UNIVERSITÄT BONN

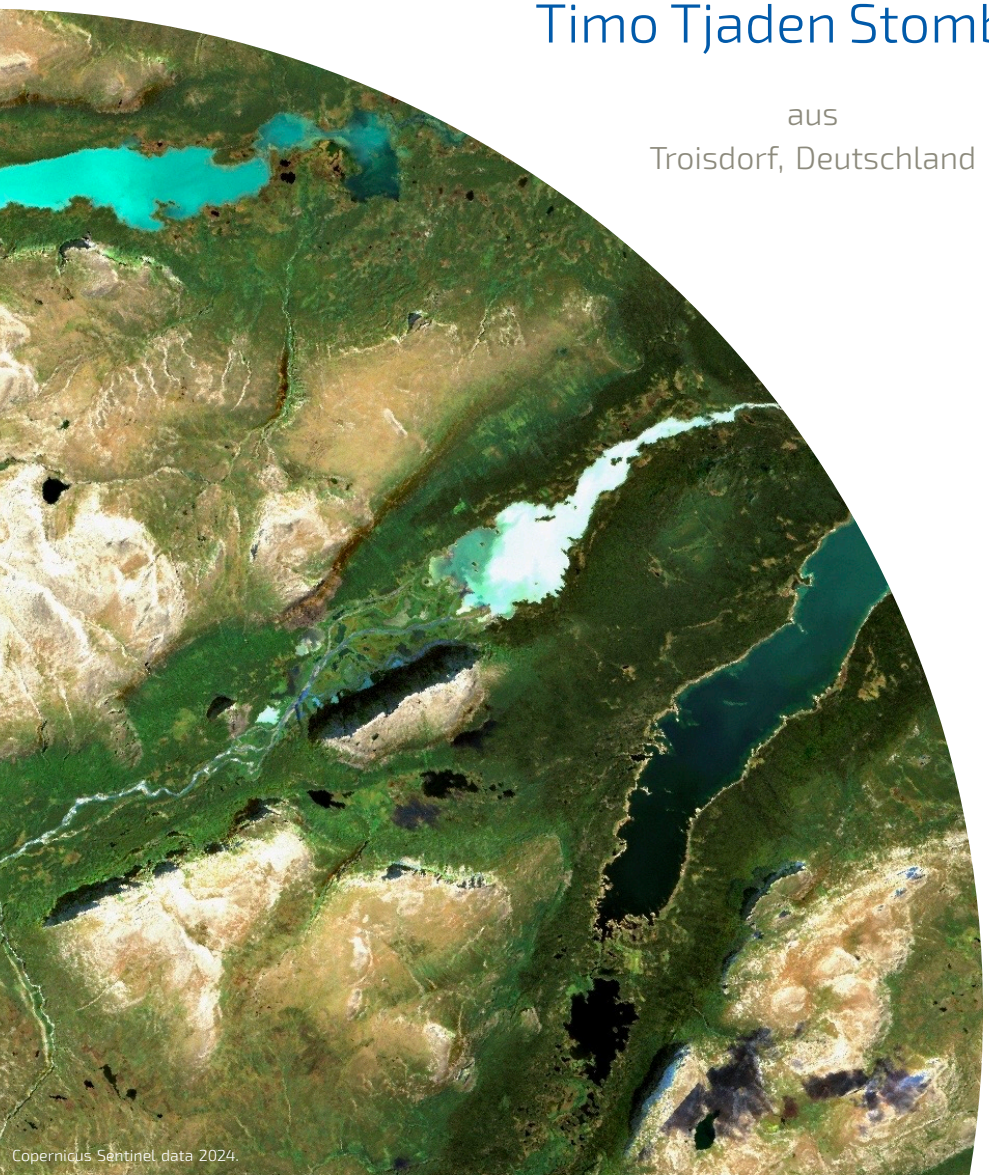
Dissertation
zur Erlangung des Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)
Agrar-, Ernährungs- und Ingenieurwissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn
Institut für Geodäsie und Geoinformation

Improving Explanations of Convolutional Neural Networks with Applications to Land Cover Mapping

von

Timo Tjaden Stomberg

aus
Troisdorf, Deutschland



Copernicus Sentinel data 2024.

Bonn 2026

Referentin:
Prof. Dr.-Ing. Ribana Roscher,
Rheinische Friedrich-Wilhelms-Universität Bonn, Deutschland

1. Korreferent:
Prof. Dr. Martin Schultz,
Universität zu Köln, Deutschland

2. Korreferent:
Prof. Dr. rer. nat. Michael Mommert,
Hochschule für Technik Stuttgart, Deutschland

Tag der mündlichen Prüfung: 27. März 2026

Angefertigt mit Genehmigung der Agrar-, Ernährungs-, und Ingenieurwissenschaftlichen Fakultät der Universität Bonn

Abstract

Convolutional neural networks (CNNs) have revolutionized computer vision and remain a key technology in many satellite imagery applications for environmental monitoring. As these models are integrated into scientific workflows and operational monitoring, questions about their interpretability arise; however, explaining how they generate their predictions remains challenging. Attribution methods like Grad-CAM and occlusion sensitivity are widely used to explain CNN predictions, yet they often yield differing explanations. These inconsistencies make it hard to assess reliable explanations and undermine overall trust in machine learning models.

This thesis addresses these challenges by investigating how explanations of CNN-based models can be made more interpretable, consistent, and reliable for remote sensing applications. First, we introduce UH-Net, an interpretable-by-design architecture that incorporates a high-resolution deep layer to combine semantic richness with spatial detail in attribution maps. Second, we conduct a systematic comparison of attribution methods across different CNN architectures and layers to better understand their behavior, strengths and limitations. Building on these insights, we propose a harmonization method that significantly reduces differences in attribution results across methods and provides more comprehensible explanations. Furthermore, we present two feature-specific attribution methods that achieve an inherent degree of harmonization by design. Finally, we apply our methods to naturalness mapping, making us among the first to do so using satellite imagery. To this end, we develop a high-quality Sentinel-2 dataset covering both protected and anthropogenic regions in Fennoscandia. Using UH-Net and harmonized attribution maps, we generate and evaluate large-scale naturalness maps and temporal changes across Fennoscandia from 2018 to 2024.

Overall, this work contributes new insights, methods, datasets, and applications for explainable machine learning in remote sensing. By improving the interpretability and consistency of CNN explanations, it advances the responsible and transparent application of machine learning in environmental science.

Zusammenfassung

Convolutional Neural Networks (CNNs) haben die Bildverarbeitung nachhaltig geprägt und sind heute ein zentraler Bestandteil zahlreicher Anwendungen der satellitenbildbasierten Umweltbeobachtung. Mit ihrer Nutzung in Forschung und operativen Anwendungen steigt jedoch auch der Bedarf an Nachvollziehbarkeit und Transparenz; denn wie CNNs zu Entscheidungen gelangen, ist häufig schwer verständlich. Zwar existieren etablierte Attributionsmethoden wie Grad-CAM und Occlusion Sensitivity, doch liefern diese oftmals voneinander abweichende Erklärungen. Diese Inkonsistenzen erschweren es, verlässliche Erklärungen zu finden, was das Vertrauen in Machine-Learning-Modelle insgesamt mindert.

Die vorliegende Dissertation untersucht daher, wie sich die Erklärbarkeit CNN-basierter Modelle verständlicher, konsistenter und verlässlicher gestalten lässt. Hierfür stellen wir zunächst das UH-Net vor — eine Architektur, die durch eine tiefe, hochauflösende Repräsentationsebene detailliertere und semantisch reiche Attributionskarten erzeugt. Anschließend vergleichen wir gängige Attributionsmethoden systematisch über verschiedene CNN-Architekturen und -Ebenen hinweg, um ihr Verhalten, ihre Stärken und ihre Einschränkungen präziser einschätzen zu können. Auf Basis dessen präsentieren wir eine Harmonisierungsmethode, die Unterschiede zwischen den Verfahren deutlich reduziert und plausiblere Erklärungen ermöglicht. Ergänzend hierzu entwickeln wir zwei neue, merkmalsorientierte Attributionsverfahren, die bereits konzeptionell auf harmonisierte Attributionen abzielen. Zuletzt wenden wir unsere Methoden zur Kartierung von Natürlichkeit an und gehören damit zu den Ersten, die dies anhand von Satellitenbildern tun. Dazu erstellen wir einen Datensatz, der Sentinel-2-Bilder aus naturgeschützten sowie anthropogen geprägten Landschaften in Fennoskandien enthält. Unter Verwendung des UH-Net und harmonisierter Attributionskarten leiten wir daraus flächendeckende Natürlichkeitskarten ab und analysieren Veränderungen im Zeitraum 2018 bis 2024.

Insgesamt liefert diese Arbeit neue Erkenntnisse, Methoden, Datensätze und Anwendungen für erklärbares maschinelles Lernen in der Fernerkundung. Durch die Verbesserung der Interpretierbarkeit und Konsistenz von CNN-Erklärungen trägt sie zu einem verantwortungsvollen und transparenten Einsatz von Maschinellem Lernen in Umweltwissenschaften bei.

Contents

Part I

Introduction and Background

Chapter 1

Introduction

1.1 Motivation

Machine learning models are widely and successfully applied to a range of tasks, including classification, object detection, and regression. They are suitable for processing large amounts of data and can find patterns and relations that may not be recognizable by humans. Convolutional neural networks (CNNs), in particular, are among the key technologies in computer vision. They apply filter operations with parameters learnt from data, thereby effectively extracting complex patterns from images. These operations are organized into structures known as layers. In remote sensing, CNNs enable large-scale land cover mapping and environmental monitoring from satellite data. Unlike conventional computer vision tasks that concentrate on object-centric images, satellite-based land cover classification involves continuous, amorphous regions characterized by fine-grained textures. Multispectral satellite images exhibit high spectral dimensionality, with multiple bands beyond the visible spectrum providing rich but complex feature information.

While CNNs are highly effective at processing the complex, high-dimensional data found in satellite imagery, their decision-making processes are often not transparent. This raises concerns, especially in applications where understanding model predictions is essential for trust and responsible use. Land cover classification comprises a wide range of environmental and societal applications, including agriculture, resource management, change detection, disaster management, and nature conservation, to name a few. However, when employed in socioeconomic and political decisions, a high level of transparency is essential, as these actions can have far-reaching impacts on society and ecosystems. Explainable machine learning aims to address this challenge. Attribution methods represent a subfield that evaluates the importance of input features based on their influence on the model's predictions. This often involves computing attribution

maps, which can be directly compared to the input image. Examples include Gradients×Features (Simonyan et al., 2014) and Grad-CAM (Selvaraju et al., 2020), which are gradient-based methods, as well as occlusion sensitivity analysis (Zeiler and Fergus, 2014), which relies on input perturbations.

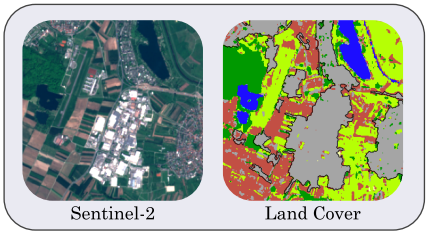
Despite their usefulness, a significant challenge with current attribution methods is their inconsistency. Different techniques often yield differing explanations for the same prediction. Discrepancies like these undermine trust in explanations and raise questions about which attribution method is best suited for a given task. However, there is currently no consensus on how to obtain consistent and reliable explanations across methods. Addressing this gap is essential for ensuring transparency in scientific and operational contexts. Also, explainable machine learning methods, if used at all, are usually applied post hoc to models designed to achieve high accuracies. However, since the decision-making process of deep learning models can differ significantly from human reasoning, the resulting explanations can be difficult to interpret. Nonetheless, specific model architectures can be designed to yield more interpretable predictions. Interpretability may become important when assessing whether a model's performance will hold in new environments or under changing conditions. Without insight into the model's reasoning, it is difficult to evaluate its reliability or to build trust in its predictions.

1.2 Scientific Contributions

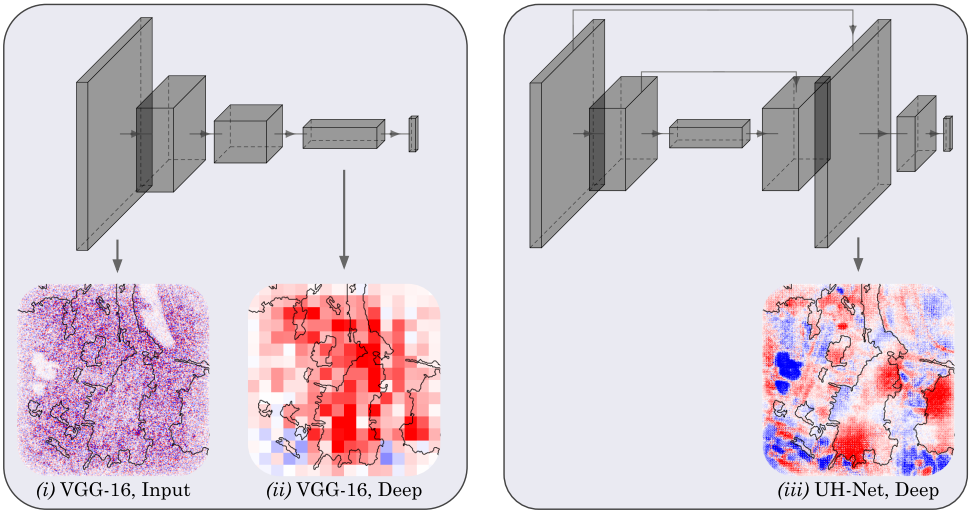
To address the challenges above, this thesis investigates how explanations of CNN-based land cover classification can be made more consistent and interpretable for environmental applications. Chapter II is method-oriented and explores novel methodologies with experiments conducted on well-established benchmark datasets for land cover and object classification. Chapter III is application-oriented and presents a novel approach to naturalness mapping that integrates these techniques, making us among the first to assess naturalness using multispectral satellite data. The following summarizes our primary scientific contributions.

UH-Net Architecture

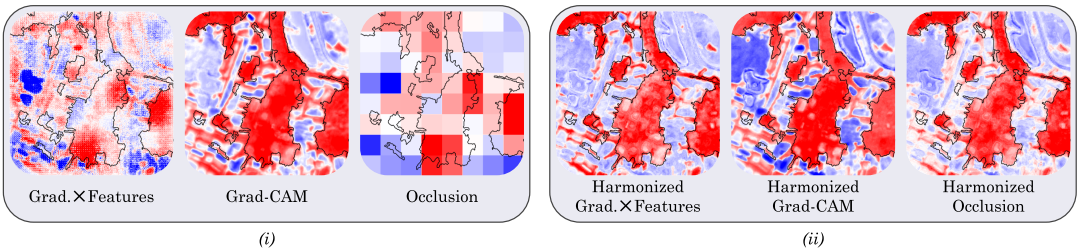
For common CNN architectures, attribution maps provide high resolution at the input level but are more informative at deeper layers (see Figure 1.1b, left-hand side). This creates a dilemma regarding which layer should be used for attribution computations. Our innovative architecture, UH-Net, addresses this issue by incorporating a deep layer that captures high-level features in high-resolution. As a result, these maps become more informative and thus easier to interpret as illustrated in Figure 1.1b, right-hand side.



(a) Sentinel-2 sample with corresponding land cover map: Forest, Grassland, Cropland, Water, and Urban/Built-up.



(b) Common CNN architectures, such as shown on the left-hand side, are designed so that the number of channels increases while the spatial resolution decreases with increasing depth. Depending on the chosen layer, the resulting attribution maps can differ significantly: *(i)* input-level analysis offers the highest resolution, whereas *(ii)* deep-layer analysis yields more meaningful results by capturing high-level semantics. *(iii)* To combine both advantages, we propose UH-Net which preserves high resolution at a specific deep layer. The attribution maps shown are computed using Gradients×Features. The illustrations of the architectures are simplified and created using PlotNeuralNet by Iqbal (2018).



(c) *(i)* Depending on the chosen method, the resulting attribution maps can differ significantly. *(ii)* Harmonization increases consistency across different attribution methods and results in more comprehensive explanations. The attribution maps shown stem from UH-Net.

Figure 1.1: Attribution maps from (b) various layers and (c) attribution methods. They are derived from image-wise, multi-label land cover classification using the DFC2020 dataset by Schmitt et al. (2019b), and computed for the Urban / Built-up class of the sample shown in (a). The attribution values range from negative to positive.

Comparing Attribution Methods

Attribution methods are compared in previous works; however, these comparisons often lack validity because different methods are applied to different layers. This inconsistency arises from the fact that some attribution methods were originally designed for input-level analysis (e.g., occlusion sensitivity), while others are intended for deeper layers (e.g., Grad-CAM). Nevertheless, most attribution methods can, in principle, be applied to any layer, and their results can vary significantly depending on the chosen layer. In this thesis, we compare several attribution methods applied to both input and deep layers across multiple CNN architectures.

Harmonizing Attributions

We introduce a novel methodology that harmonizes attribution results, thereby mitigating the issue that different attribution methods often produce differing explanations. To achieve this, we collect feature attributions across the entire training dataset within the model's learned feature space. Our approach is simple yet effective. It reduces noise in gradient-based attribution maps, enhances the resolution of occlusion-based ones, and adjusts misleading explanations. Examples are shown in Figure 1.1c. Overall, it provides the following key contributions:

- Greater consistency across different attribution methods, making the choice of attribution method less critical. Across all tested models, land cover datasets, and attribution methods, the Pearson correlation between attribution methods increases from 0.45 to 0.64.
- More comprehensible explanations, evaluated based on how well the predominant attribution class aligns with the segmentation ground truth. Across all tested models, land cover datasets, and attribution methods, we observe an increase in alignment from 50 % to 61 %.
- Enhanced transparency and traceability through a mechanistic and intuitive analysis of the feature space.

Our harmonization approach is constrained by the degree of entanglement within the learned feature space and therefore particularly suitable for deeper layers.

Feature-specific Attribution Methods

Building on the insights from our harmonization framework, we develop two feature-specific attribution methods that exhibit an inherent degree of consistency by design. Specifically, we adapt the widely used techniques Grad-CAM and occlusion sensitivity, and achieve state-of-the-art performances in our experiments. Both methods demonstrate particular suitability for our UH-Net architecture.

AnthroProtect Dataset

Landscapes with a high degree of naturalness offer important ecological and social benefits, and monitoring these areas can effectively support policy-making and land use planning. However, mapping naturalness based on satellite imagery presents significant challenges due to the difficulty of accurate labeling. We therefore introduce the AnthroProtect dataset, comprising high-quality Sentinel-2 composites of protected and anthropogenic regions in Fennoscandia, defining a relevant classification task.

Naturalness Mapping

We further present a novel approach to mapping naturalness using satellite imagery by training our UH-Net architecture on the AnthroProtect dataset and computing high-resolution attribution maps. Harmonizing the attributions ensures consistency across scenes, enabling large-scale analysis. In doing so, we generate a naturalness map for Fennoscandia and detect anthropogenically driven events between 2018 and 2024. Our results correlate with related data products. We estimate that approximately 44 % of Fennoscandia consists of natural landscapes. Anthropogenic changes are detected in 6 % of the total land area; however, this estimate is subject to certain limitations.

Chapter 2

Publications & Open Source Contributions

2.1 Research Articles

Parts of this thesis have been published in the following peer-reviewed journal and conference articles, where I was the main author:

- T. T. Stomberg, L. A. Reißner, M. G. Schultz, and R. Roscher (2025). "Building consistency in explanations: Harmonizing CNN attributions for satellite-based land cover classification". In: *Machine Learning with Applications* 20, p. 100653. ISSN: 26668270. DOI: 10.1016/j.mlwa.2025.100653

- T. T. Stomberg, J. Leonhardt, I. Weber, and R. Roscher (2023). "Recognizing protected and anthropogenic patterns in landscapes using interpretable machine learning and satellite imagery". In: *Frontiers in Artificial Intelligence* 6, p. 1278118. ISSN: 2624-8212. DOI: 10.3389/frai.2023.1278118

- T. Stomberg, I. Weber, M. Schmitt, and R. Roscher (2021). "jUngle-Net: Using explainable machine learning to gain new insights into the appearance of wilderness in satellite imagery". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-3-2021, pp. 317–324. ISSN: 2194-9042. DOI: 10.5194/isprs-annals-V-3-2021-317-2021

Parts of this thesis have also been published as a preliminary paper:

- T. T. Stomberg, T. Stone, J. Leonhardt, I. Weber, and R. Roscher (2022). "Exploring wilderness characteristics using explainable machine learning in satellite imagery". In: *arXiv (cs)*. DOI: 10.48550/arXiv.2203.00379

There are additional publications I contributed to that are not part of this thesis:

- J. Kierdorf, T. T. Stomberg, L. Drees, U. Rascher, and R. Roscher (2024). "Investigating the contribution of image time series observations to cauliflower harvest-readiness prediction". In: *Frontiers in Artificial Intelligence* 7, p. 1416323. ISSN: 2624-8212. DOI: 10.3389/frai.2024.1416323

- A. Emam, T. T. Stomberg, and R. Roscher (2024). "Leveraging activation maximization and generative adversarial training to recognize and explain patterns in natural areas in satellite imagery". In: *IEEE Geoscience and Remote Sensing Letters* 21, pp. 1–5. ISSN: 1545-598X, 1558-0571. DOI: 10.1109/LGRS.2023.3335473

- B. Ekim, T. T. Stomberg, R. Roscher, and M. Schmitt (2023). "MapInWild: A remote sensing dataset to address the question of what makes nature wild". In: *IEEE Geoscience and Remote Sensing Magazine* 11.1, pp. 103–114. ISSN: 2168-6831, 2473-2397, 2373-7468. DOI: 10.1109/MGRS.2022.3226525

- C. Betancourt, T. T. Stomberg, A.-K. Edrich, A. Patnala, M. G. Schultz, R. Roscher, J. Kowalski, and S. Stadtler (2022). "Global, high-resolution mapping of tropospheric ozone – explainable machine learning and impact of uncertainties". In: *Geoscientific Model Development* 15.11, pp. 4331–4354. ISSN: 1991-9603. DOI: 10.5194/gmd-15-4331-2022

- C. Betancourt, T. Stomberg, R. Roscher, M. G. Schultz, and S. Stadtler (2021). "AQ-Bench: a benchmark dataset for machine learning on global air quality metrics". In: *Earth System Science Data* 13.6, pp. 3013–3033. ISSN: 1866-3508. DOI: 10.5194/essd-13-3013-2021

2.2 Datasets

As part of this thesis and as a main author, I have published the following datasets:

- **AnthroProtect 2.0**:
  https://phenoroam.phenorob.de/geonetwork/srv/eng/catalog.search#/metadata/fbaac894-ce3f-4baf-89a6-c1caf9b3017c

- **AnthroProtect**:
  https://phenoroam.phenorob.de/geonetwork/srv/eng/catalog.search#/metadata/6b1b0977-9bc0-4bf3-944e-bc825e466435

2.3 Repositories

As part of this thesis and as a main author, I have published the following repositories:

- **Building consistency in explanations: Harmonizing CNN attributions for satellite-based land cover classification**:
  `https://gitlab.jsc.fz-juelich.de/kiste/harmon`
- **Recognizing protected and anthropogenic patterns in landscapes using interpretable machine learning and satellite imagery**:
  `https://gitlab.jsc.fz-juelich.de/kiste/asos`
- **AnthroProtect Dataset Export**:
  `https://gitlab.jsc.fz-juelich.de/kiste/anthroprotect`

Chapter 3

Theoretical Background

This chapter provides the theoretical foundation for understanding the methods and applications presented in this thesis, and outlines influential research in machine learning for land cover classification. It also establishes the conceptual definitions used throughout this thesis. For a reference to the notation employed in this work, see Section *Notation* on page 167.

