$ (row 3) helps to achieve diverse synthesis with smooth interpolations, but it is not enough to achieve good image quality. On the other hand, using $\mathcal{L}_{all}$ (row 4) helps to overcome instabilities and improve image quality, but it cannot maintain smooth interpolations and high diversity. Finally, our full model (row 5) allows a higher-quality, diverse synthesis with smooth and realistic latent space interpolations. Note how the image transitions mimic the behaviour of the source model (churches and shells change shapes and positions similarly), allowing to achieve diverse and realistic synthesis.

The effect of $\mathcal{L}_{all}$ is further demonstrated in Fig. 6, where we show the contribution of different D blocks to the adversarial loss at different epochs. We note the ability of the discriminator to identify correct loss contributions

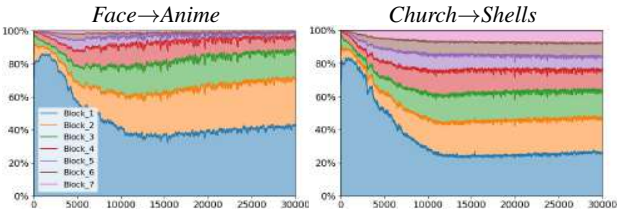

Figure 6. The contribution of features at different D blocks to the adversarial loss function $\mathcal{L}_{all}$. For two closer image domains (the left plot), the network concentrates mostly on earlier layers to compute the loss, while for less similar domains the network learns to use the later layers representing more high-level D features.

D loss	Smooth reg. for G	**Face→Anime** FID↓	LPIPS↑	**Church→Shells** FID↓	LPIPS↑
$\mathcal{L}_{all}$ (ours)	-	116.4	0.36	175.4	0.43
$\mathcal{L}_{all}$ (ours)	PPL [14]	107.8	0.46	179.4	0.44
$\mathcal{L}_{all}$ (ours)	MixDL [16]	105.9	0.50	150.4	0.51
$\mathcal{L}_{all}$ (ours)	SS (ours)	**97.3**	**0.57**	**140.5**	**0.53**

Table 4. Comparison of smoothness similarity regularization $\mathcal{L}_{SS}$ with other regularizers. Bold denotes best performance.

adaptively for different source-target domain pairs. For example for *Face→Anime*, the network concentrates mostly on the earliest D blocks to adapt the colors and textures of faces to a new style. In contrast, for the more distant domains *Church→Shells*, the network learns to attribute a higher weight to the later blocks to also adapt higher-level features, such as shapes of objects. In effect, we observe a stabilized adaptation of colors, textures, and shapes of objects across diverse source-target pairs. Using PatchGAN [23] as discriminator loss does not achieve such a balance as it focuses mostly on lower-scale features (row 6 in Fig. 5).

Our observations are confirmed by the quantitative study in Table 3: without $\mathcal{L}_{SS}$ the model does not achieve high diversity (high LPIPS), while $\mathcal{L}_{all}$ is necessary for high image quality (low FID). We conclude that both our proposed loss terms are important to achieve high-quality synthesis. More ablations on $\mathcal{L}_{SS}$ and $\mathcal{L}_{all}$ can be found in the supplementary material.

Lastly, Table 4 provides a comparison of our proposed $\mathcal{L}_{SS}$ loss term to other regularizers: path length regularization (PPL) [14] and MixDL [16]. While all regularizers help to achieve smoother latent space interpolations and thus improve the quality and diversity metrics, our smoothness similarity regularization enables the highest performance in both FID and LPIPS. While our approach transfers the learned smoothness of the source generator to the target domain, PPL and MixDL resort to gradually interpolating between the provided training samples, which leads to latent space interpolations that either look unrealistic or lack diversity (rows 7-8 in Fig. 5). This demonstrates that transferring smoothness from a pre-trained generator is beneficial to enforce image transitions that are realistic and diverse.