3.1 Multispectral Satellite Imagery

Multispectral satellites observe the Earth's surface using onboard instruments that detect sunlight reflected from the ground. Solar radiation spans a broad spectrum of wavelengths, and multispectral instruments measure this radiation across several distinct spectral bands. These bands target wavelength ranges that pass through the atmosphere with minimal absorption, including visible light, near-infrared, and shortwave infrared. In 1972, Landsat 1, the first satellite equipped with a multispectral sensor, was launched by the National Aeronautics and Space Administration (NASA) and the United States Geological Survey (USGS). It was followed by numerous identical and upgraded Landsat satellites, with the most recent, Landsat 9, entering orbit in 2021 (Wulder et al., 2022). In 2015, the first multispectral satellite of the Sentinel-2 mission was launched by the European Space Agency (ESA). Among freely available satellite data, Sentinel-2 stands out with a spatial resolution of 10 meters, a revisit time of 2 to 5 days, and 13 spectral bands (Drusch et al., 2012). Both missions play an important role in the field of land cover classification, not least because their satellite imagery is freely available and portals like Google Earth Engine (Gorelick et al., 2017) make them easily accessible. By capturing different spectral ranges, multispectral satellites are well-suited and often used for observing vegetation health, agricultural crop, water use, geological features, land cover, land use, and anthropogenic impacts (Wulder et al., 2022; Phiri et al., 2020; Misra et al., 2020;

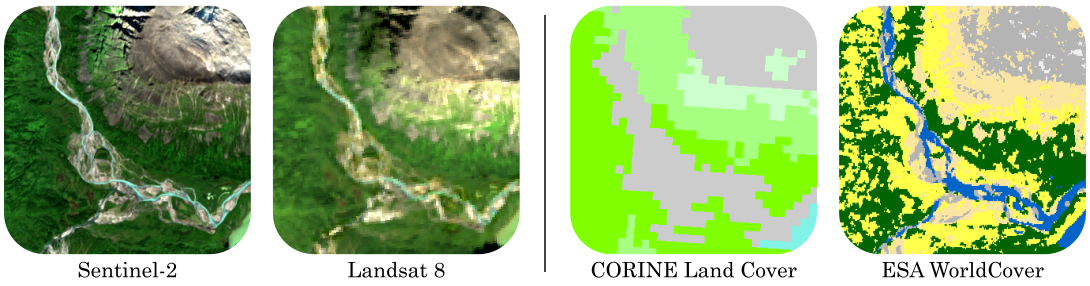


Figure 3.1: Sarek National Park, located in Lapland, northern Sweden, is one of the oldest national parks in Europe. The image shows a 3 km-wide section of one of its valleys in August 2023. The Sentinel-2 image has a spatial resolution of 10 meters, while Landsat 9 offers a resolution of 30 meters. The CORINE Land Cover product (2018) includes 44 classes with a resolution of 100 meters, whereas the ESA World Cover (2021) provides 11 classes at a 10-meter resolution. (CORINE: Broad-leaved forest, Moors and heathland, Sparsely vegetated areas, Water bodies, and Bare rock; ESA WorldCover: Trees, Grassland, Moss and lichen, Open water, Barren / sparse vegetation, and Snow and ice)

Segarra et al., 2020). In addition to multispectral data, radar images such as from Sentinel-1 (Torres et al., 2012) are also regularly used for environmental monitoring (Ienco et al., 2019). Using satellite imagery allows for continuous monitoring of broad regions, and frequent revisits over the same locations enable continuous observation and monitoring changes in Earth's land cover and land use. Figure 3.1 presents satellite images from Sentinel-2 and Landsat 9 alongside their corresponding land cover.

Sentinel-2

Sentinel-2 is a pair of multispectral satellites with a spatial resolution of 10 meters, a revisit time of 2 to 5 days, and 13 spectral bands.

3.2 Land Cover Classification

One of the most prominent applications of multispectral satellite imagery is the automatic mapping of land cover (Zhang et al., 2016). It is fundamentally based on the spectral signatures and textures of different surface types. Typical classes are forest, shrubland, grassland, cropland, built-up, water, and wetland. These can be further subdivided into subclasses such as broad-leaved forest, coniferous forest, mixed forest, and others. In contrast to land cover classes, land use describes the various ways in which different areas of land are utilized by humans, such as residential, industrial, agricultural, and recreational purposes.

Land cover classification plays a significant role in practical applications such as change detection (Chughtai et al., 2021), crop type mapping (Machichi et al., 2023), resource management (Thackway et al., 2013; Mashala et al., 2023), dis-

aster management (Bello and Aina, 2014), climate change (Roy et al., 2022), and environmental monitoring (Alotaibi and Nassif, 2024). There exist many large-scale land cover and land use datasets that are created using multispectral and radar satellite imagery, combined with expert knowledge and machine learning predictions. The Copernicus program provides the CORINE Land Cover dataset with 44 land cover classes and a minimum mapping width of 100 meters, published every six years for Europe (European Environment Agency, 2018). Copernicus also provides the Urban Atlas Land Cover/Land Use dataset for urban areas in Europe with 17 urban classes and a minimum mapping unit of about 50 meters (European Environment Agency, 2021). A particularly high resolution of 10 meters is offered by ESA WorldCover (Zanaga et al., 2022). It covers not only Europe but the entire world with 11 classes. Similarly, Dynamic World (Brown et al., 2022) delivers a global 10-meter land cover product with 9 classes, at the temporal resolution of Sentinel-2 imagery, which ranges from 2 to 5 days. Figure 3.1 presents a sample from the CORINE Land Cover and ESA WorldCover datasets.

Often based on such large-scale land cover datasets, several benchmark datasets for land-cover classification have been published. Examples are EuroSAT by Helber et al. (2019), BigEarthNet by Sumbul et al. (2021), Ben-ge by Mommert et al. (2023), Sen12MS by Schmitt et al. (2019a), and DFC2020 by Schmitt et al. (2019b). These benchmark datasets are frequently used for evaluating classification tasks (Papoutsis et al., 2022; Jain et al., 2024) or weakly-supervised learning methods (Robinson et al., 2021; Hanna et al., 2023).

Machine Learning for Land Cover Classification

Machine learning models are successfully used in remote sensing for various tasks such as classification, detection, or parameter prediction. They are well-suited for extracting complex patterns and relationships from large datasets. Traditional land cover classification methods, such as k-Nearest Neighbors, maximum likelihood classification, Support Vector Machines (Boser et al., 1992), decision tree algorithms, and Random Forests (Breiman, 2001), have been widely used for land cover classification (Richards, 2022). However, these methods are prone to overfitting in high-dimensional spaces (Richards, 2022; Pal and Mather, 2003), a challenge commonly attributed to the curse of dimensionality (Bishop, 2006). Therefore, these methods are usually applied on a pixel-wise basis, overlooking important information such as texture and structure, which are crucial for accurate predictions. In contrast, deep learning has emerged as a powerful tool capable of processing vast amounts of remote sensing data, using receptive fields to effectively capture spatial context. Ma et al. (2019) review on the variety of remote sensing applications in deep learning. For image analysis including satellite

imagery, convolutional neural networks are most commonly used (Song et al., 2019; Kattenborn et al., 2021). Since the proposal of the Vision Transformer by Dosovitskiy et al. (2021), transformer-based architectures have also become increasingly popular for remote sensing analyses (Aleissaee et al., 2023).

3.3 Deep Learning

Deep learning is a subfield of machine learning that uses artificial neural networks. These models are composed of multiple, interconnected layers, each performing relatively simple mathematical operations. Nonetheless, the depth and size of these networks enable them to learn complex patterns and solve challenging tasks.

A deep learning model is trained to approximate a mapping from input data $\mathcal{X}$ to target data $\mathcal{Y}$. It consists of neural layers that define a nonlinear, differentiable function $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\boldsymbol{y}}$ with $\boldsymbol{x} \in \mathcal{X}$. The model parameters $\boldsymbol{\theta}$ are optimized so that the predictions $\hat{\boldsymbol{y}}$ closely match the targets $\boldsymbol{y} \in \mathcal{Y}$. This is realized by minimizing a loss function $L(\hat{\boldsymbol{y}}, \boldsymbol{y})$ that measures the dissimilarity between predictions and ground truth. Commonly used loss functions are mean square error (MSE) for regression tasks; cross entropy and binary cross entropy for classification tasks.

3.3.1 Fully Connected Neural Networks

The simplest structure of an artificial neural network is a fully connected one, also called multilayer perceptron. It is built of neurons organized in layers as shown in Figure 3.2. Every neuron i in one layer l is connected to every neuron in the previous layer, performing a linear operation, and followed by a non-linear activation function σ. The weights $\boldsymbol{w}_{l,i}$ and biases $\boldsymbol{b}_l$ represent the trainable parameters $\boldsymbol{\theta}$. Given the inputs $\boldsymbol{x}_{l-1,i}$, the output of neuron j in layer l is given by:

$$x_{l,j} = \sigma\left(\left(\sum_i w_{l,ij} x_{l-1,i}\right) + b_l\right) \tag{3.1}$$

Commonly used activation functions include the sigmoid function, hyperbolic tangent (tanh), and rectified linear unit (ReLU).

Fully connected neural networks can process any data that can be represented as vectors. However, data types like images or sequences benefit from specialized architectures with mathematical operations adapted to their structure. In computer vision, convolutional neural networks have been the main choice for many years and remain highly relevant today. Transformer-based models have been widely applied to sequential data and, since the introduction of the Vision Transformer (Dosovitskiy et al., 2021), to image data as well. Fully connected layers are often components of these architectures. In many convolutional neural

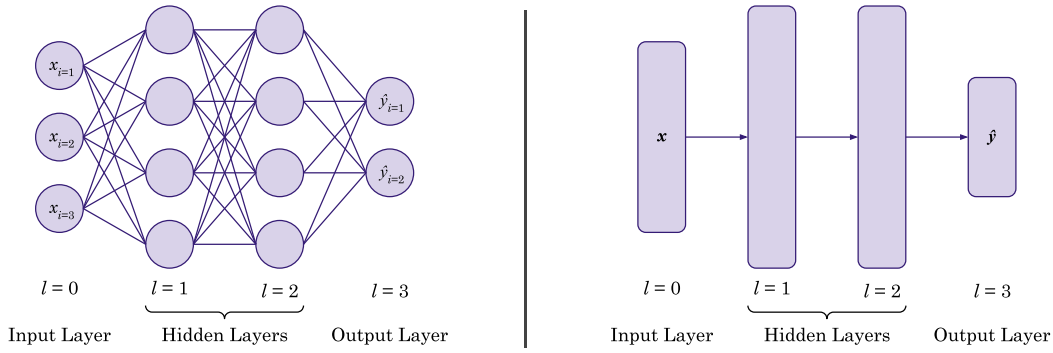


Figure 3.2: Fully connected neural networks consist of layers in which every neuron is connected to every neuron in the previous layer. There are two common ways to illustrate such networks: on the left, each neuron and its connections are shown individually; on the right, all neurons within a layer are represented collectively as a single box. Within this thesis, fully connected layers are consistently visualized in ■ purple.

networks they are used at the final stage to reduce the feature map to an output vector.

3.3.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are one of the preferred approaches for image analysis. Images are three-dimensional tensors $\boldsymbol{X}_0 \in \mathbb{R}^{K_0 \times H_0 \times W_0}$, where K denotes the number of channels, and H and W represent the height and width, respectively.[1] They can be effectively analyzed by applying convolutional operations with filters in $\mathbb{R}^{K_l \times H_{F_l} \times H_{F_l}}$, where H_{F_l} is the filter's kernel size at layer l. These filters perform a two-dimensional convolution across the spatial dimensions of the input. At each spatial location, the convolutional operation performs a linear transformation of the input values, followed by a non-linear activation function — similarly to Equation 3.1. In this way, filters can, for example, respond to colors, edges or textures. The filters' parameters in a CNN are learned. A convolutional layer typically consists of multiple filters, each producing one channel of the resulting feature map $\boldsymbol{X}_l \in \mathbb{R}^{K_l \times H_l \times W_l}$. The three-dimensional shape of a feature map is comparable to that of an image. In addition to convolutional layers, CNNs commonly include pooling layers, which reduce the spatial dimensions of feature maps by aggregating local regions, typically using average or maximum operations. See the box below for our definitions of the terms *feature map*, *feature vector*, and *feature* in CNNs.

[1] We use K to denote the number of channels, as C is reserved for the number of classes in classification tasks.

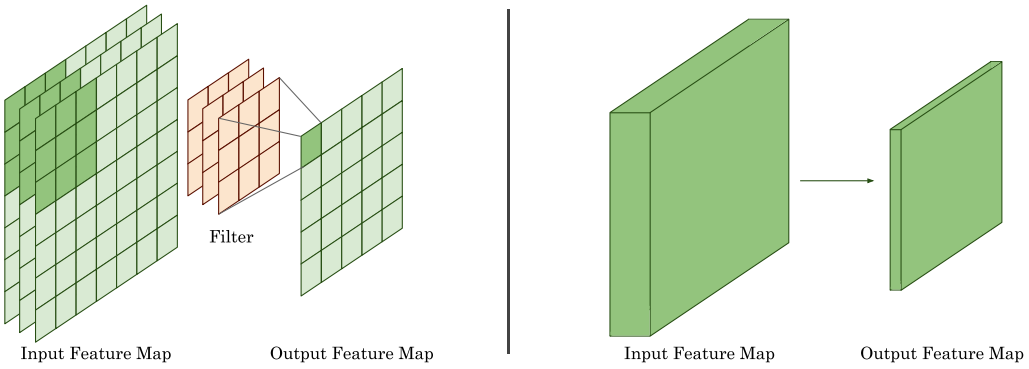


Figure 3.3: Within a CNN, filters perform two-dimensional convolutions across the spatial dimensions of the input. On the left is an illustration of a single convolution using one filter. For simplicity, feature maps are commonly represented as boxes, as shown on the right. Within this thesis, convolutions are consistently visualized in green; pooling operations (optionally with convolutions) in red; strided convolutions in yellow.

Feature Map, Feature Vector, and Feature

We denote the output of a convolutional or pooling layer l as *feature map* $\boldsymbol{X}_l \in \mathbb{R}^{K_l \times H_l \times W_l}$ with K_l channels, height H_l and width W_l. The channel-wise vectors at each spatial location (h_l, w_l) are referred to as *feature vectors* in $\mathbb{R}^{K_l}$. They are the "pixels" of the feature map. A single *feature* is a scalar of $\mathbb{R}$ at position (k_l, h_l, w_l). We also term the input image $\boldsymbol{X}_0$ an (input) feature map.

Each layer in a CNN outputs a feature map in which the feature vectors encode increasingly complex properties of the input data. Convolutions and poolings progressively reduce spatial dimensionality, thereby increasing the receptive field of the filters. Beginning with pixel colors, this enables the network to capture edges, textures, and, with deeper layers, more task-specific patterns. Since the features must represent increasingly complex patterns, CNNs are typically designed so that as the spatial dimensions decrease, the number of channels increases. A corresponding architecture, the VGG-16, is illustrated in Figure 3.4. Finally, fully connected layers integrate these high-level feature representations into a final prediction. A well-trained CNN thus learns highly informative feature representations, allowing it to make reliable predictions such as classification.

Feature Representation

Feature representations are structured encodings of information that capture the essential characteristics and properties of data relevant to a specific task. For example, the feature map of a CNN's layer is a feature representation of the input image. The corresponding feature vectors represent concepts that may be interpretable by humans.

Numerous CNN architectures have been developed over the years. Among them, we introduce the most relevant to this thesis: VGG, ResNet, and U-Net, all of which are well-established and widely adopted in computer vision. While VGG and ResNet are commonly used for image-level classification or regression, U-Net is built for pixel-wise tasks, such as segmentation. All three follow the common architectural design described earlier. Even U-Net adheres to this approach, creating a spatial bottleneck before restoring the original spatial dimensions. It has been shown that deeper layers tend to capture more abstract, high-level feature representations (Zeiler and Fergus, 2014; Olah et al., 2017; Olah et al., 2018; Carter et al., 2019; Selvaraju et al., 2020).

Image-level and Pixel-level Task

In computer vision, image-level tasks involve making predictions for an entire image as a whole, while pixel-level tasks involve making predictions for each individual pixel within the image. Commonly used CNN architectures for image-level tasks include VGG and ResNet, while for pixel-level tasks, U-Net is widely used.

VGG

The VGG architecture, developed by the **V**isual **G**eometry **G**roup (Simonyan and Zisserman, 2015), consists of convolutional and pooling layers followed by fully connected layers. It is typically implemented in 11-, 13-, 16-, or 19-layer variants, where the count refers to the number of trainable layers. Figure 3.4 illustrates the VGG-16 version. The convolutional layers are batch normalized and all layers are ReLU activated.

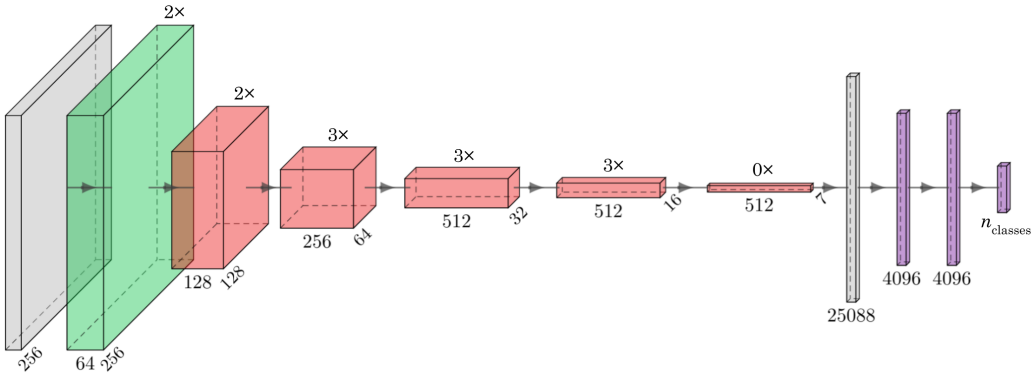


Figure 3.4: VGG-16 consists of 16 convolutional and fully connected layers. Additionally, max pooling operations are applied throughout the network. The final pooling layer is adaptive and produces a feature map with a height and width 7, regardless of the input image size. (convolution, pooling and convolution, copy, fully connected. The number of convolutions in each illustrated block is indicated above it. The illustration is created using PlotNeuralNet by Iqbal, 2018.)

ResNet

A ResNet, developed by He et al. (2016), incorporates **Res**idual connections that add feature maps from earlier layers to deeper ones. This helps mitigate the vanishing or exploding gradient problem (Bengio et al., 1994) and thus enables the training of much deeper neural networks. ResNet is proposed in 18-, 34-, 50-, 101-, and 152-layer configurations, with the number referring to the total count of trainable layers. The architecture of ResNet-18 is depicted in Figure 3.5. Each convolutional layer is followed by batch normalization and ReLU activation is used throughout the network. The spatial dimensions are mostly reduced through strided convolutions instead of pooling. While ResNet-18 and ResNet-34 use 2-layer residual blocks, ResNet-50 and deeper variants use 3-layer residual blocks.

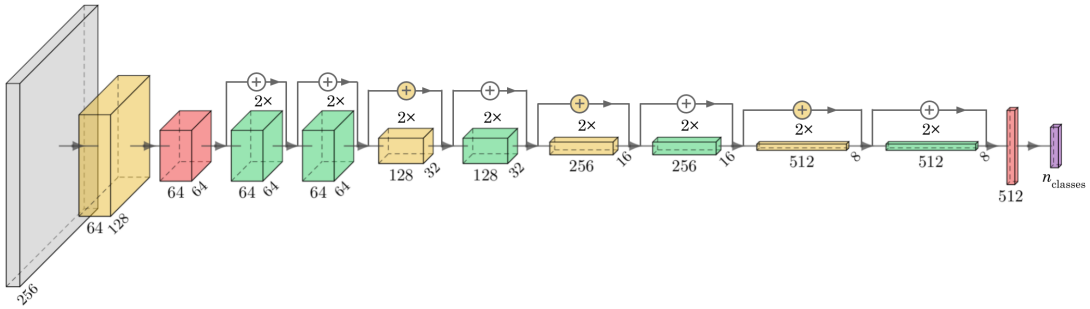


Figure 3.5: A ResNet includes residual connections that add feature maps from earlier layers to deeper ones, enabling the training of much deeper CNNs. The illustrated ResNet-18 consists of 18 trainable layers and downsampling is mostly realized by strided convolutions. The final pooling layer performs average pooling over the spatial dimensions, which is called global average pooling. (■ convolution, ■ strided convolution ■ pooling and convolution, ■ fully connected. The number of convolutions in each illustrated block is indicated above it. The illustration is created using PlotNeuralNet by Iqbal, 2018.)

U-Net

The U-Net was developed for medical image segmentation by Ronneberger et al. (2015). Today, it is widely used for pixel-wise classification and regression tasks. The original U-Net architecture consists of four downsampling steps, creating a spatial bottleneck with high-dimensional feature vectors. This bottleneck is then upsampled through four corresponding upsampling steps. Feature maps from earlier layers are copied to deeper layers, which helps recover information about small structures. The U-Net architecture is commonly used in various forms and in this thesis, we apply the following modifications compared to the original one: 1) We add padding to each convolution to preserve the spatial dimensions, ensuring consistent sizes at each skip connection. 2) We replace deconvolutional upsampling with bilinear upsampling, as proposed by Odena et al. (2016), to prevent checkerboard artifacts. 3) We apply batch normalization

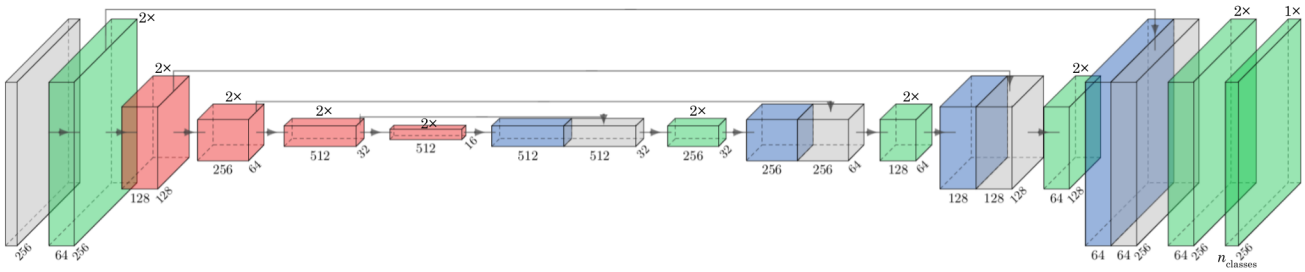


Figure 3.6: A U-Net is used for pixel-wise tasks. The feature maps are first downsampled and then upsampled back to their original size. Skip connections copy feature maps from earlier layers to deeper ones. (■ convolution, ■ pooling and convolution, ■ upsampling, ■ copy. The number of convolutions in each illustrated block is indicated above it. The illustration is created using PlotNeuralNet by Iqbal, 2018.)

after each convolutional layer. The modified version of U-Net is illustrated in Figure 3.6.

3.4 Explainable Machine Learning

Explainable machine learning techniques aim to make the predictions and internal workings of machine learning models understandable and interpretable to humans. The field has been actively promoted in research in recent years (Roscher et al., 2020b) and is also widely applied in the field of remote sensing (Höhl et al., 2024; Roscher et al., 2020a). There exist various techniques for explanations in deep learning (Samek et al., 2021) which can be broadly categorized into global and local approaches: Global explanations aim to capture the overall behavior of a model, such as how it performs on average across an entire dataset, revealing general patterns, feature importance, or decision rules. In contrast, local explanations focus on single predictions, clarifying why the model made specific decisions, e.g. by highlighting the impact of features in that instance.

Global and Local Explanations

Global and local explanations refer to explainable machine learning techniques that explain the overall behavior of the model or the predictions for individual samples, respectively.