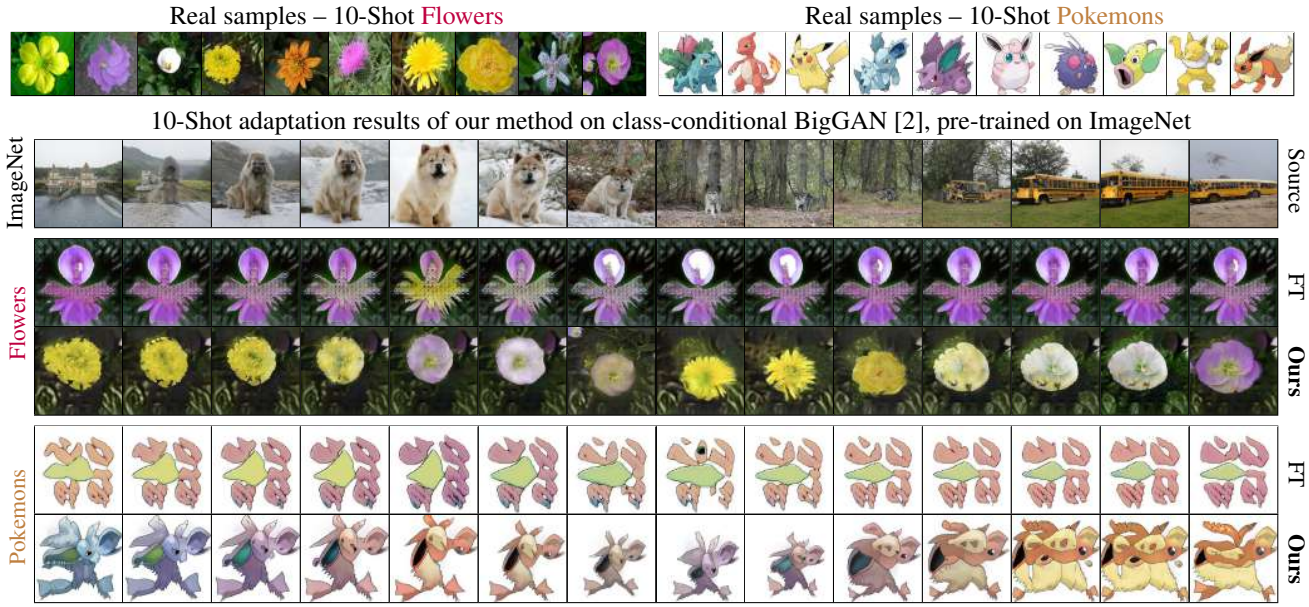

Figure 7. 10-shot adaptation results for the class-conditional BigGAN [2] pre-trained on ImageNet. While simple fine-tuning (FT) suffers from training instabilities and mode collapse, our method helps to achieve much higher image quality and diversity, transferring smooth and realistic image transitions from the source domain, e.g., objects smoothly changing their locations, size, and shape.

D loss	Smooth reg. for G	**ImageNet→Flowers** FID↓	LPIPS↑	**ImageNet→Pokemons** FID↓	LPIPS↑
BigGAN	-	213.3	0.29	226.8	0.15
BigGAN	SS (ours)	225.6	0.47	208.3	0.47
$\mathcal{L}_{all}$ (ours)	-	123.9	0.28	129.4	0.27
$\mathcal{L}_{all}$ (ours)	SS (ours)	**106.4**	**0.55**	**89.6**	**0.56**

Table 5. Ablation on the performance when adapting the class-conditional BigGAN model [2] pre-trained on ImageNet.

4.2. Adaptation of class-conditional GAN

Our approach is not limited to unconditional GANs, but it can also be applied to a class-conditional GAN model. We selected BigGAN [2] for our experiments as it is a popular backbone architecture for class-conditional image synthesis on ImageNet [7]. We make two modifications to enable the adaptation of the model to unconditional few-shot datasets. Firstly, we remove the conditioning of the discriminator via the projection layer [19]. Secondly, we treat the generator's learned continuous class embedding as part of the latent space, thus sampling a Gaussian vector in the joint noise-class space at each fine-tuning epoch. This way, the generator produces an image based on a single input vector in an unconditional fashion. We then fine-tune the pre-trained model using our loss terms $\mathcal{L}_{SS}$ and $\mathcal{L}_{all}$ as presented in Sec. 3. We use image resolution 256×256 and batch size of 32. The hyperparameters for $\mathcal{L}_{SS}$ are the same as for StyleGANv2: intermediate features G^l at resolution (32×32) and $\lambda_{SS} = 5.0$. We train for 30k epochs and select checkpoints by best FID.