3.4.1 Mechanistic Interpretability

Research in mechanistic interpretability contributes to a general understanding of how neural networks process information. The field examines the internal mechanisms of neural networks, focusing on two aspects: the features that represent learned concepts and the circuits — groups of neurons responsible for specific

computations (Saphra and Wiegreffe, 2024). While the term "mechanistic interpretability" is mostly associated with large language models, related research by Olah et al. (2017), Olah et al. (2018), and Carter et al. (2019) explores how individual CNNs process image data.

3.4.2 Attribution Methods

While mechanistic interpretability focuses on the internal mechanisms of neural networks, attribution methods primarily aim to examine individual input-output relationships. They evaluate the importance of features based on how much they influence the model's predictions, assigning significance scores to features according to their impact.

Feature Attribution

A feature attribution quantifies how much a feature influences the model's prediction, assigning an importance score based on its impact.

Attribution methods can be largely divided into two fundamental groups: Perturbation-based and gradient-based attribution methods. Perturbation-based attribution methods modify the features and evaluate resulting changes in the prediction. A first attempt was made with occlusion sensitivity by Zeiler and Fergus (2014) who occluded patches in the input image to identify changes in the model's output. These changes are a measure of the occluded areas' sensitivities in regards to the observed class. Gradient-based attribution methods involve the gradients of the model's prediction with respect to the features. Here, a first attempt was the computation of "Saliency Maps" by Simonyan et al. (2014), who computed the absolute gradients for an input image. Since Saliency Maps tend to be noisy, numerous extensions, variations and more complex methods have been developed over time, leading to a large number of available techniques. Examples are Layer-wise Relevance Propagation (Bach et al., 2015), Gradients×Features (Shrikumar et al., 2017), Integrated Gradients (Sundararajan et al., 2017), DeepLift (Shrikumar et al., 2017), Gradient SHAP (Lundberg and Lee, 2017), DeepLift SHAP (Lundberg and Lee, 2017), and Grad-CAM (Selvaraju et al., 2020).

Different attribution methods can produce significantly different results, which is one of the main topics discussed in this thesis. Furthermore, attribution results can vary depending on whether they are computed for the input features or for features from a deeper layer. This variation arises from the way feature representations evolve across layers in neural networks, as explained in Section 3.3.2, and is also examined in this thesis.

Attributions are computed using the non-activated output of the model because activation functions like sigmoid can saturate the output range, potentially

obscuring the model's true sensitivity to input features. Therefore, throughout this section (unless specified otherwise), *gradients* refer to the partial derivatives of the non-activated output with respect to the features. Similarly, occlusion-based methods are also computed using the non-activated output. All attribution methods are explained in the context of feature maps in CNNs; however, most of them can be analogously applied to other neural networks with different data structures.

Gradients in the Context of Attributions

Throughout this section (unless specified otherwise), *gradients* refer to the partial derivatives of the non-activated output with respect to the features.

Occlusions

Occluding sections of a feature map influences the model's prediction score. If the occluded part is relevant for predicting a certain class, it reduces the prediction score; if it contradicts the class, the score increases. The attribution for each occluded pixel is then computed as the negative of this score change, divided by the number of occluded pixels. A sketch illustrating this principle is shown in Figure 3.7. By systematically applying this process, one obtains an attribution map. Zeiler and Fergus (2014) were the first to apply this approach to CNNs using a sliding window. Petsiuk et al. (2018) extended this method by randomly occluding parts of the image. For deeper feature maps, we propose occluding similar feature vectors simultaneously, for example, by applying k-means clustering (see Section 5.3).

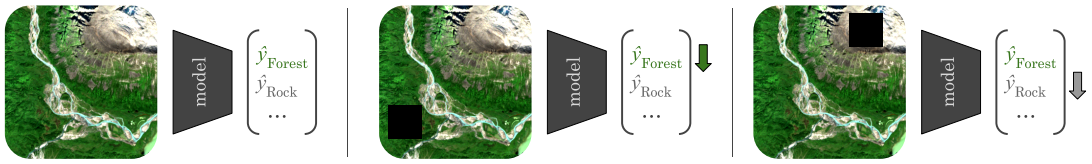


Figure 3.7: By systematically covering sections of a feature map (here the input image), one can observe how the model's prediction changes, revealing the attributions of the occluded regions.

Gradients×Features

Simonyan et al. (2014) were the first to compute gradient maps for model explanations, referring to them as Saliency Maps. They use the maximum absolute gradient value across channels to indicate attribution, without distinguishing between positive and negative contributions. Alternatively, Saliency Maps are often computed using the sum of gradients across channels considering both, negative and positive values. Shrikumar et al. (2017) propose to multiply the gradients

with the corresponding feature vectors, known as Gradients×Features. For class c and feature map $\boldsymbol{X}$ with channels k, the attribution map $\boldsymbol{R}^c$ is defined as:

$$\boldsymbol{R}^c = \sum_k \left[\frac{\partial \hat{y}^c}{\partial \boldsymbol{X}_k} \right]_{i,j} \circ \boldsymbol{X}_k \tag{3.2}$$

The Hadamard product $\circ$ is the element-wise product of two matrices, and the derivative is also computed element-wise as indicated by $[\,]_{i,j}$. Multiplying attributions by their corresponding feature values is generally a common practice, as discussed in Section 3.4.3.

Gradient-weighted Class Activation Mapping (Grad-CAM)

Zhou et al. (2016) propose a method which leverages a global average pooling layer to produce attribution maps, called Class Activation Mapping (CAM). This approach works with CNNs whose last two layers are global average pooling followed by a fully connected layer, such as a ResNet. For a class c, the attribution map $\boldsymbol{R}^c$ is the sum of the feature map channels $\boldsymbol{X}_k$ from the last convolutional layer, weighted by the corresponding weights w_k^c of the fully connected layer:

$$\boldsymbol{R}^c = \sum_k w_k^c \boldsymbol{X}_k \tag{3.3}$$

Since CAM is limited to specific architectures only, Selvaraju et al. (2020) generalize this approach computing global averaged gradients for each channel of a feature map instead of leveraging the global average pooling values:

$$\boldsymbol{R}^c = \sum_k \alpha_k^c \boldsymbol{X}_k \;, \quad \alpha_k^c = \frac{1}{H\,W} \sum_i^{H \times W} \frac{\partial \hat{y}^c}{\partial X_{k,i}} \tag{3.4}$$

H and W are the height and width of the feature map $\boldsymbol{X}$, and $\boldsymbol{X}_{k,i}$ is the feature value at channel k and position i. The authors call their method Gradient-weighted Class Activation Mapping (Grad-CAM) and it can be applied to any CNN. Originally only the positive Grad-CAM values are considered for the sake of clearer visualization. In this thesis, however, both positive and negative values are analyzed.

Layer-wise Relevance Propagation (LRP)

In Layer-wise Relevance Propagation (LRP) by Bach et al. (2015), the model's prediction is backpropagated through the network layers using specific propagation rules that assign a relevance score to each feature. These rules ensure that the relevance is preserved across layers throughout the network. Each feature

passes as much relevance to the lower layer as it received from the deeper one. A commonly used propagation rule is the Epsilon Rule. For class c, it defines the relevance $A^c_{l-1,i}$ of feature $X_{l-1,i}$ corresponding to neuron i in layer $(l-1)$ as:

$$R^c_{l-1,i} = \sum_j \frac{w_{l,ij} X_{l-1,i}}{\epsilon + \sum_i w_{l,ij} X_{l-1,i}} R^c_{l,j} \ , \quad \epsilon \geq 0 \tag{3.5}$$

The equation estimates the total relevance that feature $X_{l-1,i}$ has on the features of the subsequent layer l. It further relates this to the relevance of the other features in layer $(l-1)$. The epsilon ϵ ensures numerical stability when the denominator is small. Setting $\epsilon = 0$, the rule is referred to as the Basic Rule or z-Rule. For an attribution map, the relevance of the features are summed up across the channels.

LRP is limited to specific neural network layers and activation functions. Although there are recent extensions — for example, to backpropagate through skip connections in ResNets (Otsuki et al., 2025) — LRP remains only conditionally flexible in its applicability. Under the Basic Rule, and provided that all activations are piecewise linear, LRP is equivalent to Gradients×Features (Shrikumar et al., 2017).

Integrated Gradients

To compute Integrated Gradients (Sundararajan et al., 2017), a baseline feature map X' must be chosen, which is often set to zeros. Gradients are then averaged as the features transition linearly from the baseline map to the original feature map.

$$R^c_i = (X_i - X'_i) \times \int_{\alpha=0}^{1} \frac{\partial f(X' + \alpha \times (X - X'))}{\partial X_i} \partial\alpha \tag{3.6}$$

The result is multiplied by the difference between the feature map and the baseline map. The integral is approximated using a sum computed discretely over a set of data points with sufficiently small intervals. Also, the attributions of the features are summed up across the channels.

Deep Learning Important FeaTures (DeepLIFT)

Similar to Layer-wise Relevance Propagation (LRP), DeepLIFT by Shrikumar et al. (2017) defines backpropagation rules to compute *contribution scores*. Unlike the relevance scores in LRP, DeepLIFT computes contribution scores relative to a baseline feature map, which is often set to zeros. While LRP ensures that the sum of relevance scores is conserved across layers, DeepLIFT preserves contribution by ensuring that the sum of contributions at a given layer l equals the difference

between the predictions $\Delta\hat{y}^c$ of the original input and the baseline input:

$$\Delta\hat{y}^c = \sum_i R^c_{l,i} \quad \forall \ l \in \{0, ..., L\} \tag{3.7}$$

Two rules are defined — one for linear operations and one for non-linear operations. For convolutional or fully connected layers, the Linear Rule applies. The relevance of neuron i in layer $(l-1)$ for neuron j in layer l is given by:

$$R^{l,j}_{l-1,i} = w_{l,ij}\Delta X_{l-1,i} \tag{3.8}$$

For non-linear computations, such as activation functions, the Rescale Rule is applied. Here, positive and negative contributions are computed separately before combining them. This approach is necessary for certain non-linear operations, particularly those involving functions like ReLU.

$$R = R^+ + R^- \ , \quad (R^{l,j}_{l-1,i})^{+/-} = \frac{\Delta X_{l,j}}{\Delta X_{l-1,i}}\Delta X^{+/-}_{l-1,i} \tag{3.9}$$

Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) by Ribeiro et al. (2016) operates by constructing an interpretable surrogate model that approximates the behavior of the original model in the vicinity of the given features. To do this, LIME generates a set of perturbed data points around the selected input and evaluates the original model on each of them. The resulting predictions are then used to train a simpler, more transparent model — often a linear one. This surrogate model offers explanations without requiring access to the original model's internal mechanisms. Since a separate surrogate model must be trained for each instance to be explained, LIME can be computationally intensive. For image data, individual pixels are typically grouped into superpixels to reduce dimensionality, with each superpixel serving as an interpretable feature in the explanation.

SHapely Additive exPlanations (SHAP)

SHapely Additive exPlanations (SHAP), proposed by Lundberg and Lee (2017), is an attribution method based on the game-theoretic concept of Shapely values introduced by Shapley (1953). Shapley values aim to fairly distribute the total gain from a cooperative game among players, based on their individual contributions. The contributions are determined by evaluating all possible subsets of players. SHAP applies this concept to models, where the players are the features

and the game's gain is the model's prediction. Since every possible subset of features must be evaluated, features need to be omitted. For neural networks, this poses a challenge, as they must be replaced with neutral values. While methods like Occlusions, Integrated Gradients, and DeepLIFT often use zeros, SHAP takes a different approach: it samples features from the training data distribution to simulate neutrality. This process is usually repeated multiple times, and the results are averaged.

An important characteristic of SHAP is that feature attributions are additive, similar to DeepLIFT (Equation 3.7): The model's prediction for certain features equals the baseline prediction (all features omitted) plus the sum of the feature attributions. For additive attributions, the term *contributions* is also commonly used. SHAP cannot be exactly computed for neural networks, in part because the number of possible feature subsets increases factorially with the number of features. However, the following approaches provide approximate solutions, all proposed by Lundberg and Lee (2017):

- **Kernel SHAP** is a technique built upon the LIME framework. By using a linear surrogate model, selecting an appropriate loss function and weighting kernel, and omitting regularization, the conditions for SHAP are satisfied, allowing LIME to serve as an approximation for computing SHAP values. However, the computational intensity of LIME is also a disadvantage for Kernel SHAP.

- **DeepLift SHAP** combines concepts from SHAP and DeepLIFT. It applies DeepLIFT's backpropagation rules while sampling multiple reference baselines from the training data distribution. The resulting attributions are averaged to approximate SHAP values. DeepLift SHAP assumes that the analyzed features are independent and that the model behaves linearly.

- **Gradient SHAP** combines principles from SHAP, Integrated Gradients, and SmoothGrad. SmoothGrad reduces the noise in gradient-based attribution methods by averaging the attributions across multiple noisy versions of the features (Smilkov et al., 2017). Gradient SHAP applies integrated gradients while sampling reference baselines from the training data distribution and adding noise. The resulting attributions are averaged. Gradient SHAP assumes that the analyzed features are independent and that the model behaves linearly.

3.4.3 Global and Local Attribution Methods

Figure 3.8 illustrates that a feature x can have a different effect on a function $f(x)$ depending on whether the function is considered locally or globally.[2] Accordingly, there are different approaches to attribution methods: Local attribution methods evaluate how a prediction changes in response to infinitesimally small perturbations, whereas global attribution methods assess the effect of a feature on a prediction relative to a reference baseline (Ancona et al., 2018). While Saliency Maps simply compute gradients and thus indicate local attributions, most attribution methods rely on an explicit or implicit reference baseline, measuring global attributions:

- Integrated Gradients and DeepLIFT explicitly require a baseline. SHAP approximations extend these methods by, among other things, selecting baselines from the training data distribution.
- Gradients×Features is a special case of Integrated Gradients with a zero baseline and a step size of one.
- Grad-CAM also multiplies gradients with features, implicitly using a zero baseline.
- Occlusions can be interpreted as the product of a feature and the average gradient between the occlusion value (baseline) and the actual feature value.

Global and Local Attributions

Local attribution methods evaluate how a prediction changes in response to infinitesimally small perturbations, whereas global attribution methods assess the effect of a feature on a prediction relative to a reference baseline. Here, *local* and *global* follow the nomenclature of Ancona et al. (2018), based on mathematical terminology describing the behavior of functions. It is unrelated to the terms local and global *explanations*.

[2] Here, the terms *local* and *global* refer to the mathematical behavior of a function and are unrelated to local and global *explanations*.

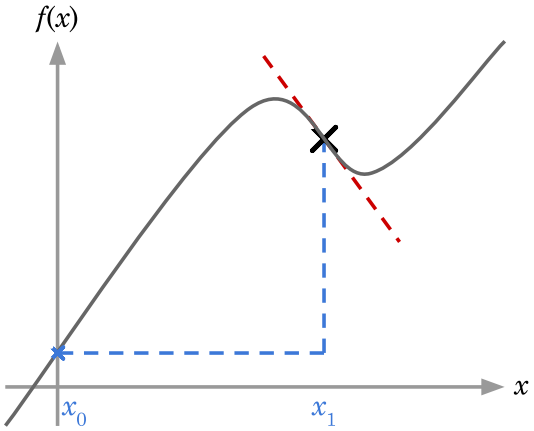


Figure 3.8: ■ Globally, feature x_1 has a positive impact on the illustrated function $f(x)$; ■ locally, the impact is negative. Most attribution methods aim to evaluate the global effect of a feature relative to a reference baseline x_0. The illustration also shows that greater non-linearity leads to increased variation in local gradients, resulting in noisier gradient-based attribution maps. Integrating over the gradients, as done in the Integrated Gradients method, is one approach to reduce this noise.

Part II

Improving Explanations of CNNs

Chapter 4

Related Work and Research Gaps

Machine learning models are able to find patterns and relations in large datasets and are successfully applied to a wide range of tasks. However, despite their widespread use, they frequently raise concerns due to lacking interpretability. Especially for scientific applications, the comprehensibility and reliability of machine learning results are crucial for ensuring scientific validity and increasing trust in the models. Therefore, there is a pressing need for more interpretability in machine learning, especially in domains where decisions have significant scientific, societal, or environmental implications.

To achieve these goals, explainable machine learning has been actively promoted in recent years (Samek et al., 2020; Roscher et al., 2020b) and is also widely applied in the field of remote sensing (Höhl et al., 2024; Roscher et al., 2020a). It aims to uncover the underlying reasons behind models' decisions and functioning, thereby building trust in model predictions, identifying obvious flaws (Lapuschkin et al., 2019), and helping to improve model performance.

4.1 Interpretability-by-Design Architectures

When solving a specific task, research often prioritizes predictive accuracy over interpretability which can, however, be detrimental for ensuring scientific validity. Decision-making processes of deep learning models may differ significantly from how humans approach a given task, making their explanations difficult to interpret. Concerning this matter, the design of specific architectures can result in more interpretable results. This may also become important when assessing whether a model's performance will hold in new environments or under changing conditions. Without insight into the model's reasoning, it is difficult to evaluate its reliability or to build justified trust in its predictions.

Interpretability-by-design architectures aim to constrain the training process so that the resulting internal representations and decision-making steps are in-

herently more transparent and aligned with human-understandable concepts — for example, by imposing a specific representation on model components or latent variables. A prominent work in this field is network dissection, where semantic labels are assigned to individual units based on their alignment with visual concepts (Zhou et al., 2019). *Bag-of-local-features models* by Brendel and Bethge (2018) are modified to disregard the spatial ordering of small, local image features, enabling better analysis of how different parts of an image influence classification. Koh et al. (2020) introduce concept bottleneck models that predict intermediate human-understandable concepts first (e.g. colors of a bird's feathers), before computing the final output (bird species). Similarly, Marcos et al. (2020) build a model with a semantic bottleneck that explicitly learns interpretable attributes that are meaningful for predicting landscape scenicness. Levering et al. (2020) expand this approach by predicting scenicness using Sentinel-2 imagery with an interpretable bottleneck layer for land cover classes.

Embedding existing concepts into the features of a neural network increases its interpretability but requires a labelling of these concepts and restricts the flexibility of training. We introduce a different approach to improve interpretability that does not explicitly derive conceptual insights but enables the generation of comprehensible, high-resolution attribution maps. Our innovative architecture, UH-Net, addresses the dilemma that attribution maps applied at the input level offer high resolution but are less informative, whereas those computed in deeper layers capture high-level features but at lower resolution (see Section 5.1).

4.2 The Diversity of Attribution Methods

To interpret CNN predictions, various attribution techniques have been developed that highlight important feature vectors via attribution maps. These are also regularly applied in remote sensing (Höhl et al., 2024). Well-known and relevant attribution methods are Sliding Window Occlusions (Zeiler and Fergus, 2014), Gradients×Features (Shrikumar et al., 2017), Layer-wise Relevance Propagation (LRP, Bach et al., 2015), Gradient-weighted Class Activation Mapping (Grad-CAM, Selvaraju et al., 2020), Integrated Gradients (Sundararajan et al., 2017), Deep Learning Important FeaTures (DeepLift, Shrikumar et al., 2017), Local Interpretable Model-agnostic Explanations (LIME, Ribeiro et al., 2016), and SHapely Additive exPlanations (SHAP, Lundberg and Lee, 2017) such as Gradient SHAP and DeepLift SHAP. The mechanisms of these methods are introduced in Section 3.4.2.

A key challenge is that different attribution methods often produce divergent explanations for the same prediction, which undermines confidence and raises questions about which method is most reliable for specific tasks. Kakogeorgiou

and Karantzalos (2021) evaluate their appearance and robustness on multi-label remote sensing benchmark datasets; Hsu and Li (2023) compare several qualitative abilities for geospatial datasets; Mohan and Peeples (2023) assay the robustness, faithfulness, randomization, complexity, localization, and axiomatic; and Nieradzik et al. (2024) assess their similarities with imagery from unmanned aerial vehicles. Some studies combine attribution methods to unite their respective advantages (Selvaraju et al., 2020; Gulum et al., 2021; Dhore et al., 2024).

From our perspective, previous comparisons have limited validity because the methods are applied to different layers. This is due to the fact that some attribution methods were originally designed for input-level analysis (e.g., Sliding Window Occlusions), while others target deeper layers (e.g., Grad-CAM). However, most attribution methods can, in principle, be applied to any layer, and the results can vary significantly depending on whether they are computed at the input or at a deeper layer. Therefore, in this work, we compare all attribution methods applied to the same layers. We also propose a harmonization technique in Section 5.2 to address the challenge of inconsistent attribution results across different methods.

4.3 Occlusions and Grad-CAM

The two attribution methods, Sliding Window Occlusions and Grad-CAM, are very popular. They have proven effective in various comparisons (Kakogeorgiou and Karantzalos, 2021; Adebayo et al., 2018; Yang and Kim, 2019) and are theoretically easy to understand. Several extensions to both methods have already been proposed.

Occlusion-based Attribution Methods

Occlusion-based attribution methods work by occluding parts of a feature map and measuring the sensitivity to these occluded regions. In this context, studies have explored which occlusion values are most appropriate, aiming for values that are as neutral as possible. Common choices include zeros, the average value of the occluded area, blurring, or Gaussian noise (Fong and Vedaldi, 2017). More advanced approaches generate the infilled regions using generative models, such as variational autoencoders or generative adversarial networks (Chang et al., 2019).

In contrast, previous research has paid less attention to which regions should be occluded simultaneously. Common implementations include a sliding window (Zeiler and Fergus, 2014) or random rectangular patches (Petsiuk et al., 2018). These procedures constrain the resolution of the resulting attribution map; and different or even opposing concepts can end up included within a single occlusion area. Approaches that segment the input into meaningful regions, such as

superpixels used in LIME (Ribeiro et al., 2016), have not been applied to occlusion methods. However, since occlusions are commonly applied to the input and colors carry only limited conceptual information, a superpixel approach is likely not very advantageous. This assumption, though, changes when occlusions are applied to deep layers, where similar features represent similar concepts. This becomes particularly important when the deep feature maps are high-resolution — as in our interpretable-by-design UH-Net. We therefore propose using feature-specific clustering to determine occlusion regions as presented in Section 5.3.

Grad-CAM

There are numerous extensions of Grad-CAM, particularly aimed at improving the computation or interpretation of the average channel gradients α_k^c (explained in Section 3.4.2). Grad-CAM++ (Chattopadhay et al., 2018) incorporates higher-order gradients for weighting the first-order gradients, thereby assigning greater weight to features with non-linear effects. Ablation-CAM (Desai and Ramaswamy, 2020) estimates the importance of each channel by ablating (completely occluding) one channel at a time and observing the impact on the output score. Eigen-CAM (Muhammad and Yeasin, 2020) finds the main direction of variation (principal eigenvector) in the feature maps; and Score-CAM (Wang et al., 2020a) works by creating soft masks from the feature maps, applying them to the input, and determining how much each mask affects the class score.

However, previous research has not addressed one limitation of Grad-CAM: Grad-CAM performs global average pooling of the gradients, resulting in a single mean gradient α_k^c for each channel k. In vector form, $\boldsymbol{\alpha}^c$ defines a fixed direction in the feature space onto which the feature vectors $\boldsymbol{X}_j$ are projected — as follows from Equation 3.4:

$$R_j^c = \sum_k \alpha_k^c \cdot X_{k,j} = \boldsymbol{\alpha}^c \cdot \boldsymbol{X}_j \tag{4.1}$$

In other words, Grad-CAM selects a consistent direction in the feature space and evaluates the attribution of each spatial feature vector based on its alignment with this direction, as illustrated in Figure 4.1. Therefore, its application is most appropriate when computed for a layer that is followed by global average pooling. The authors of Grad-CAM recommend applying it to the last convolutional layer, and this certainly also provides meaningful insights. However, for other layers it is possible that not just one direction in the feature space is relevant for the predicted class, but several. We propose a way to overcome this limitation in Section 5.3.

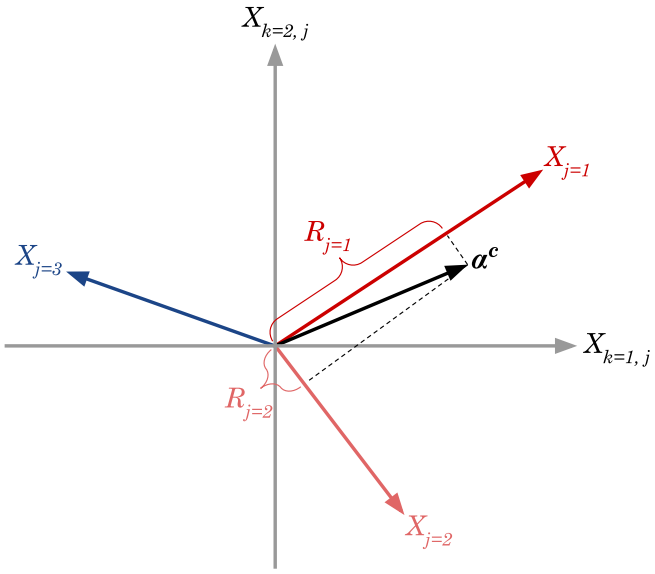


Figure 4.1: In Grad-CAM, a constant vector $\boldsymbol{\alpha}^c$ is computed for each feature map and defines a fixed direction in the feature space. The dot product between $\boldsymbol{\alpha}^c$ and feature vector $\boldsymbol{X}_{j=1}$ results in a high attribution score $\boldsymbol{R}_{j=1}$ because the two vectors point in a similar direction. In contrast, $\boldsymbol{X}_{j=2}$ is nearly orthogonal to $\boldsymbol{\alpha^c}$, yielding a low attribution $\boldsymbol{R}_{j=2}$. $\boldsymbol{X}_{j=3}$ points in the opposite direction, leading to a negative attribution.

Chapter 5

Novel Methodologies

In this section, we propose methodologies to improve the explainability of convolutional neural networks (CNNs). These include a novel architecture, UH-Net; a methodology for harmonizing attributions; and two new feature-specific attribution methods.

5.1 UH-Net Architecture

Attribution maps for CNNs can be computed at any layer including the input and deeper layers. Each approach has distinct advantages: input-level analysis offers the highest resolution, whereas deep-layer analysis yields more meaningful results by capturing high-level semantics. To combine both advantages, we propose an interpretable-by-design architecture combining a U-Net and a task-specific Head, which we refer to as UH-Net. An illustration is shown in Figure 5.1. The encoder-decoder based U-Net transforms input data of the form $\mathbb{R}^{K_0 \times H_0 \times W_0}$ into a new representation of the form $\mathbb{R}^{K_L \times H_0 \times W_0}$. This intermediate representation is then passed to the classification head, which classifies it into C classes. At the intermediate layer, attributions exhibit both high resolution and high-level semantics. Skip connections within the U-Net establish a strong link between deep and input features, ensuring that attributions remain closely tied to the input and are thus more interpretable.

As described in Section 3.3.2, we modify the original U-Net proposed by Ronneberger et al. (2015) by using bilinear upsampling and adding padding and batch normalization. If the input image size is 128 or smaller, we omit the second and fourth pooling. The strided convolutions in the head double the number of channels and use a kernel size of 6 and a stride of 3, with padding applied if data would otherwise be ignored. They are followed by two fully connected layers. All hidden layers are activated with ReLU. The number of channels K_L of the intermediate layer should be treated as a hyperparameter. UH-Net is trained end-

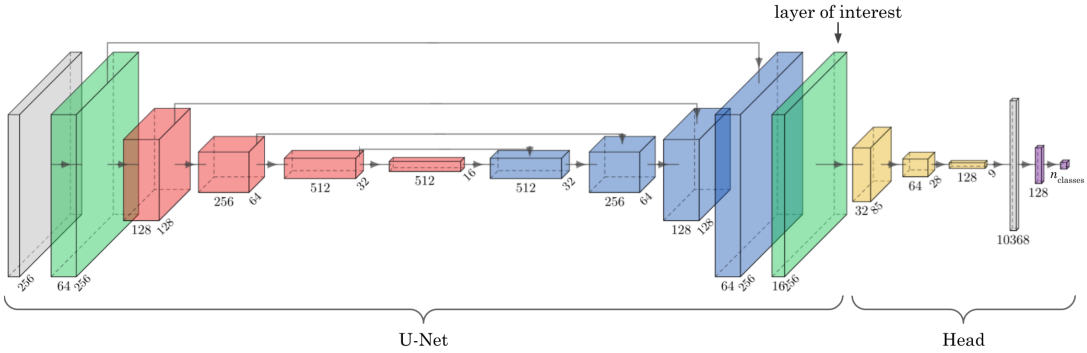


Figure 5.1: The interpretable-by-design UH-Net architecture combines a U-Net with a shallow classifier Head. The intermediate feature map matches the resolution of the input image and is of interest for computing the attribution maps. Shown is the architecture for an input image of size 256 × 256 and 16 channels at the intermediate layer. The illustration simplifies the U-Net architecture shown in Figure 3.6. (■ convolutions, ■ pooling and convolutions, ■ upsampling and convolutions, ■ strided convolution, ■ copy, ■ fully connected. The illustration is created using PlotNeuralNet by Iqbal, 2018.)

to-end. In Section 7.2, we investigate architectural variations whose explanations, however, are not as interpretable as those of the architecture proposed here.

5.2 Attribution Harmonization

The deeper the feature map being analyzed, the more high-level are the feature representations with respect to the task at hand. Here, a common assumption in mechanistic interpretability is that similar feature vectors represent similar concepts. Consequently, we argue that similar feature vectors have similar attributions. We propose a harmonization technique that leverages this property to produce meaningful and reliable explanations across diverse attribution methods.

To predict harmonized attributions of a model's layer l for a chosen attribution method, we first compute the feature maps $\boldsymbol{X}_{l,n} \in \mathbb{R}^{K_l \times H_l \times W_l}$ of this layer for all training samples, where n indexes the samples. Across the training dataset with N samples, this yields N feature maps that form a set of feature vectors, expressed as $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{N \times H_l \times W_l}\}, \boldsymbol{x}_i \in \mathbb{R}^{K_l}$ which define the learned feature space. In the feature space, similar feature vectors — and thus similar feature representations — are positioned close together, while dissimilar ones are spaced apart. This proximity enables us to average the attributions from nearby feature vectors, forming the foundation of our harmonization approach, illustrated in Figure 5.2.

Original Attributions

To find the relationship between feature maps and model predictions, we compute the original attribution maps for each feature map and class utilizing a desired

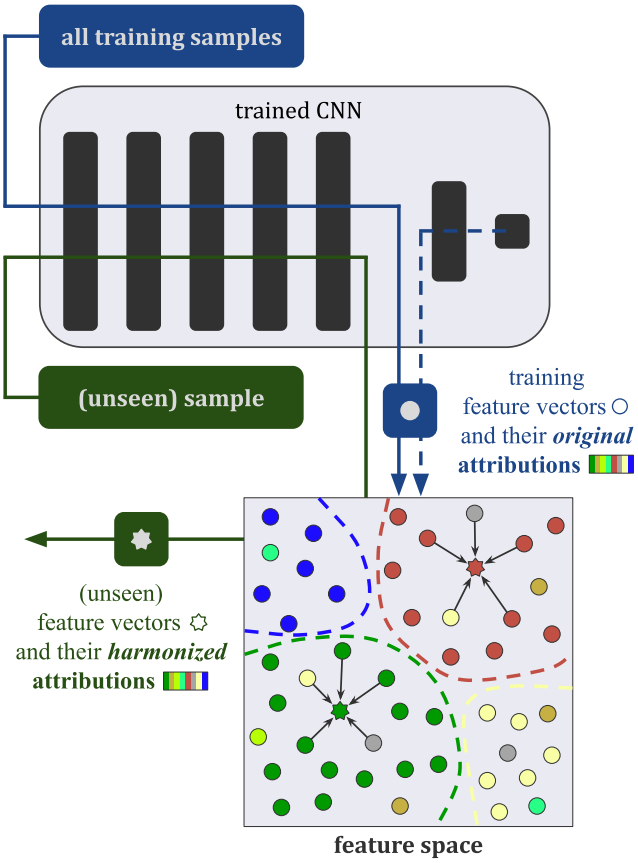


Figure 5.2: The feature vectors of all training samples' feature maps build the feature space. As similar feature vectors represent similar concepts, regions with comparable attributions emerge, as indicated by the dashed lines. The color represents the predominant class of an attribution vector (argmax). Using k-nearest neighbors, (unseen) feature vectors are assigned the average value of the surrounding attributions, resulting in more coherent and improved explanations.

attribution method. In this thesis, *original* attributions refer to the unmodified outputs of the attribution methods, whereas *harmonized* attributions are generated by our proposed approach. We denote the original attribution map for layer l and training sample n as $\boldsymbol{R}_{l,n} \in \mathbb{R}^{C \times H_l \times W_l}$ where C is the number of classes in the given task. As feature maps and attribution maps share identical spatial dimensions, each feature vector $\boldsymbol{x}_i$ can be assigned directly to an attribution, resulting in a set of training attributions $\{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_{N \times H_l \times W_l}\}, \boldsymbol{r}_i \in \mathbb{R}^C$.

Our procedure to compute original attribution maps diverges from conventional procedures in three ways:

1. Attributions are calculated for all classes, not just the predicted ones.

2. Negative attribution values are retained instead of being thresholded using ReLU as was originally done in methods like GradCAM.

3. Attributions are left unnormalized instead of transforming them to the $[0, 1]$ range, preserving their original state.

These choices ensure that the raw attribution values reflect the model's internal logic without introducing biases from post-processing.

Harmonized Attributions

To predict the harmonized attributions for the m-th (unseen) test sample, we first compute its feature map $\boldsymbol{X}'_{l,m}$ with feature vectors $\{\boldsymbol{x}'_1, \boldsymbol{x}'_2, ..., \boldsymbol{x}'_{H_l \times W_l}\}, \boldsymbol{x}'_i \in \mathbb{R}^{K_l}$. For each feature vector $\boldsymbol{x}'_i$, we identify the k nearest neighbors in the training feature space $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{N \times H_l \times W_l}\}$ and the corresponding attributions $\{\boldsymbol{r}_j, ...\}$ are averaged. Averaging the attributions of nearby training feature vectors provides relevant information about the feature representations in that region of the feature space leading to more meaningful attributions.

Original and Harmonized Attributions

Original attributions denote the unmodified, raw outputs of the attribution methods, while *harmonized* attributions are derived by our proposed methodology.

5.3 Feature-specific Attribution Methods

Building on the assumption that similar feature vectors represent similar concepts, we argue that existing attribution methods can be improved by making them more feature-specific.

Feature-specific Occlusions

Occluding features is a technique used to assess their sensitivity to the model's prediction. We propose specifying the occluded regions based on the feature map's feature vectors. Clustering them into groups of similar vectors can, for example, be achieved using k-means (Lloyd, 1982). Each cluster is then occluded separately and the attributions are divided by the number of occluded pixels. Our approach is illustrated in Figure 5.3, alongside the sliding window (Zeiler and Fergus, 2014) and random patches (Petsiuk et al., 2018) methods.

Sliding Window Occlusion | RISE Occlusion | k-means Occlusion (*ours*)

Figure 5.3: Sliding Window Occlusions by Zeiler and Fergus (2014) and RISE by Petsiuk et al. (2018) occlude the input using rectangular patches. In contrast, our Feature-specific Occlusion method simultaneously occludes similar feature vectors within the whole feature map. Although demonstrated here on an input image, our proposed method shows its true strength in the deep, intermediate layer of UH-Net, where the feature representations already correspond to meaningful concepts.

Feature-specific Grad-CAM

Grad-CAM computes the average gradient for each channel k of the feature map, denoted by $\boldsymbol{\alpha}^c = (\alpha_1^c, ..., \alpha_K^c)^{\mathrm{T}}$, and projects the feature vectors onto this direction. In Section 4.3 we discuss that this works well if only one direction in the feature space is relevant for the corresponding class. However, there might be more than one relevant direction, which cannot be captured by the existing Grad-CAM method. We therefore propose to replace the global average pooling of the gradients with a pooling that is based on clusters of the feature vectors. Similar to Feature-specific Occlusions, a clustering into groups of similar vectors can be performed using k-means (Lloyd, 1982). Thus, there is not a single $\boldsymbol{\alpha}^c$, but one for each group of feature vectors.

Our proposed method can be seen as an intermediate between Grad-CAM and the attribution method Gradients×Features. Compared to Gradients×Features, the only difference is that the gradients are averaged over similar feature vectors before being multiplied with them. Gradients×Features does not average the gradients at all and Grad-CAM averages the gradients globally.

Chapter 6

Explaining Land Cover Classification

In this section, we evaluate our proposed methodologies involving two well-known benchmark datasets for land cover classification, DFC2020 and Ben-ge; three CNN architectures, VGG-16, ResNet-18, and UH-Net; and ten attribution methods. The experiments are repeated ten times for each model and the results are averaged.

6.1 Experimental Setup

6.1.1 Datasets

Both benchmark datasets, DFC2020 by Schmitt et al. (2019b) and Ben-ge by Mommert et al. (2023), present unique challenges typical of remote sensing applications: DFC2020 includes imagery sampled from diverse geographic regions, meaning that models trained on this dataset must generalize across different climatic zones, vegetation types, and seasonal variability. Ben-ge, though detailed, suffers from label noise due to the reliance on automated classification products.

Image-Wise, Multi-Class Labels

Both remote sensing datasets include segmentation ground truth data. However, for our experiments, we define image-wise multi-labels $\boldsymbol{y} \in \{0,1\}^C$ for both datasets to train the CNNs. For this, we label a land cover class 1, if it covers more than 10% of the segmentation ground truth; and 0, if it covers less.

DFC2020

The DFC2020 dataset has been created for the 2020 IEEE GRSS Data Fusion Contest (Schmitt et al., 2019b) and land cover classes have been semiau-

tomatically annotated with a 10-meter resolution for 6114 patches with a size of 256 × 256 pixels each. For each patch, a Sentinel-1 and a Sentinel-2 image are provided, of which only the Sentinel-2 images are used for our experiments. They have 13 spectral bands, with values ranging from 0 to 10,000, and are sampled from seven globally distributed regions, specifically from Mexico, Germany, South Africa, Iran, India, Russia, and Australia. We divide the Sentinel-2 image values by 10,000 so that they range from 0 to 1. The land cover classification scheme is based on the International Geosphere-Biosphere Program (IGBP) scheme by Loveland and Belward (1997) and includes 10 classes in a simplified version, of which 8 are present in the dataset, as listed in Table 6.1. A sample is shown in Figure 6.1 (page 50).

Due to the task of the contest, the provided data split contains no training samples but 986 validation and 5128 test samples. Therefore, we conduct experiments using a random split among all samples, with 60 % for training, 20 % for validation, and 20 % for testing.

Table 6.1: DFC2020 land cover classes and their relative areas within the DFC2020 dataset.

DFC2020 LC Class	Rel. Area
Forest	23 %
Shrubland	6 %
Grassland	10 %
Wetland	5 %
Cropland	18 %
Urban/Built-up	10 %
Barren	3 %
Water	25 %

Ben-ge (BigEarthNet)

Ben-ge, introduced by Mommert et al. (2023), extends the BigEarthNet dataset by Sumbul et al. (2021) by adding geographical and environmental data. The dataset contains about 590,000 Sentinel-1 and Sentinel-2 images from Europe. The Sentinel-2 images have 12 spectral bands and a size of 120 × 120 pixels each, with values ranging from 0 to 10,000. We divide the image values by 10,000 so that they range from 0 to 1. We also reduce the dataset size by randomly selecting 20 % of the images. Using the provided data split, this results in approximately 54,000 training samples, and 25,000 samples each for validation and testing. Ben-ge provides ESA WorldCover (Zanaga et al., 2022) for each Sentinel scene which is a global land cover product with a resolution of 10 meters and 11 land cover classes. Three land cover classes are nearly not covered by BigEarthNet (< 0.01 %) and are ignored in our experiments. The remaining 8 classes and their frequencies are listed in Table 6.2. A sample is shown in Figure 6.1 (page 50).

Table 6.2: ESA WorldCover classes and their relative areas within the Ben-ge dataset.

ESA WorldCover Class (Ben-ge)	Rel. Area
Tree cover	41 %
Shrubland	1 %
Grassland	22 %
Cropland	15 %
Built-up	2 %
Bare/sparse vegetation	0.1 %
Permanent water bodies	17 %
Herbaceous wetland	0.5 %

6.1.2 Training

We train the CNNs end-to-end on the given datasets with image-wise multi-labels $\boldsymbol{y} \in \{0,1\}^C$ as described in Section 6.1.1. Each model uses sigmoid activation, binary cross-entropy loss, the AdamW optimizer, and a batch size of 32. The learning rate is linearly warmed up over 5 epochs and reduced by a factor of 10 if the validation loss stagnates for 5 epochs. The model state with the lowest validation loss is selected. The number of model parameters; the learning rates, weight decays, dropout probabilities, and epochs corresponding to the lowest validation loss; as well as the performances of the trained CNNs are listed in Table 6.3. Each architecture is trained 10 times with different initializations and randomizations. The results are averaged and standard deviations are indicated. Training is performed using PyTorch (Paszke et al., 2019) on an NVIDIA A100 40 GB PCIe GPU.

Table 6.3: The number of the trainable parameters (#Params) in the CNNs is on the order of millions (M). Learning rate (LR), weight decay (WD), and dropout (DO) are tuned to minimize the validation loss. The lowest validation loss is achieved after #Epochs. The test dataset metrics, including accuracy and F1-score, are reported as percentages (%).

Dataset	Model	#Params	LR	WD	DO	#Epochs	Accuracy	F1 (macro)
DFC2020	VGG-16	134.3 M	$1e-5$	$1e-4$	-	41 ± 7	95 ± 0	87 ± 0
	ResNet-18	11.2 M	$1e-4$	$1e-4$	-	47 ± 8	95 ± 0	87 ± 1
	UH-Net	19.0 M	$1e-4$	$1e-4$	0.3	35 ± 14	95 ± 0	83 ± 1
Ben-ge	VGG-16	134.3 M	$1e-6$	$1e-4$	-	57 ± 16	96 ± 0	77 ± 1
	ResNet-18	11.2 M	$1e-6$	$1e-4$	-	58 ± 9	96 ± 0	71 ± 1
	UH-Net	17.7 M	$1e-4$	$1e-4$	0.3	36 ± 7	97 ± 0	76 ± 1

Metrics: Accuracy, F1-score, Micro, and Macro

Accuracy is the proportion of correct predictions made by a model. The F1-score is defined as the harmonic mean of precision and recall. Precision measures how many of the predicted positives are actually correct, while recall measures how many of the actual positives are correctly identified. When a metric is described as *micro*, it is computed globally across all classes. In contrast, *macro* indicates that the metric is calculated independently for each class and then averaged. In this context, we report accuracy only globally (micro). For multi-class problems, the micro-averaged F1-score is equal to the micro-averaged accuracy; therefore, we report the F1-score as macro.

6.1.3 Original Attributions

We compute attributions utilizing ten attribution methods, namely:

- Gradients×Features (Shrikumar et al., 2017),
- Feature-specific Grad-CAM (proposed by us in Section 5.3),
- Grad-CAM (Selvaraju et al., 2020),
- Layer-wise Relevance Propagation (Bach et al., 2015),
- Integrated Gradients (Sundararajan et al., 2017),
- DeepLift (Shrikumar et al., 2017),
- Gradient SHAP (Lundberg and Lee, 2017),
- DeepLift SHAP (Lundberg and Lee, 2017).
- Sliding Window Occlusions (Zeiler and Fergus, 2014),
- Feature-specific Occlusions (proposed by us in Section 5.3),

We compute the attributions for most of the methods using the Captum library by Kokhlikyan et al. (2020). For Layer-wise Relevance Propagation, we apply the Epsilon Rule. For Integrated Gradients, we use zeros as the baseline and approximate the integral over 10 steps using the Gauss-Legendre method. For DeepLift, we use zeros as the baseline. For Gradient SHAP, we involve 100 random baseline samples, and each run is repeated five times with a noise of the standard deviation scaled by 0.01 added to the input sample. For DeepLift SHAP, we use 10 random baseline samples. For Sliding Window occlusions, we cover a total of $8 \times 8 = 64$ non-overlapping, equally-sized squares. Similarly, in Feature-specific Occlusions, we cover 64 clusters. For both occlusion methods, we set the occlusion value to zero. For Feature-specific Grad-CAM, we choose a number of 64 clusters as well. As constituted in Section 5.2, we do *not* threshold attributions to positive values using ReLU and we do *not* normalize attributions.

6.1.4 Harmonized Attributions

We perform the following steps independently for each model, dataset, layer of interest, and attribution method. First, we compute feature vectors and their attributions from the training dataset and select a representative random subset of approximately 100,000 training feature vectors for performance reasons. Next, we use the k-nearest neighbor regressor of the RAPIDS cuML Python library by Raschka et al. (2020) to find the $k = 100$ nearest neighbors of a vector. Since Euclidean distances become less meaningful in high-dimensional spaces (Aggarwal et al., 2001), we employ cosine similarity as our distance metric.

6.1.5 Architecture-Specific Details

VGG-16

We use a standard VGG-16 as introduced in Section 3.3.2, modified for the multi-channel input images and the number of targets, totaling 134.3 million parameters. VGG-16 has 13 convolutional layers from which we select the last convolutional layer with 512 channels for computing the attributions. For DFC2020 with an input size of 256×256 pixels, its resolution is 16×16 feature vectors; for Ben-ge (120×120), it is 7×7. The layer is followed by a Max Pooling operation with a kernel size of 2×2. For the odd Ben-ge feature maps (7×7), this causes all attribution methods — except Grad-CAM — to assign zero attributions to the right column and bottom row. Additionally, due to the low resolution of Ben-ge, Sliding Window and Feature-specific Occlusions occlude each feature vector independently, yielding identical attribution maps. Similarly, Gradients×Features and Feature-specific Grad-CAM are identical in this case.

ResNet-18

We use a standard ResNet-18 as introduced in Section 3.3.2, modified for the multi-channel input images and the number of targets, totaling 11.2 million parameters. ResNet-18 consists of four basic blocks. We select the last layer of the second block for computing the attributions due to its higher resolution compared to the layers in blocks three and four. The layer has a resolution of 32×32 for DFC2020 and 15×15 for Ben-ge, and 128 channels. We do not apply Layer-wise Relevance Propagation as originally no rules are defined for skip connections.

UH-Net

We use a UH-Net as proposed in Section 5.1. For Ben-ge, pooling is omitted in the first and third encoding steps due to the small image size. The U-Net's last layer is set to 16 channels with batch normalization but no activation function

and used for computing the attributions. For DFC2020, the full model has 19.0 million trainable parameters — 17.3 million in the U-Net and 1.7 million in the classification head. For Ben-ge, the model has 17.7 million parameters — 17.3 million in UH-Net and 0.4 million in the head. The difference in the number of head parameters is due to the two linear layers at the end, which have significantly more parameters for higher-resolution images.

6.1.6 Evaluation

Similarity between Attribution Methods

To compare two attribution methods, we calculate the Pearson correlation coefficient for their attributions across the test dataset. The Pearson correlation coefficient is the covariance of two random variables A and B normed by the product of their standard deviations:

$$\text{sim}_P = \frac{\text{cov}(A, B)}{\sigma_A \sigma_B} \tag{6.1}$$

In our case, the random variables A and B represent attribution values from two sets of attribution vectors $\{\boldsymbol{a}_1, ..., \boldsymbol{a}_n\}$ and $\{\boldsymbol{b}_1, ..., \boldsymbol{b}_n\}$. The vectors $\boldsymbol{a}_i$ correspond to one attribution method, while the vectors $\boldsymbol{b}_i$ correspond to the other. The Pearson correlation coefficient can have values between -1 (negative correlation) and 1 (positive correlation), where zero indicates no correlation.