Datasets. As the source generator, we use the BigGAN checkpoint pre-trained on class-conditional ImageNet at resolution 256×256. We demonstrate 10-shot adaptation results with two commonly used few-shot generation datasets: Oxford-Flowers [21] and Pokemons [18]. We use the same model configuration for both datasets.

Results. Fig. 7 demonstrates latent space interpolations of the source and target generators. We note that a simple fine-tuning of BigGAN suffers from training instabilities and mode collapse. In contrast, our method successfully adapts BigGAN to generate diverse images in the target domains. We highlight that our method transfers smooth and realistic image transitions from the well-learned BigGAN's noise-class space, despite significant dissimilarities between ImageNet and the few-shot datasets, in particular Pokemons. For example, it can be noticed how the latent space interpolations in the target domains mimic the source domain, e.g., the generated flowers and pokemons change their position and size similarly to dogs and wolves (5^{th}-10^{th} columns in Fig. 7) or stretch their shape to mimic the proportions of busses (11^{th}-14^{th} columns).

Table 5 shows the importance of our proposed loss terms. Our observations are consistent with the ablations with StyleGANv2: $\mathcal{L}_{all}$ is necessary to avoid instabilities and achieve a good image quality (low FID), while $\mathcal{L}_{SS}$ is required to achieve smooth latent space interpolations and good diversity (high LPIPS). We conclude that our method successfully extends to the adaptation of class-conditional models, where target domains benefit from the rich noise-class space learned on a multi-class dataset such as ImageNet. More details and results are provided in the supplementary material.

5. Conclusion

In this work, we presented a new method for few-shot adaptation of GAN models. It transfers the smooth latent space of a pre-trained GAN, which was trained on a large dataset, to a new domain with very few images. We addressed the case of few-shot GAN adaptation when the source and target domains are structurally dissimilar, which is a common issue in applications. Our extensive results demonstrate that in this setting our approach outperforms previous works in terms of image quality and diversity.

Acknowledgement

The work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 -390732324 and the ERC Consolidator Grant FORHUE (101044724). We thank Jinhui Yi for providing and analyzing the sugar beet data used in supplementary experiments.

References

[1] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[5] Kaiwen Cui, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Fangneng Zhan, and Shijian Lu. Genco: Generative co-training for generative adversarial networks with limited data. In *Conference on Artificial Intelligence (AAAI)*, 2021.

[6] Yann Dauphin, Harm De Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Tero Karras, Miika Aittala, Janne Hellsten, S. Laine, J. Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] Chaerin Kong, Jeesoo Kim, Donghoon Han, and Nojun Kwak. Few-shot image generation with mixup-based distance learning. In *European Conference on Computer Vision (ECCV)*, 2022.

[17] Yijun Li, Richard Zhang, Jingwan Cynthia Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[18] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[19] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations (ICLR)*, 2018.

[20] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. In *CVPR AI for Content Creation Workshop*, 2020.

[21] Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[22] Atsuhiro Noguchi and T. Harada. Image generation from small datasets via batch statistics adaptation. In *International Conference on Computer Vision (ICCV)*, 2019.

[23] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[24] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[25] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv:2010.11943*, 2021.

[26] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023.

[27] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH*, 2022.

[28] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[29] Tamar Rott Shaham, Tali Dekel, and T. Michaeli. SinGAN: Learning a generative model from a single natural image. In *International Conference on Computer Vision (ICCV)*, 2019.

[30] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[31] Vadim Sushko, Juergen Gall, and Anna Khoreva. One-Shot GAN: Learning to generate samples from single images and videos. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[32] Vadim Sushko, Dan Zhang, Juergen Gall, and Anna Khoreva. Learning to generate novel scene compositions from single images and videos. *arXiv:2103.13389*, 2021.

[33] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[34] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning (ICML)*, 2020.

[35] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, L. Herranz, F. Khan, and Joost van de Weijer. Minegan: Effective knowledge transfer from gans to target domains with few images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] Yaxing Wang, Chenshen Wu, L. Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and B. Raducanu. Transferring gans: generating images from limited data. In *European Conference on Computer Vision (ECCV)*, 2018.

[37] Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[38] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365*, 2015.

[39] Zhao Yunqing, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-man Cheung. Few-shot image generation via adaptation-aware kernel modulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[42] Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.