Segmentation Ground Truth

We compare the predominant class of the attribution vectors with the segmentation ground truth of the test data. Assuming the model's predictions are rational, attribution methods that more closely align with the segmentation ground truth are more likely to provide meaningful explanations. We apply bilinear interpolation to the attribution maps if their sizes differ from the ground truth maps, which is required for VGG-16 and ResNet-18. As the segmentation classes of both datasets are imbalanced (see Tables 6.1, 6.2, pages 41 f.), we consider the F1-score in addition to the accuracy.

6.2 Results

This section presents the results obtained from the CNNs' deep layers. These include numerous figures, tables and matrices (pages 50 ff.) which are broadly summarized in the text to provide a high-level narrative. A detailed discussion of the results is presented in Section 8.

The following two boxes define our use of the term *feature-vector-wise attribution methods*, and list abbreviations often used in figures and tables, respectively.

Feature-vector-wise Attribution Methods

By feature-vector-wise attribution methods, we refer to methods that compute attributions independently for each individual feature vector: Gradients×Features, Layer-wise Relevance Propagation, Integrated Gradients, DeepLift, Gradient SHAP, and DeepLift SHAP.

The following methods are *not* considered feature-vector-wise: Feature-specific Grad-CAM, Grad-CAM, Sliding Window Occlusions, and Feature-specific Occlusions.

Abbreviations

Within visualizations and tables, we regularly abbreviate attribution methods as follows: Gradients×Features (Gr.×Ft.), Feature-specific Grad-CAM (Ft. Gr.-CAM), Grad-CAM (Gr.-CAM), Layer-wise Relevance Propagation (LRP), Integrated Gradients (Int. Gr.), Gradient SHAP (Gr. SHAP), DeepLift SHAP (DL SHAP), Sliding Window Occlusions (Sl. W. Occ.), Feature-specific Occlusions (Ft. Occ.). These are also listed in Section *Abbreviations* on page 165.

6.2.1 Visual Appearance

Figure 6.2 (page 50) compares the visualization of attribution maps as class-specific heatmaps — common in multi-class classification — with our visualizations of the predominant class attributions, which is more intuitive for mutli-label classification. Attribution maps for all architectures, datasets, and attribution methods are shown in Figures 6.3 and 6.4 (pages 51 f.) The visualized attribution maps are from a randomly selected model out of the ten ones trained per architecture.

Visualization of Attribution Maps

Given our multi-label tasks, we focus on visualizing the predominant class attributions (Figure 6.2b, page 50). All visualized attribution maps correspond to the dataset samples in Figure 6.1 (page 50), which are not repeatedly displayed. Similarly, the land cover legends provided in Tables 6.1 and 6.2 (pages 41 f.) are not reiterated at any point.

Original Attributions

Among the CNNs studied, the deep layer selected from VGG-16, and consequently its attribution maps, have the lowest spatial resolution, followed by those of ResNet-18, while UH-Net's attribution maps preserve the input resolution. The feature-vector-wise attribution methods produce noisy attribution maps, especially when applied to VGG-16 and ResNet-18. For UH-Net, cluster-like areas with reduced noise occur. Grad-CAM and Feature-specific Grad-CAM are less noisy as gradients are averaged across channels. Sliding Window Occlusions result in a low resolution which is particularly noticeable for UH-Net. In contrast, Feature-specific Occlusions is not limited in resolution and less noisy than gradient-based methods.

The extreme noise of feature-vector-wise methods for ResNet-18 stems from strided convolutions (3×3 kernel with 2×2 stride) which include every second feature vector twice. This creates a checkerboard pattern within gradient maps as shown in Figure 6.5 (page 53) and transfers to the gradient-based attribution maps.

Harmonized Attributions

Harmonization leads to greater similarity across attribution methods by mainly compensating for two aspects: It reduces attribution noise, particularly for feature-vector-wise methods; and it can adjust misleading explanations — significant attributions to classes with low prediction scores. For example, in Grad-CAM, Figure 6.4c, initial attributions to *Shrubland* are reassigned to *Grassland*, aligning with the model's predictions. For Sliding Window Occlusions, harmonization significantly improves the resolution, which is clarified in the next section, Section 6.2.2.

6.2.2 Feature Space

Figure 6.6 (page 53) compares the same feature space for two exemplary attribution methods: DeepLift and Grad-CAM. The original DeepLift method shows noisy attribution-feature alignment. Harmonization smoothens this mapping and aligns DeepLift's attributions closer to Grad-CAM, which already exhibits strong attribution-feature alignment.

For feature-vector-wise methods like DeepLift, the noisy alignment within the feature space is caused by noisy gradients. In contrast, Sliding Window Occlusions results in a noisy alignment due to the attribution maps' low resolution. Harmonization mitigates this, improving resolution within the image space for this attribution method.

6.2.3 Similarity between Attribution Methods

The Pearson correlation coefficients between attribution methods are illustrated in Figure 6.7 for DFC2020 and Figure 6.8 for Ben-ge (pages 54 f). Harmonization significantly increases the similarity between attribution methods, with only few exceptions. For VGG-16, we observe an average increase in Pearson correlation of 0.12 across both datasets and all attribution methods, resulting in an average correlation of 0.73 for the harmonized attributions. For ResNet-18, the increase is 0.17, yielding an average correlation of 0.44. For UH-Net, harmonization has the strongest effect, with an average increase of 0.27 and a resulting average correlation of 0.76. Across all architectures, Integrated Gradients and DeepLift show the lowest similarity to the other methods, even after harmonization.

For Ben-ge and VGG-16, Grad-CAM exhibits low similarity with the other methods because the low-resolution, odd-sized feature maps (7×7) are followed by a 2×2 Max Pooling operation. This causes all attribution methods — except Grad-CAM — to assign zero attributions to the rightmost column and bottom row of the attribution maps, affecting 27 % of the feature vectors.

6.2.4 Segmentation Ground Truth

The evaluation metrics comparing the predominant attributions with the segmentation ground truth are listed in Tables 6.4 and 6.5 (pages 56 f.) Through harmonization, all metrics improve (for all architectures and attribution methods), with a few exceptions where the metrics remain unchanged. UH-Net shows a particularly strong improvement. Assuming the model's predictions are rational, attribution methods that more closely align with the segmentation ground truth are more likely to provide meaningful explanations.

Among the original attribution methods, Grad-CAM performs best on VGG-16; for ResNet-18, it is DeepLIFT SHAP; and for UH-Net, the best methods are Feature-specific Grad-CAM, Grad-CAM, and Feature-specific Occlusions. After harmonization the best-performing methods on VGG-16 and UH-Net are: Gradients×Features, Feature-specific Grad-CAM, Grad-CAM, and Layer-wise Relevance Propagation. Additionally, DeepLIFT SHAP ranks among the top methods for VGG-16, while Sliding Window Occlusions and Feature-specific Occlusions are among the best on UH-Net. For ResNet-18, harmonized Gradient SHAP and DeepLIFT SHAP show the best performance.

6.2.5 Similarities between Original and Harmonized Attributions

The Pearson correlation coefficients between the original and harmonized attributions are presented in Figure 6.9. A high correlation indicates strong similarity, indicating that harmonization has little effect and that the attribution method is already aligned with the feature vectors.

VGG-16 exhibits the highest correlations for all methods due to a relatively simple relationship between the analyzed layer and the prediction. The layer is not followed by further convolutions, but Max Pooling, Global Average Pooling, and three fully connected linear layers, resulting in a more straightforward alignment between feature vectors and original attributions. Across all architectures, Grad-CAM exhibits the highest correlation, indicating the highest alignment with the feature vectors. This strong alignment is also evident by the feature space visualization in Figure 6.6. Feature-specific Grad-CAM and Feature-specific Occlusions exhibit high correlation, especially for higher-resolution feature maps as given for UH-Net (both datasets) as well as ResNet-18 (DFC2020).

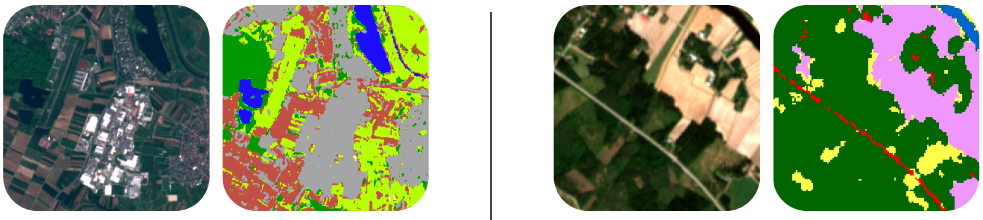
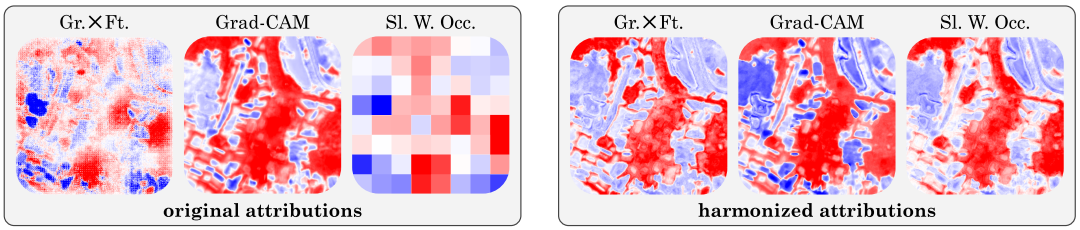
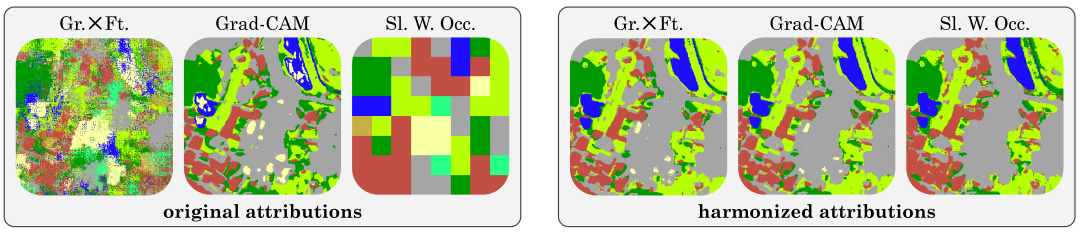


Figure 6.1: Sentinel-2 images and their corresponding segmentation ground truths from the DFC2020 (left) and Ben-ge (right) datasets. Color legends are shown in Tables 6.1 and 6.2 (pages 41 f.)

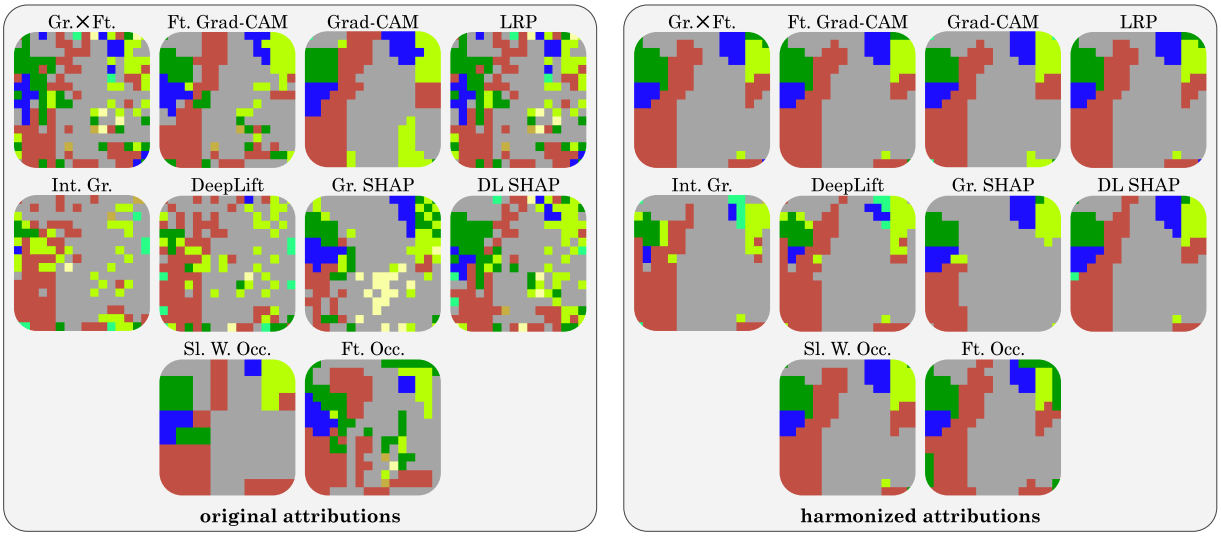


(a) Attribution maps for the ■ Urban/Built-up class with ■ negative and ■ positive attributions. Class-specific heatmaps are the common visualization of attribution maps for multi-class classification.

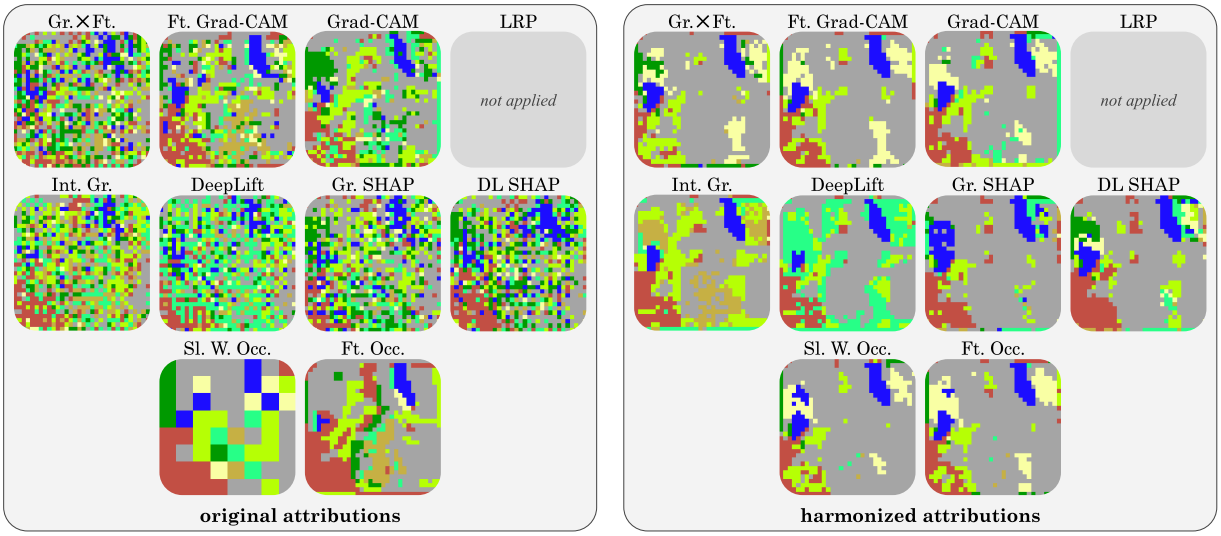


(b) Attribution maps showing the predominant classes of the attribution vectors (argmax). The land cover legend is provided in Table 6.1 (page 41). We favor this visualization for a multi-label task as it consolidates all class attributions into a single representation.

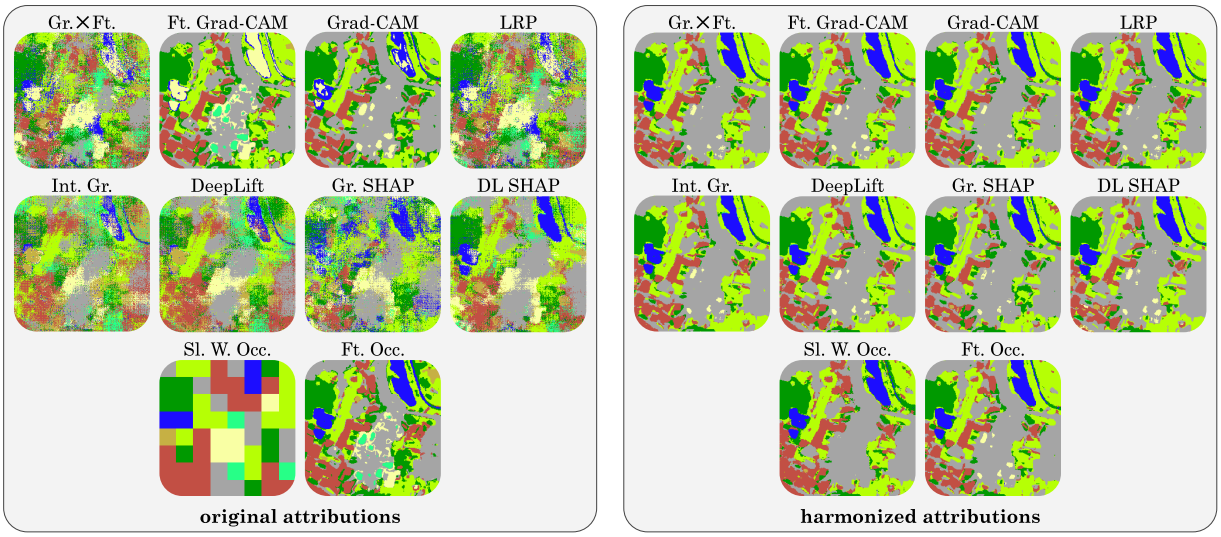
Figure 6.2: Original (left) and harmonized (right) attribution maps of the deep UH-Net layer for the sample from DFC2020 (Figure 10.1) for three attribution methods. The harmonized attributions show greater similarity across attribution methods than the original attributions.



(a) **VGG-16.**

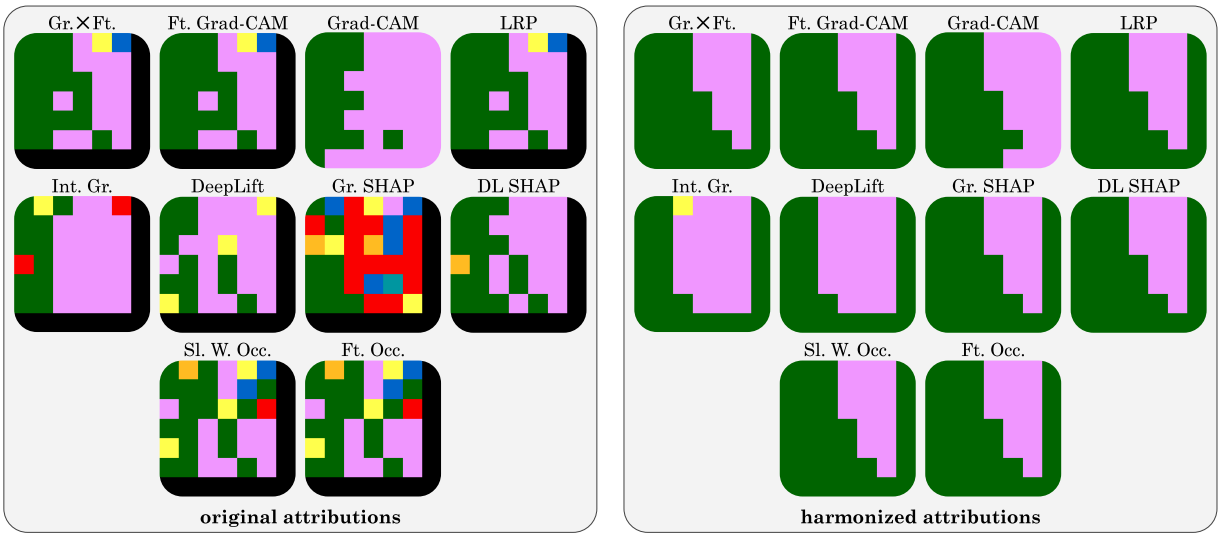


(b) **ResNet-18.** We do not apply Layer-wise Relevance Propagation as originally no rules are defined for skip connections.

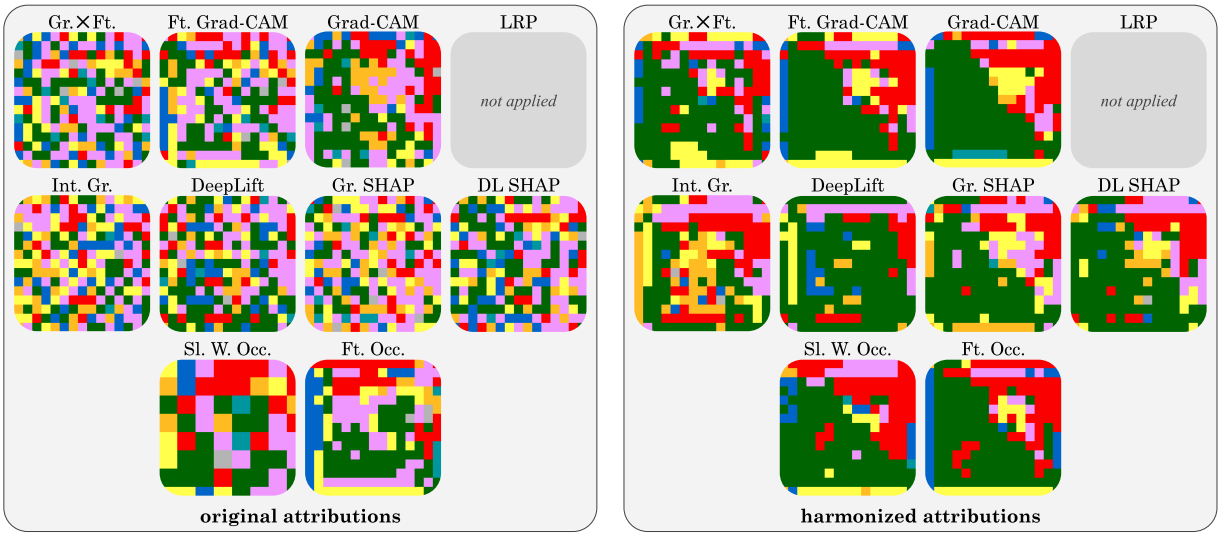


(c) **UH-Net.**

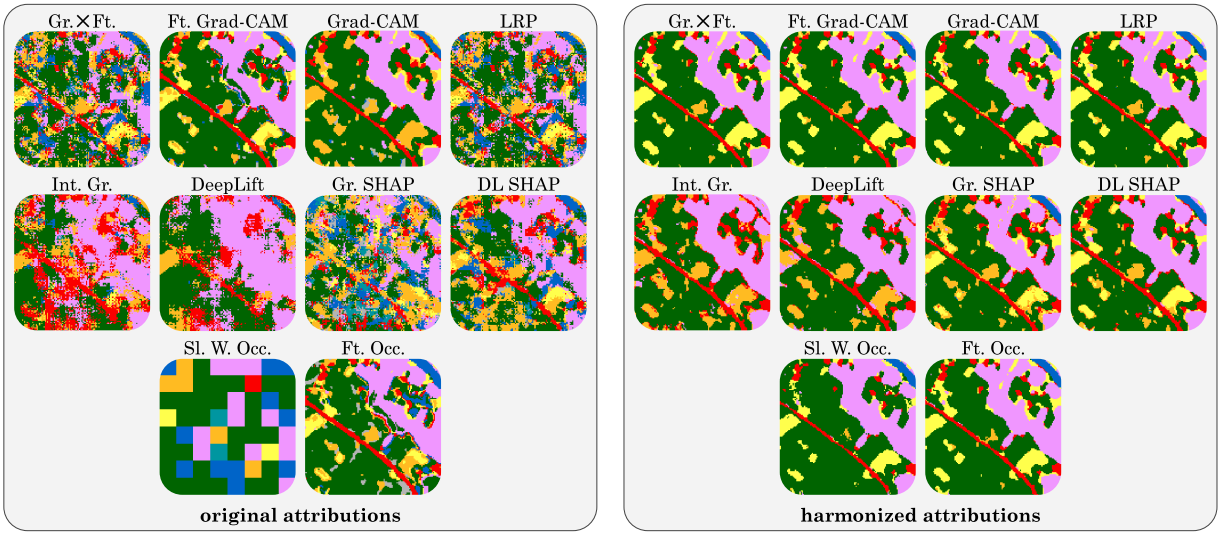
Figure 6.3: Original (left) and harmonized (right) attribution maps for a sample of **DFC2020**. Satellite image and ground truth are shown in Figure 10.1. The harmonized attributions show greater similarity across attribution methods than the original attributions. All models correctly predict the classes ■ Forest, ■ Grassland, ■ Cropland, and ■ Urban/Built-up as these classes occur with a relative area >10 %.



(a) **VGG-16.** The model correctly predicts the classes ■ Tree cover and ■ Cropland as these classes occur with a relative area >10 %. ■ Black indicates that all attributions are zero which occurs due to the odd feature map size of 7 × 7, followed by pooling with a 2 × 2 kernel.



(b) **ResNet-18.** The model predicts the classes ■ Tree cover, ■ Cropland, and ■ Grassland. Grassland is not included in the label as the relative area threshold is >10 %. We do not apply Layer-wise Relevance Propagation as originally no rules are defined for skip connections.



(c) **UH-Net.** The model predicts the classes ■ Tree cover, ■ Cropland, and ■ Grassland. Grassland is not included in the label as the relative area threshold is >10 %.

Figure 6.4: Original (left) and harmonized (right) attribution maps for a sample of **Ben-ge**. Satellite image and ground truth are shown in Figure 10.1. The harmonized attributions show greater similarity across attribution methods than the original attributions.

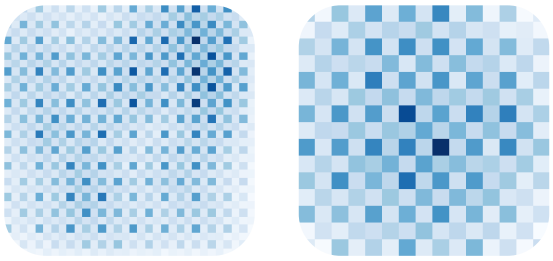


Figure 6.5: The absolute, summed gradient maps of the ResNet-18's deep feature vectors show checkerboard patterns also transferring to the gradient-based attribution maps. This is due to the strided convolutions (3×3 kernel with 2×2 stride) which include every second feature vector twice. (■ Low to ■ high gradients.)

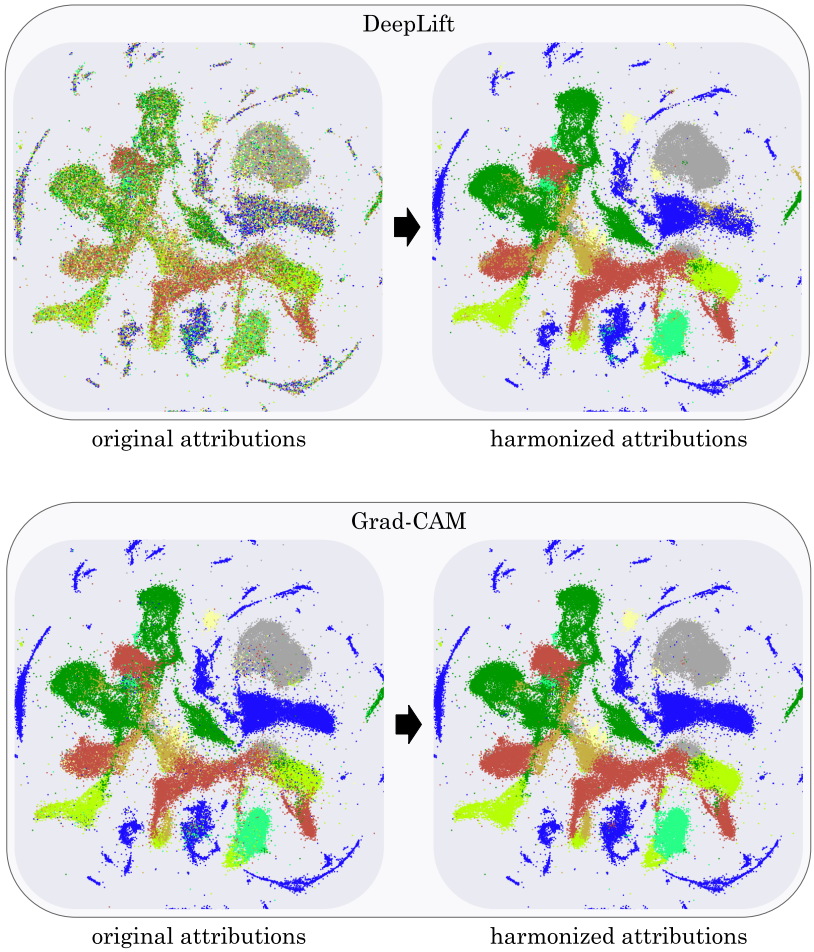


Figure 6.6: Feature space of the deep UH-Net layer for the DFC2020 training dataset and two exemplary attribution methods (DeepLift and Grad-CAM). The four illustrations differ only in the colorization of the feature vectors. On the left, colors indicate the predominant classes of the attribution vectors (argmax) within the *original* attributions. The *harmonized* attributions (right) result in a smoother alignment. DeepLift attributions are significantly smoothed through harmonization, whereas Grad-CAM attributions are already well-aligned prior to harmonization. For visualization purposes, the 16-dimensional feature vectors are reduced to two dimensions using Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2020; Raschka et al., 2020).

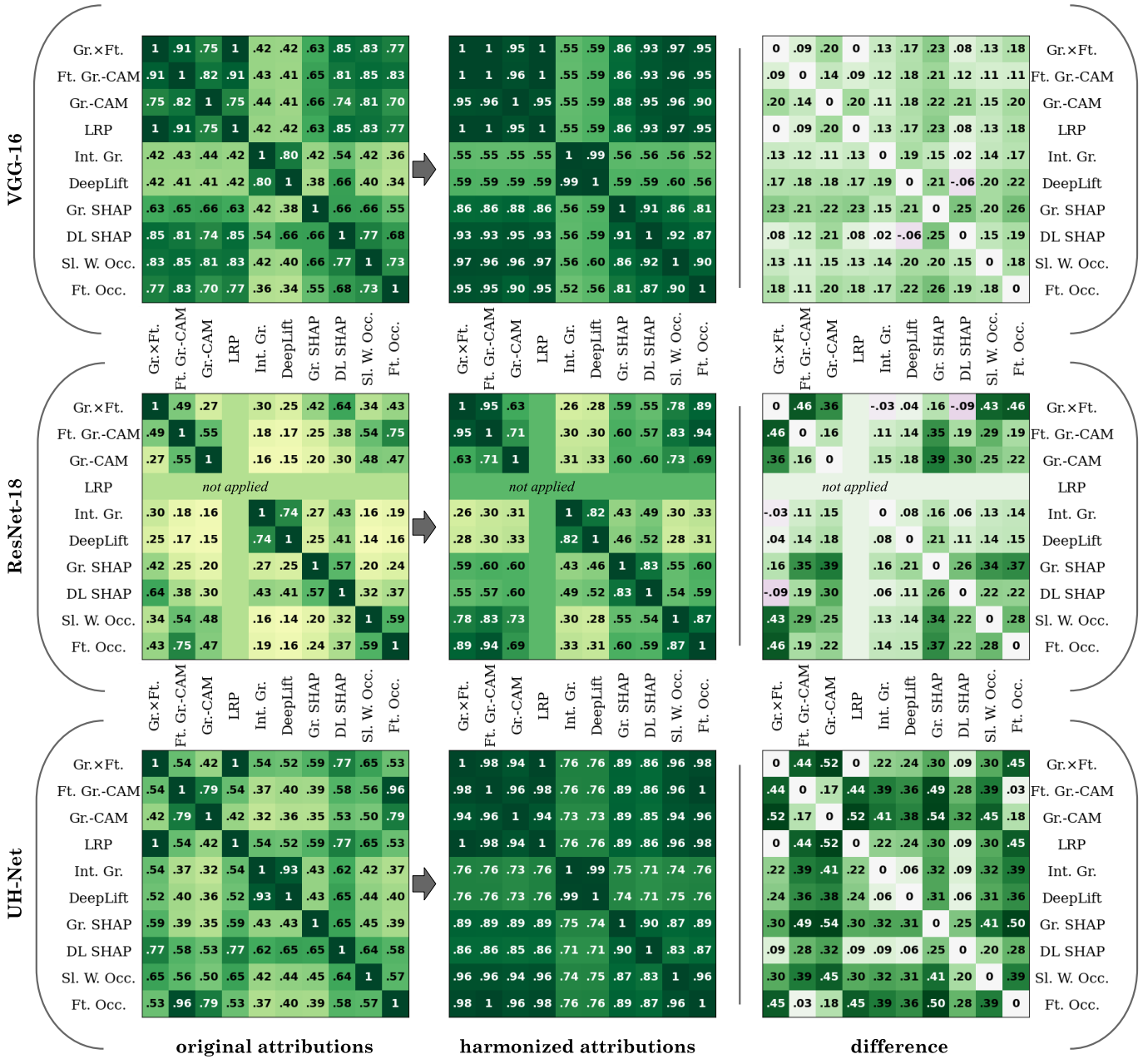


Figure 6.7: Pearson correlations between attribution methods for **DFC2020** and the three architectures: VGG-16 (top), ResNet-18 (middle), and UH-Net (bottom). The values are averaged over ten models. From left to right: coefficients for the original attributions, the harmonized attributions, and their difference. Layer-wise Relevance Propagation is not computed for the ResNet-18 as originally no rules are defined for skip connections.

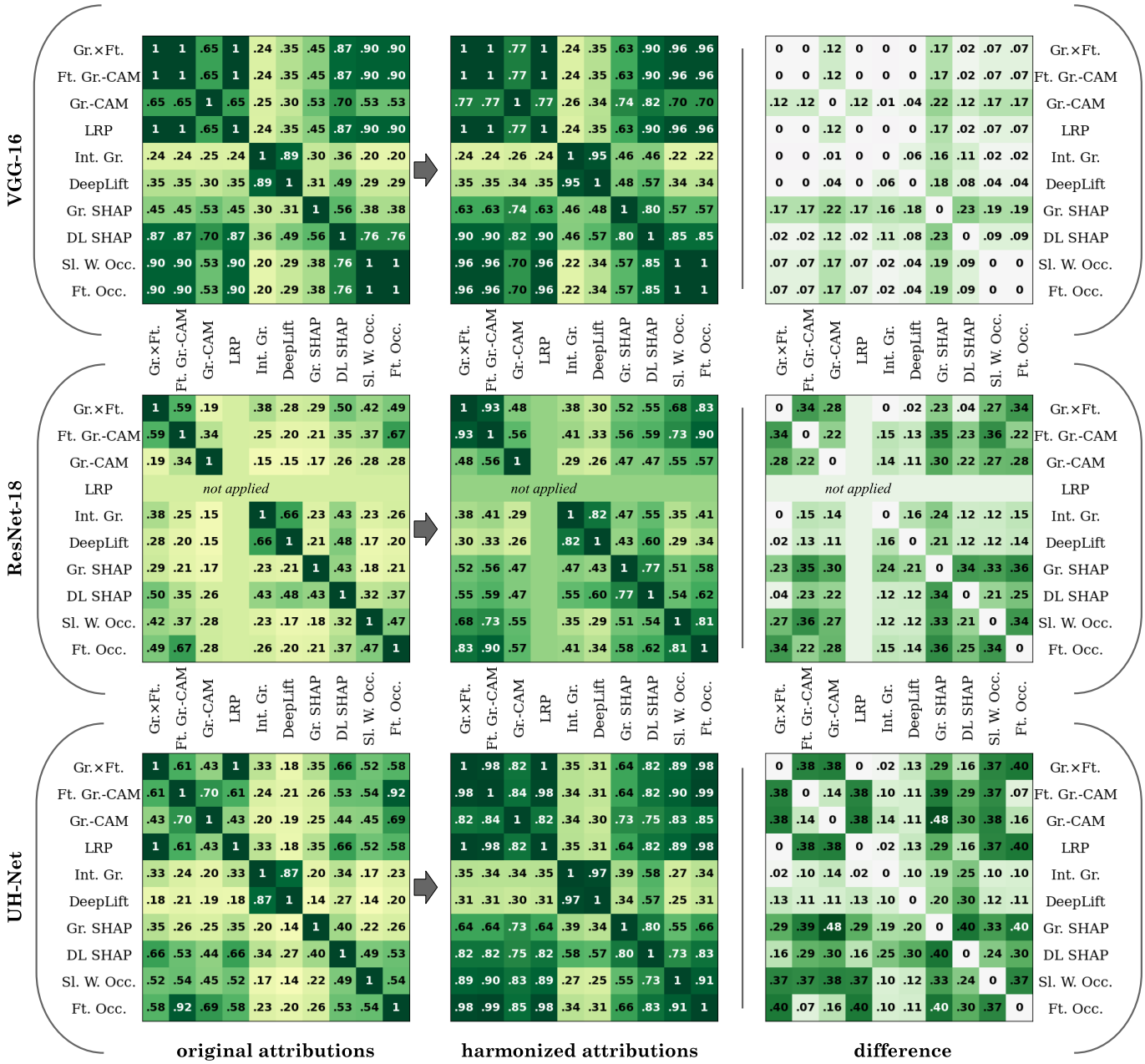


Figure 6.8: Pearson correlations between attribution methods for **Ben-ge** and the three architectures: VGG-16 (top), ResNet-18 (middle), and UH-Net (bottom). The values are averaged over ten models. From left to right: coefficients for the original attributions, the harmonized attributions, and their difference. Layer-wise Relevance Propagation is not computed for the ResNet-18 as originally no rules are defined for skip connections.

Table 6.4: Metrics comparing the predominant class attributions with the segmentation ground truth for **DFC2020**. Values are averaged over the ten models and reported as percentages, including the standard deviation. For clarity, the numbers are rounded to whole integers. The best values in a column (per CNN), along with those up to 2% lower, are highlighted in bold. Deviations in the differences may result from rounding.

CNN	Attribution Method	Accuracy			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.
VGG-16	Gradients×Features	69 ± 1	**75** ± 0	7 ± 1	57 ± 1	**65** ± 1	8 ± 1
	Ft. Grad-CAM	**73** ± 1	**76** ± 0	3 ± 1	62 ± 0	**65** ± 1	3 ± 1
	Grad-CAM	**75** ± 0	**75** ± 0	0 ± 0	**65** ± 0	**65** ± 0	0 ± 0
	LRP	69 ± 1	**75** ± 0	7 ± 1	57 ± 1	**65** ± 1	8 ± 1
	Integrated Gradients	45 ± 11	55 ± 12	10 ± 2	38 ± 8	47 ± 9	9 ± 2
	DeepLift	47 ± 12	57 ± 13	10 ± 2	40 ± 9	49 ± 10	9 ± 2
	Gradient SHAP	63 ± 1	71 ± 1	9 ± 1	51 ± 1	60 ± 1	9 ± 1
	DeepLift SHAP	71 ± 1	**74** ± 0	3 ± 1	61 ± 1	**65** ± 0	4 ± 1
	Sl. W. Occlusions	72 ± 1	**76** ± 0	4 ± 1	61 ± 1	**65** ± 1	4 ± 1
	Ft. Occlusions	68 ± 1	**74** ± 0	7 ± 1	56 ± 1	**63** ± 1	7 ± 1
ResNet-18	Gradients×Features	34 ± 4	46 ± 6	12 ± 3	27 ± 3	35 ± 5	7 ± 2
	Ft. Grad-CAM	**41** ± 5	46 ± 6	5 ± 3	33 ± 4	35 ± 5	2 ± 2
	Grad-CAM	39 ± 7	42 ± 8	3 ± 2	34 ± 5	35 ± 6	2 ± 1
	Integrated Gradients	32 ± 3	37 ± 6	5 ± 3	28 ± 3	33 ± 5	4 ± 3
	DeepLift	34 ± 4	42 ± 6	8 ± 3	30 ± 3	37 ± 5	7 ± 4
	Gradient SHAP	31 ± 3	**51** ± 5	21 ± 4	27 ± 3	**45** ± 3	18 ± 3
	DeepLift SHAP	**43** ± 5	**51** ± 4	8 ± 4	**38** ± 5	**46** ± 3	8 ± 3
	Sl. W. Occlusions	40 ± 5	43 ± 6	3 ± 2	33 ± 3	34 ± 4	1 ± 1
	Ft. Occlusions	39 ± 4	41 ± 5	2 ± 2	33 ± 3	33 ± 4	1 ± 1
UH-Net	Gradients×Features	42 ± 3	**80** ± 0	37 ± 3	35 ± 2	**69** ± 1	33 ± 2
	Ft. Grad-CAM	**77** ± 1	**80** ± 0	4 ± 1	**65** ± 2	**69** ± 1	4 ± 1
	Grad-CAM	**78** ± 1	**80** ± 1	2 ± 1	**67** ± 2	**69** ± 1	2 ± 1
	LRP	42 ± 3	**80** ± 0	37 ± 3	35 ± 2	**69** ± 1	33 ± 2
	Integrated Gradients	35 ± 6	72 ± 7	37 ± 4	30 ± 4	62 ± 5	32 ± 3
	DeepLift	40 ± 6	73 ± 7	33 ± 4	34 ± 5	63 ± 5	29 ± 3
	Gradient SHAP	36 ± 2	77 ± 1	42 ± 2	30 ± 2	66 ± 1	36 ± 2
	DeepLift SHAP	56 ± 2	73 ± 2	17 ± 2	46 ± 2	64 ± 1	18 ± 2
	Sl. W. Occlusions	50 ± 3	**81** ± 0	30 ± 3	42 ± 3	**69** ± 1	27 ± 3
	Ft. Occlusions	**77** ± 1	**80** ± 0	4 ± 1	**65** ± 2	**69** ± 1	4 ± 1

Table 6.5: Metrics comparing the predominant class attributions with the segmentation ground truth for **Ben-ge**. Values are averaged over the ten models and reported as percentages, including the standard deviation. For clarity, the numbers are rounded to whole integers. The best values in a column (per CNN), along with those up to 2% lower, are highlighted in bold. Deviations in the differences may result from rounding.

CNN	Attribution Method	Accuracy			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.
VGG-16	Gradients×Features	**63** ± 1	**69** ± 0	6 ± 1	40 ± 1	**43** ± 0	3 ± 1
	Ft. Grad-CAM	**63** ± 1	**69** ± 0	6 ± 1	40 ± 1	**43** ± 0	3 ± 1
	Grad-CAM	**63** ± 2	**69** ± 1	5 ± 1	**43** ± 1	**45** ± 1	2 ± 1
	LRP	**63** ± 1	**69** ± 0	6 ± 1	40 ± 1	**43** ± 0	3 ± 1
	Integrated Gradients	48 ± 3	54 ± 3	5 ± 1	27 ± 3	29 ± 4	2 ± 1
	DeepLift	49 ± 3	55 ± 3	6 ± 1	27 ± 2	30 ± 3	3 ± 1
	Gradient SHAP	51 ± 2	64 ± 1	13 ± 1	32 ± 1	41 ± 1	10 ± 1
	DeepLift SHAP	**62** ± 1	**68** ± 0	6 ± 1	**41** ± 1	**45** ± 0	4 ± 1
	Sl. W. Occlusions	53 ± 1	65 ± 1	12 ± 1	32 ± 1	40 ± 1	8 ± 1
	Ft. Occlusions	53 ± 1	65 ± 1	12 ± 1	32 ± 1	40 ± 1	8 ± 1
ResNet-18	Gradients×Features	21 ± 3	26 ± 7	5 ± 5	13 ± 1	15 ± 3	2 ± 2
	Ft. Grad-CAM	25 ± 4	30 ± 8	4 ± 4	15 ± 2	17 ± 4	2 ± 2
	Grad-CAM	**31** ± 7	33 ± 9	2 ± 3	**20** ± 4	20 ± 5	0 ± 1
	Integrated Gradients	26 ± 2	34 ± 7	8 ± 5	16 ± 1	19 ± 3	3 ± 2
	DeepLift	**30** ± 2	37 ± 5	7 ± 4	16 ± 1	19 ± 3	3 ± 2
	Gradient SHAP	26 ± 2	**44** ± 8	18 ± 6	17 ± 1	**27** ± 4	10 ± 3
	DeepLift SHAP	**31** ± 3	**43** ± 7	12 ± 5	**20** ± 2	**26** ± 4	6 ± 2
	Sl. W. Occlusions	27 ± 4	32 ± 8	5 ± 4	17 ± 2	19 ± 4	2 ± 2
	Ft. Occlusions	27 ± 5	30 ± 7	3 ± 3	17 ± 3	18 ± 4	1 ± 1
UH-Net	Gradients×Features	55 ± 4	**78** ± 1	23 ± 3	32 ± 2	**53** ± 1	21 ± 2
	Ft. Grad-CAM	**73** ± 2	**79** ± 1	5 ± 1	**46** ± 1	**54** ± 1	8 ± 1
	Grad-CAM	**74** ± 1	**79** ± 1	5 ± 1	**47** ± 1	**54** ± 1	7 ± 0
	LRP	55 ± 4	**78** ± 1	23 ± 3	32 ± 2	**53** ± 1	21 ± 2
	Integrated Gradients	38 ± 4	46 ± 9	7 ± 6	26 ± 2	28 ± 5	2 ± 3
	DeepLift	40 ± 5	44 ± 12	4 ± 7	26 ± 3	26 ± 6	0 ± 4
	Gradient SHAP	30 ± 3	67 ± 3	37 ± 2	20 ± 2	45 ± 2	25 ± 1
	DeepLift SHAP	56 ± 2	72 ± 1	16 ± 2	38 ± 1	49 ± 1	11 ± 1
	Sl. W. Occlusions	60 ± 6	**80** ± 2	20 ± 5	36 ± 4	**55** ± 1	19 ± 4
	Ft. Occlusions	**73** ± 2	**79** ± 1	6 ± 2	**46** ± 2	**54** ± 1	8 ± 1

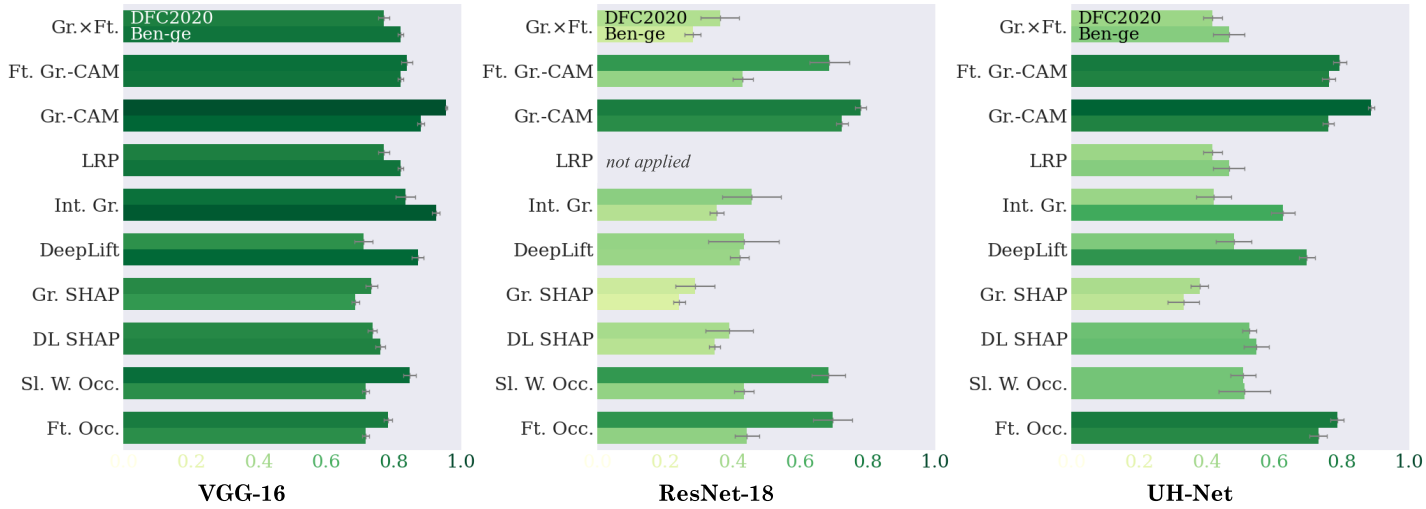


Figure 6.9: Similarities between original and harmonized attributions for the three architectures (left to right) and all attribution methods (top to bottom within each plot). For each attribution method, the upper bar represents the Pearson correlation for DFC2020; the lower bar the one for Ben-ge. The color intensity highlights the bar values, with darker shades indicating higher ones. The values are averaged across ten models, and grey lines represent the standard deviations. We do not apply Layer-wise Relevance Propagation for the ResNet-18 as originally no rules are defined for skip connection.

Chapter 7

Additional Experiments

In this section, we address additional questions related to the methods introduced in Section 5. While the previous section, Section 6, emphasized completeness — covering various CNN architectures, datasets, attribution methods, and averaging results over multiple training runs — this section focuses on individual aspects, often with a particular emphasis on our UH-Net architecture and the DFC2020 dataset.

7.1 Varying Harmonization Parameters

We evaluate the robustness of the parameter k in the k-nearest neighbors search applied when attributions are harmonized utilizing the training feature space. To do so, we select one of the ten UH-Net models trained on DFC2020 and repeat the harmonization for $k = 5$, 20, 100, 300. Some resulting attribution maps are shown in Figure 7.1. The Pearson correlations among the attribution methods can be found in Appendix A.1, Figure A.1 (page 137).

For attribution methods that closely align with the features (e.g., Grad-CAM), the choice of k has little to no effect. However, for other attribution methods, a very low k limits the harmonization effect, as averaging over a small number of neighbors does not fully eliminate noise within the feature space. Regarding the similarity between harmonized attribution methods, $k = 5$ already has a positive impact. However, there is a significant increase in similarity at $k = 20$. We observe a further, though less pronounced, increase at $k = 100$, while at $k = 300$, the improvement is minimal. We make similar conclusions when observing the accuracies and F1-scores for different k's. Overall, our findings suggest that the choice of k is quite robust. However, it should exceed a certain threshold, which we find to be around $k = 20$. For our experiments we have chosen $k = 100$ as a good compromise between performance and computational efficiency.

Another choice we made was to use cosine similarity as the distance metric

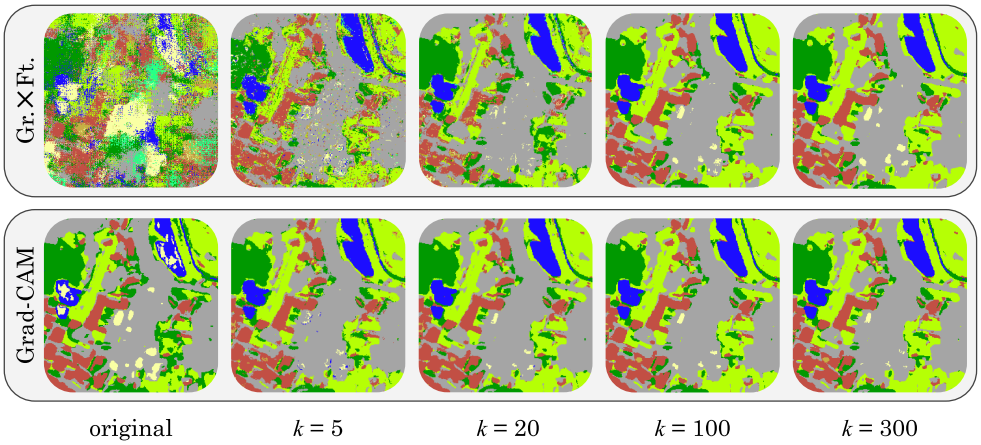


Figure 7.1: Harmonized attribution maps for DFC2020 and one of the UH-Net models (deep layer) for varying numbers of nearest neighbors. The parameter is quite robust for values above $k = 20$ and we have chosen $k = 100$ for our experiments.

within the feature space when searching for nearest neighbors. However, we find that the choice between cosine similarity and Euclidean distance has no significant effect for all tested architectures.

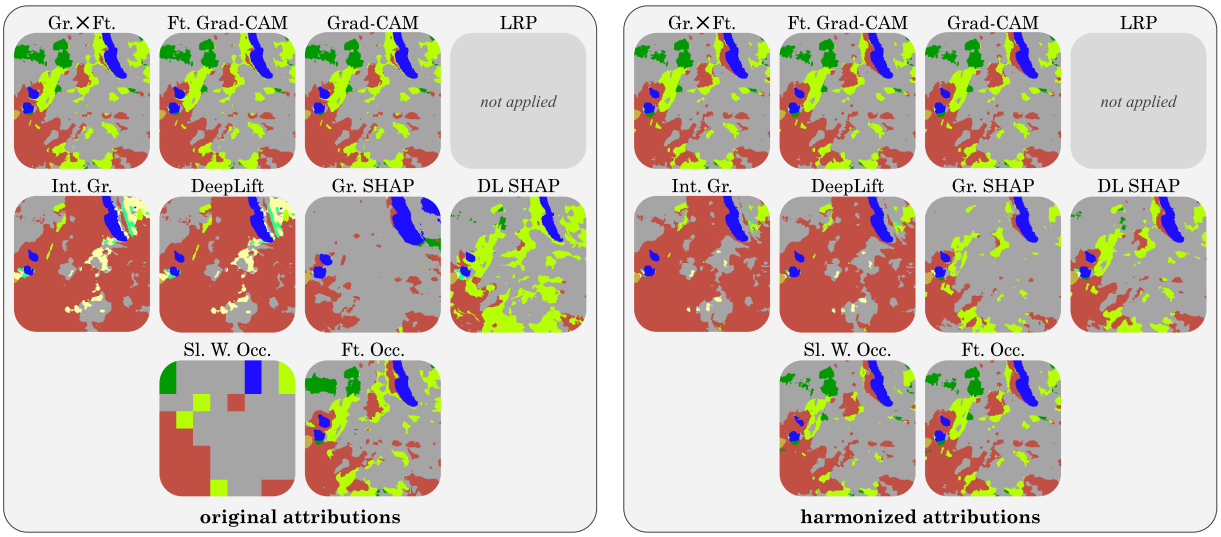
7.2 Variants of UH-Net

Our proposed UH-Net architecture, as described in Section 5.1, employs a shallow classifier head to ensure that the feature vectors at the intermediate layer capture high-level concepts. In this section we investigate the attributions for two modified versions of this architecture: A UH-Net with an even simpler classifier head, and a UH-Net with a more complex classifier head. For both architectures, ten models are trained and the results are averaged. Attribution maps for one of the ten models are visualized in Figure A.2. The Pearson correlations and evaluation metrics can be found in Appendix A.1, Figure A.2 (page 138); and Table A.1 (page 139).

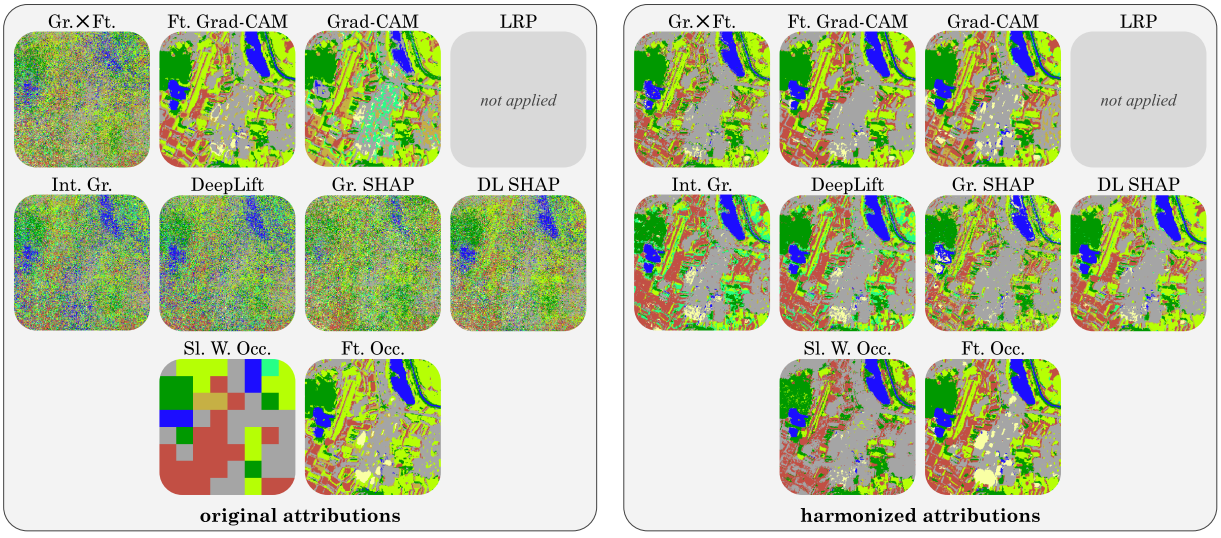
UH-Net with a Simple Head

As an example of a simple head, we use a final layer as proposed by Zhou et al. (2016) to enable Class Activation Mapping (CAM). CAM is an attribution method applicable to specific CNN architectures and introduced in Section 3.4.2. Our layer of interest has 16 channels, as in our proposed UH-Net, but it is not batch-normalized. The head performs global average pooling (channel-wise mean) followed by a single linear layer with weights only (no bias). Thus, the head contains only 128 trainable parameters, while the U-Net still has 17.3 million.

With this architecture, CAM, Grad-CAM, Feature-specific Grad-CAM, and Gradients×Features yield identical results. These are also similar to both occlusion-



(a) **UH-Net with a simple head**. For most methods, harmonization has minimal impact.



(b) **UH-Net with a complex ResNet-18 head**. The more complex head causes noisier gradient-based attribution maps than our proposed UH-Net described in Section 5.1; however, harmonization can largely mitigate this effect.

Figure 7.2: Original (left) and harmonized (right) attribution maps for a sample of DFC2020 and two **variants of UH-Net** (deep layer).

based and both SHAP methods. Integrated Gradients and DeepLift produce identical results to each other but differ from the other methods. The original attributions align better with the segmentation ground truth than those of our proposed UH-Net. However, harmonization has only a small effect and the harmonized attributions align less well.

UH-Net with a Complex Head

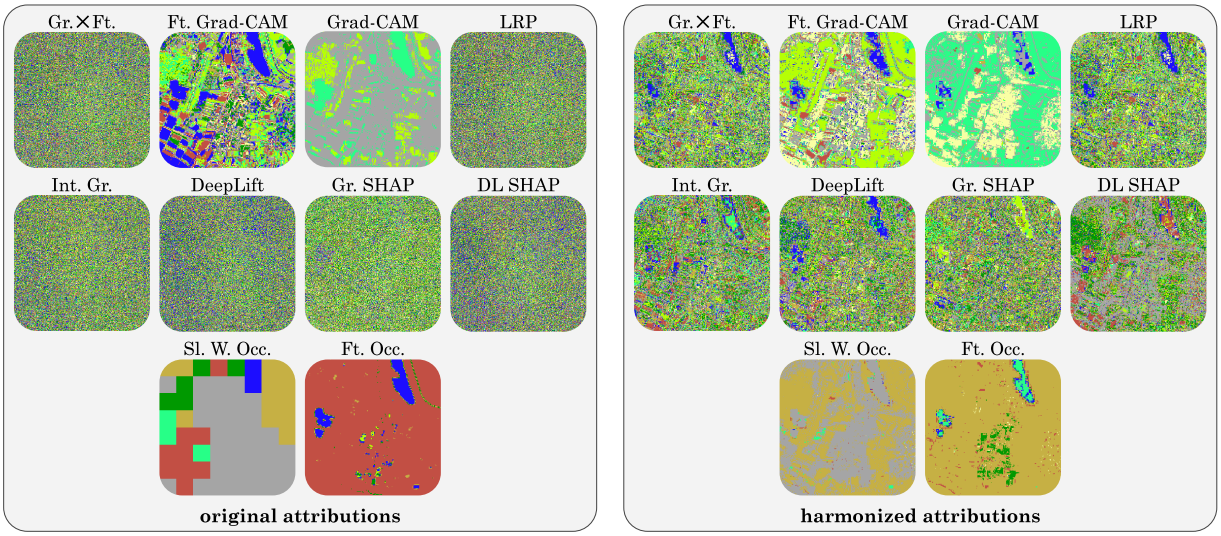
To demonstrate a UH-Net with a complex head, we append a ResNet-18 after the U-Net. The intersection layer has 16 channels, as in our proposed UH-Net. In this configuration the U-Net has 17.3 million trainable parameters, while the ResNet-18 head has 11.2 million. In the case of the complex head, gradient-based attribution maps of the intersection layer are significantly noisier compared to our proposed UH-Net architecture. This noise results in low Pearson correlations, although the attribution maps appear visually similar. Harmonization reduces the noise and significantly improves the Pearson correlations. The harmonized attributions of UH-Net with a ResNet-18 head align less with the segmentation ground truth compared to those of our proposed UH-Net architecture.

7.3 Attributions for the Input Layers

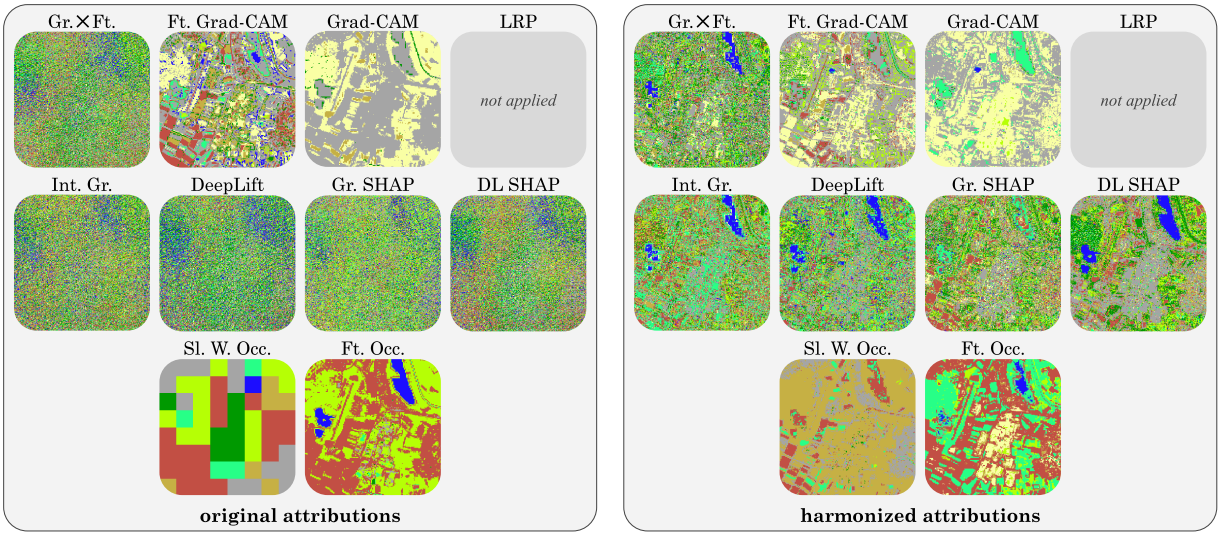
In Section 6.2, we focus on the results of the deep CNN layers. Here, we present the attributions for the *input* layers of the same ten CNNs per architecture for DFC2020. Layer-wise Relevance Propagation is not computed for ResNet-18 and UH-Net, as originally no rules are defined for skip connections.

Original and harmonized attribution maps for the inputs are shown in Figure 7.3. The feature-vector-wise attribution maps are highly noisy for all architectures, and harmonization can compensate for this only to a limited extent. UH-Net exhibits the lowest noise, likely due to the skip connections in the U-Net architecture, which allow gradients to backpropagate more directly to the input. The attributions show low correlation across different methods for all architectures (Figure A.3 in Appendix A.1, page 140). While harmonization increases their similarity, the effect is usually minor. The most notable improvements are observed with UH-Net. Comparing the predominant attributions with the segmentation ground truth, some metrics show a significant upward trend after harmonization, with the highest impact observed for UH-Net (Table A.2 in Appendix A.1, page 141). However, there are also cases where harmonization leads to a decline in metrics. Harmonized DeepLift SHAP yields the best performance across all architectures.

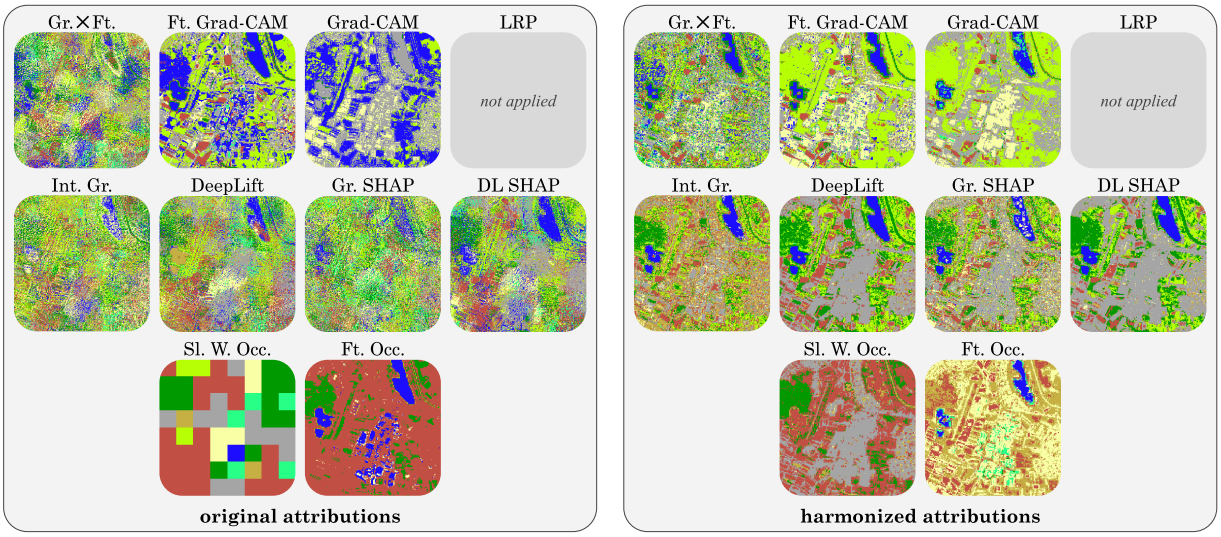
Correlations between original and harmonized attributions are present for Grad-CAM, Feature-specific Grad-CAM, Sliding Window Occlusions, and Feature-



(a) **VGG-16.**



(b) **ResNet-18.**



(c) **UH-Net.**

Figure 7.3: Original (left) and harmonized (right) **input** attribution maps for DFC2020. Harmonization frequently falls short of producing a result that seems any more meaningful than the original attributions especially for the occlusion-based and both Grad-CAM methods. All models correctly predict the classes ■ Forest, ■ Grassland, ■ Cropland, and ■ Urban/Built-up.

specific Occlusions, but hardly for the other methods (Figure A.4 in Appendix A.1, page 142). The likely reason is the high level of noise for the feature-vector-wise methods. For UH-Net, noise is generally lower due to the presence of skip connections. It thus exhibits higher similarity values also for the gradient-based methods.

7.4 Object Classification

Caltech 101 Dataset

In this section, we evaluate our methodologies using the Caltech 101 dataset by Li et al. (2022), which contains illustrations and photos taken with a handheld camera. It is a multi-class dataset with 101 categories, including objects, faces, and animals. Although the dataset is imbalanced, most categories have around 50 images, each approximately 300 × 200 pixels in size, with either red-green-blue (RGB) channels or a single gray channel. In total, the dataset contains 9,146 images.

We resize the images to 256 × 256 pixels, replicate the channels of single-channel images to obtain three channels, and normalize each image to values between 0 and 1. We apply a random data split across all samples: 60% for training, 20% for validation, and 20% for testing.

Experimental Setup

We use the same three model architectures as for DFC2020 (VGG-16, ResNet-18, and UH-Net), adjusting only the number of input channels to 3. For UH-Net, we further choose the intermediate layer to have 32 channels instead of 16, as the number of classes is significantly larger. As Caltech 101 provides a multi-class task, we use softmax activation and cross-entropy loss. We employ the AdamW optimizer with a batch size of 32. The learning rate is linearly warmed up over 5 epochs and reduced by a factor of 10 if the validation loss stagnates for 20 epochs. Other hyperparameters are listed in Table 7.1. We select the model state with the lowest validation loss. Other than for the land cover classification tasks, UH-Net has a significantly lower test accuracy and F1-score (macro) than VGG-16 and ResNet-18.

Due to the high number of classes and the multi-class nature of this task, we do not compute attributions for all classes in each sample, as described in Section 5.2. Instead we compute the attributions of the predicted class only. To harmonize attributions, we set up a separate feature space for each class. Each feature space includes the feature vectors of the training samples that were predicted as belonging to the corresponding class. For an unseen test sample, we

Table 7.1: Learning rate (LR), weight decay (WD), and Dropout (DO) are tuned to minimize the validation loss. The lowest validation loss is achieved after #Epochs. The test dataset metrics, including accuracy and F1-score, are reported as percentages (%).

Dataset	Model	LR	WD	DO	#Epochs	Accuracy	F1 (macro)
Caltech 101	VGG-16	$1e-3$	$1e-2$	-	43	76	62
	ResNet-18	$1e-3$	$1e-1$	-	34	78	66
	UH-Net	$1e-4$	$1e-1$	0.7	61	64	33

refer to the feature space of its predicted class.

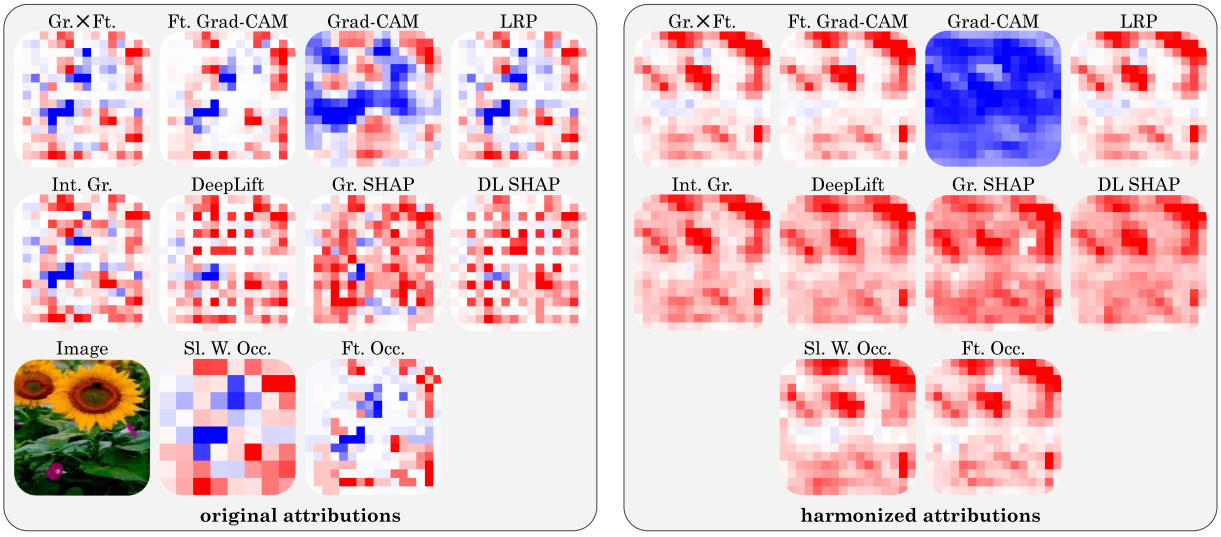
Results

Figure 7.4 shows the attribution heatmaps of a sample, that was correctly predicted as a sunflower by all models, while Figure 7.5 presents the Pearson correlations computed over the entire test dataset. Similar to the multi-label land cover classification, harmonization leads to an increase in similarity across all models and nearly all attribution methods. A notable exception is Grad-CAM with VGG-16. We make the following observations:

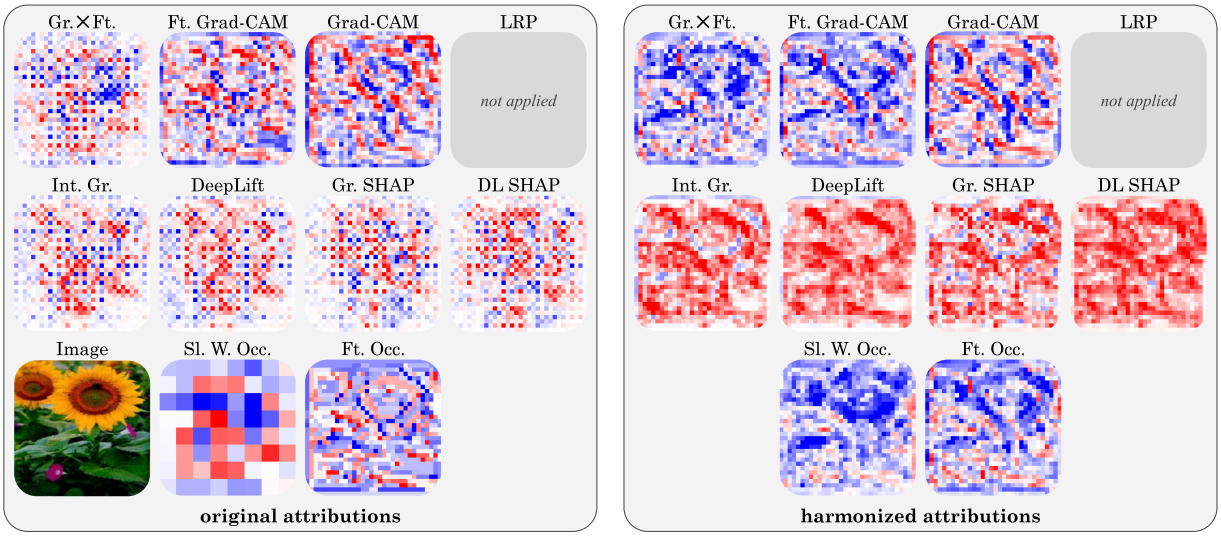
- **Positive vs. negative values in harmonized attributions:** For VGG-16, the harmonized attributions are mostly positive across all attribution methods, except for Grad-CAM. For ResNet-18, Integrated Gradients, DeepLift, Gradient SHAP, and DeepLift SHAP they are predominantly positive. The other methods yield a mix of negative and positive harmonized attributions. A similar trend is also present for UH-Net, though less pronounced.

- **Object vs. background attributions:** UH-Net often focuses on specific parts of the object, especially if they contrast with the background. In our example, it focuses on the yellow petals of the sunflower. VGG-16 and ResNet-18 tend to include more of the background in their attributions. Attribution maps of the ResNet-18 can look abstract.

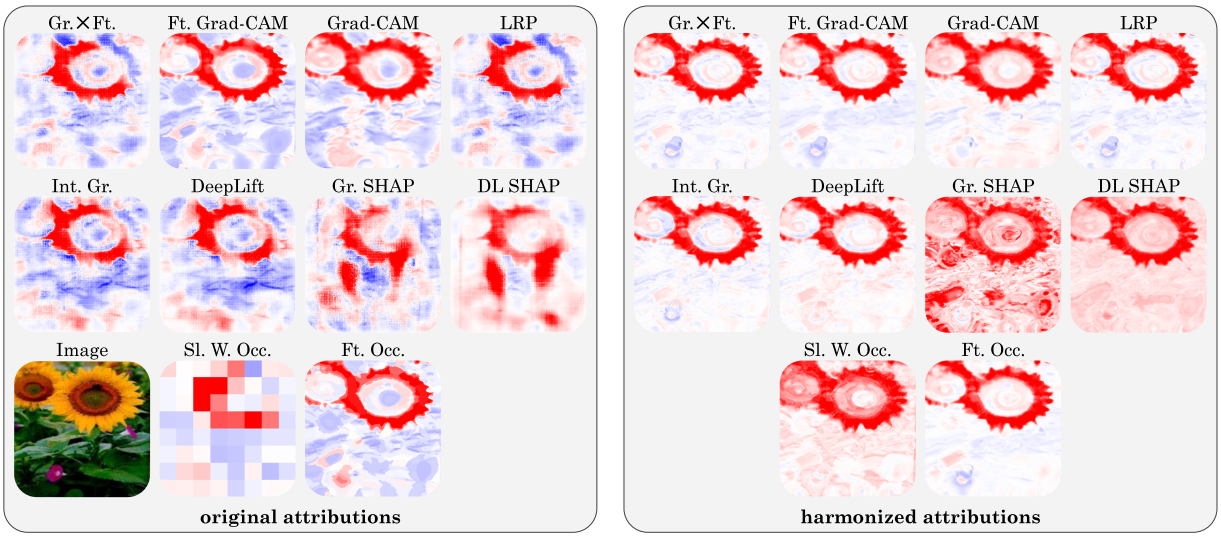
- **Grad-CAM:** Grad-CAM produces results that are relatively dissimilar from those of the other attribution methods across all architectures. This effect is most pronounced for VGG-16, where the original Grad-CAM attributions often exhibit strong negative values. Thus, after harmonization, most attributions become negative.



(a) **VGG-16.**



(b) **ResNet-18.** We do not apply Layer-wise Relevance Propagation as originally no rules are defined for skip connections.



(c) **UH-Net.**

Figure 7.4: Original (left) and harmonized (right) attribution maps for a sample of the sunflower class of **Caltech 101** with ■ negative and ■ positive attributions. The sample was correctly predicted by all models.

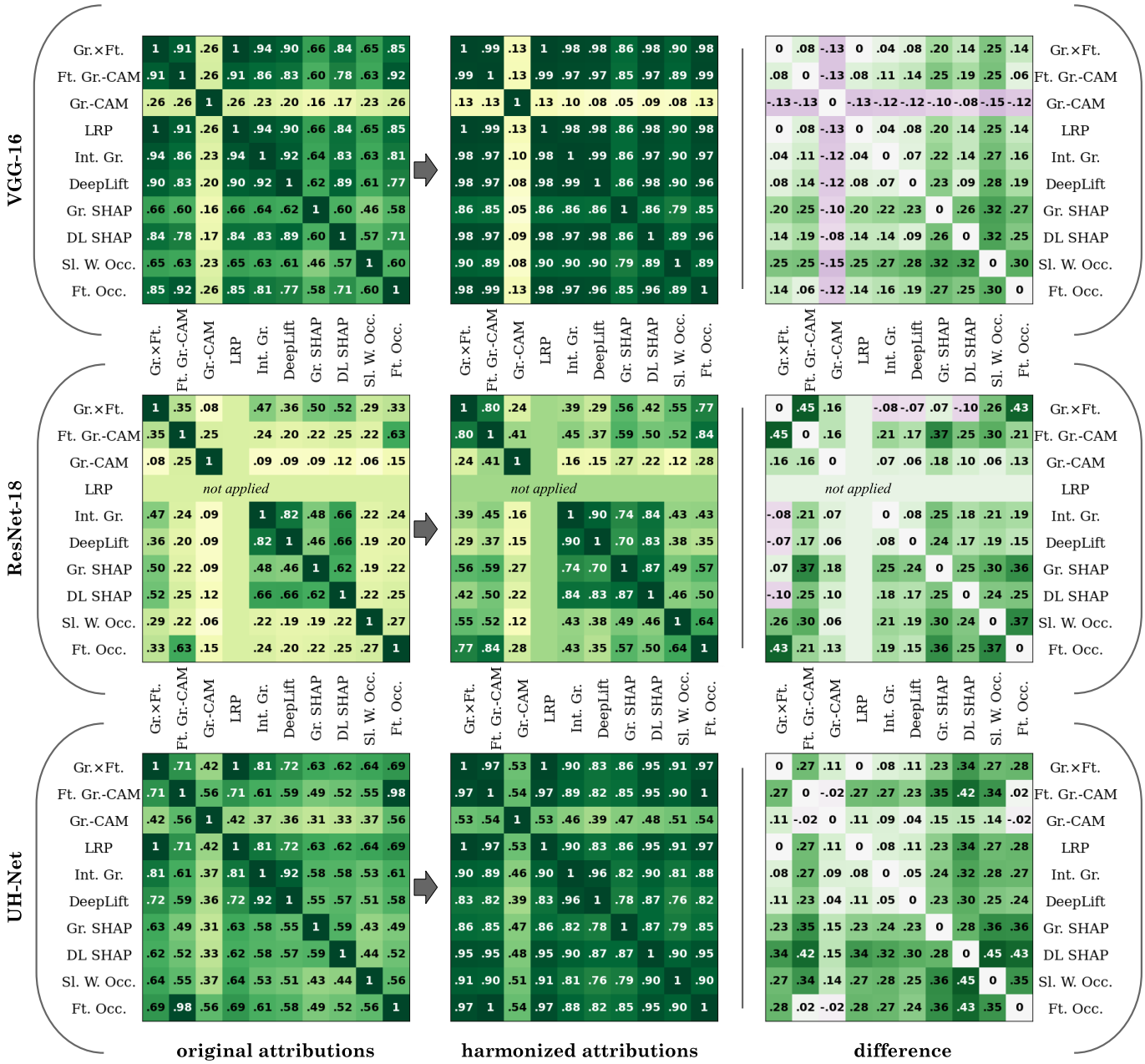


Figure 7.5: Pearson correlations between attribution methods for **Caltech 101** and three architectures: VGG-16 (top), ResNet-18 (middle), and UH-Net (bottom). From left to right: coefficients for the original attributions, the harmonized attributions, and their difference. Layer-wise Relevance Propagation is not computed for the ResNet-18 as originally no rules are defined for skip connections.

Chapter 8

Discussion

8.1 Input vs. Deep Layer Attributions

In the majority of works, most of the attribution methods are applied at the input level. This practice generally follows the original formulations of the methods, including Gradients×Features, Layer-wise Relevance Propagation, Integrated Gradients, DeepLift, Gradient SHAP, DeepLift SHAP, and Sliding Window Occlusions. An exception is Grad-CAM, whose inventors, Selvaraju et al. (2020), explicitly suggest an application to the last convolutional layer. This recommendation likely stems from its conceptual origins in Class Activation Mapping (CAM, Zhou et al., 2016), which is inherently restricted to the final convolutional layer due to its reliance on global average pooling over spatial feature maps.

Although not common practice, applying attribution methods to deeper layers appears generally reasonable from findings in mechanistic interpretability which conclude that representations in deeper layers are more refined and closely aligned with the output (see Section 3.3.2). In contrast, the input feature vectors (pixels) represent only the concept of (multispectral) color, not even structure. However, especially in remote sensing, structure plays a crucial role in differentiating specific land cover classes. Accordingly, our results demonstrate that computing attributions in deeper layers yields more intuitive explanations. While input attributions tend to be noisy for feature-vector-wise methods and often lead to misleading explanations, attributions from deep layers have proven to produce better explanations.

To the best of our knowledge, no comparable study has systematically examined and compared multiple attribution methods at both the input and deeper layers. We fill this research gap and our findings not only support the theoretical foundations established in mechanistic interpretability but also offer the practical recommendation to apply attribution methods in deeper layers. This raises an important critique of many previous studies: numerous works comparing at-

tribution methods evaluate different methods at different layers (Section 4.2). Grad-CAM, which is typically applied to the last convolutional layer, frequently performs well in these comparisons, which is further discussed in Section 8.4. However, such evaluations are flawed and potentially misleading, as it remains unclear whether the strong performance is due to the method itself or the specific layer to which it is applied.

Key Points: Input vs. Deep Layer Attributions

- Computing attributions in deeper layers yields more intuitive explanations.
- No previous study has systematically compared multiple attribution methods at both input and deep layers, and many prior comparisons are flawed by evaluating methods at different layers.

8.2 UH-Net Architecture

While attributions are more meaningful at deeper layers, these layers typically have lower resolution. Concerning this matter, our proposed UH-Net architecture (introduced in Section 5.1) provides an advantage in interpretability by enabling meaningful high-resolution attributions through the incorporation of an unconventional bottleneck layer. Typical bottleneck layers, such as those in autoencoders, contain a small number of high-dimensional feature vectors. In contrast, our representation preserves a large number of feature vectors with relatively low dimensionality. Our design comes with a potential trade-off of reduced classification accuracy compared to VGG-16 and ResNet-18. While this effect is negligible in the multi-label land cover classification tasks evaluated (DFC2020 and Ben-ge; see Table 6.3, page 42), it becomes more pronounced in the multi-class classification involving 101 classes (Caltech 101; see Table 7.1, page 65). Nevertheless, if high transparency and interpretability is desired, this trade-off might be often worthwhile.

Considering the *original* attributions for UH-Net, both attribution methods we propose in Section 5.3 — Feature-specific Grad-CAM and Feature-specific Occlusions — are among the best explaining methods for land cover classification and also yield reasonable results on Caltech 101. For the feature-vector-wise attribution methods, resolution is often limited by the presence of cluster-like regions in the attribution maps; and for Sliding Window Occlusions, resolution is constrained by the occlusion patch size. Here, harmonization is recommended as it addresses these limitations. The *harmonized* attribution maps generated by our UH-Net show the highest concordance with the segmentation ground truth compared to the other architectures (VGG-16 and ResNet-18). It is thus the

model that provides the most comprehensible feature maps for humans. This is certainly also due to the skip connections within the U-Net, which establish a strong link between deep and input features, ensuring that attributions remain closely tied to the input and are therefore more interpretable.

Key Points: UH-Net Architecture

- UH-Net enables meaningful attributions at high resolution.
- It provides the most comprehensible feature maps for humans.
- Feature-specific Grad-CAM and Feature-specific Occlusions emerge as the most suitable methods for UH-Net considering the *original* attributions.
- UH-Net's high interpretability may come at the cost of reduced classification accuracy.

Variants of UH-Net

Using a simpler classification head than our proposed one, the harmonized attributions are less aligned with the segmentation ground truth. While these attributions might accurately reflect the behavior of the model's layer, the restricting head (global average pooling and a single linear layer) forces the model to learn representations that are less interpretable to humans. We thus recommend a head that is lightly restrictive.

On the other hand, using a complex ResNet-18 head, the *original* attributions are less informative because the longer backpropagation path introduces noise and artifacts. Harmonization significantly improves interpretability and reveals a moderate feature complexity at the layer of interest — although not as high-level as in our proposed UH-Net variant. This is because ResNet-18 has nearly as many parameters as U-Net, allowing the whole CNN to learn similarly complex features in later stages.

8.3 Harmonization

Our harmonization approach (proposed in Section 5.2) compensates for the inherent inconsistencies of individual attribution methods, resulting in more coherent and reliable explanations. It thus makes the choice of method less critical. This is achieved by averaging attributions within the learned feature space using the training data. We assess our approach by analyzing attribution maps, inspecting the learned feature space, comparing similarities between attribution methods as well as between original and harmonized attributions, and evaluating alignment with segmentation ground truth. All these aspects indicate improved explana-

tions when attributions are harmonized. Further, harmonization requires only a few hyperparameters, specifically those for the k-nearest neighbor search, and these prove to be highly robust in our experiments.

Local vs. Global Explanations

Original attribution methods provide local explanations by primarily focusing on visualizing input–output relationships. They do not attempt to capture the model's overall behavior or to explain its internal mechanisms — both of which are addressed by global explainable machine learning techniques and central to the field of mechanistic interpretability. Definitions of local and global explanations, and mechanistic interpretability are provided in Section 3.4.

By harmonizing attributions, we propose a more global approach that uses the entire training dataset to evaluate the feature space and its representations in relation to the predicted classes. This enables a deeper understanding of the model's internal mechanisms, as features can be directly mapped to attributions — independent of individual samples. Because harmonized attributions align with global feature representations, they provide more interpretable explanations for individual samples. However, these global feature explanations may not fully capture individual nuances as the following examples point out:

- ResNet-18 exhibits checkerboard-like gradient maps due to strided convolutions. These patterns suggest that every second feature appears more important for the model's prediction, resulting in noisy attribution maps when using original attribution methods (see Figure 6.5, page 53). In contrast, harmonized attributions provide global explanations for these features, independent of their spatial location, thereby ignoring the checkerboard-like gradients.
- In the case of VGG-16 and Ben-ge, the bottom row and right column of a feature map are excluded from the prediction, as it is odd-sized and followed by pooling with an even-sized kernel. Most original attribution methods correctly assign zero attribution to these features (see Figure 6.4a, page 52). Harmonized attributions, on the other hand, evaluate how features generally contribute and therefore attribute importance to these features as well.

We suggest that harmonized attributions are preferable in most cases as their explanations are more interpretable to humans. However, in certain situations, original attributions may be more appropriate. It should be noted, though, that not all attribution methods capture all details — for example, Grad-CAM and Sliding Window Occlusions are unable to represent the checkerboard pattern in ResNet18.

Limitations

For harmonization, the separation of distinct feature representations is crucial; thus its effectiveness is constrained by the degree of disentanglement within the learned feature space. Similar feature vectors representing different concepts or classes cannot be well harmonized. In our experiments on land cover classification, the high agreement between attributions and segmentation ground truth suggests that the deep feature vectors are disentangled with respect to the eight classes. This appears to hold for Caltech 101 as well, as evidenced by the successful convergence of the attribution methods after harmonization. However, as the number of classes increases and interactions between them grow, the likelihood of entanglement rises. Therefore, entanglement is also more common in the initial layers, as they process low-level features, such as color and texture; which explains why harmonization is not effective at the input layer. For example, when different land cover classes share similar (multispectral) colors, their attributions become mixed in the input space. Even in deeper layers, the same semantics can support different classes — for example, depending on their position, quantity, or surroundings.

Another limitation are domain shifts between training and test data posing limitations not only for model predictions but also for the reliability of feature attributions. If a CNN encounters out-of-distribution input features, they may be mapped to out-of-distribution deep features or mistakenly to in-distribution deep features. In both cases, the resulting features may be misattributed both by the original and the harmonized attributions.

Key Points: Harmonization

- Harmonization yields more coherent and comprehensive explanations.
- It makes the choice of attribution method less critical.
- Original attribution methods provide local explanations, whereas harmonization produces explanations with a more global character.
- Harmonization is constrained by the degree of entanglement within the learned feature space.

8.4 Attribution Methods

Gradients×Features

Gradients×Features is one of the simplest attribution methods: it computes gradients and multiplies them with the corresponding features. Since gradients in neural networks tend to be noisy, the resulting attribution maps are often noisy as well. Methods such as SmoothGrad (Smilkov et al., 2017) address this by

averaging attributions across multiple noisy versions of the features. In practice, however, Gradients×Features is rarely used nowadays, as its attribution maps are often difficult for humans to interpret. It is therefore surprising how well this method performs after harmonization: It not only shows high similarity with other attribution methods but also achieves a high overlap with segmentation ground truth. For VGG-16 and UH-Net, it even ranks among the best-performing methods — despite its extreme simplicity. This further underscores the strength of harmonization.

Layer-wise Relevance Propagation

Across all experiments, we observe that Layer-wise Relevance Propagation is equivalent to Gradients×Features. This aligns with Shrikumar et al. (2017), who state that the two attribution methods are equivalent under the Basic Rule (z-Rule), provided all activations are piecewise linear. Using ReLU activations and the Epsilon Rule in our experiments, these conditions are approximately satisfied.

Grad-CAM

Grad-CAM is the method least affected by harmonization as its attributions are already well aligned with the feature vectors. This can be explained by the fact that averaging the gradients across channels defines a distinct direction $\boldsymbol{\alpha}^c$ in feature space, which is then projected onto the feature vectors (see Section 4.3 and Figure 4.1, page 33). However, as also discussed in Section 4.3, this leads to a problem whenever multiple directions are relevant for a class. For this reason, Grad-CAM fails to provide meaningful explanations in our experiments using Caltech 101.

In related work, Grad-CAM has generally been found to perform well, even on tasks with many classes. It is one of the few gradient-based attribution methods passing the sanity checks by Adebayo et al. (2018) and performs best under various evaluations by Yang and Kim (2019). Grad-CAM is also among the best-performing attribution methods for land cover classification when evaluated with the Most Relevant First approach of Samek et al. (2017), which measures how quickly the prediction score decreases as information is progressively removed (Kakogeorgiou and Karantzalos, 2021). Further, it performs best when testing the robustness of explanations by perturbating the input in a semantically meaningful way (Yang and Kim, 2019). One explanation for the strong performance of Grad-CAM in related work is that it is typically applied to the last convolutional layer. At this stage, the likelihood of multiple relevant directions in feature space is lower than in our experiments, where we mostly do not use the very last convolutional layer. Moreover, in related comparisons, other attribution methods are applied

at the input level, which makes for an unfair comparison (Section 8.1).

Feature-specific Grad-CAM

Feature-specific Grad-CAM, proposed in Section 5.3, can be seen as an approach that bridges Gradients×Features and Grad-CAM. While Gradients×Features does not perform any gradient averaging, Feature-specific Grad-CAM averages gradients over similar feature vectors, whereas Grad-CAM averages them across entire feature map channels. In this way, Feature-specific Grad-CAM substantially reduces noise in the attribution maps and, by design, generates attributions that align closely with the feature vectors. It thus provides less noisy explanations than Gradients×Features. After harmonization, Gradients×Features exhibits a strong correlation with Feature-specific Grad-CAM.

Feature-specific Grad-CAM is particularly effective for our UH-Net architecture, as it reaches its full potential with high-resolution feature maps. The limitations of this attribution method are similar to those of harmonization (Section 8.3), namely that its effectiveness depends on the degree of disentanglement in the learned feature space, which is, for example, less distinct in shallow layers.

Sliding Window Occlusions

Sliding Window Occlusions provides meaningful explanations, but their resolution is limited by the size of the occluding window. Harmonization increases the resolution, thereby improving both the alignment with the segmentation ground truth and the consistency with other attribution methods.

Feature-specific Occlusions

Feature-specific Occlusions, proposed in Section 5.3, simultaneously occludes similar feature vectors instead of using a sliding window. By design, this approach leads to attributions that are better aligned with the feature vectors, thus yielding more accurate explanations. It is particularly valuable for high-resolution feature maps. Accordingly, significant improvements over Sliding Window Occlusions are observed with UH-Net. Harmonizing Feature-specific Occlusions has a relatively minor effect, since the attributions are already closely aligned with the feature vectors. The limitations of this attribution method are similar to those of harmonization (Section 8.3), namely that its effectiveness depends on the degree of disentanglement in the learned feature space, which is, for example, less distinct in shallow layers.

Correlation between Feature-specific Grad-CAM and Feature-specific Occlusions

Feature-specific Grad-CAM and Feature-specific Occlusions exhibit high Pearson correlations with one another already prior to harmonization. We explain this behavior by the following argumentation: Based on Equation 3.4 (page 22), the Grad-CAM attribution for a feature vector $\boldsymbol{X}_j$ at position j with features $X_{k,j}$ is given by:

$$R_j^{\text{Grad-CAM}} = \frac{1}{N_{\text{all}}} \sum_{k,i}^{\text{all}} \frac{\partial f(\boldsymbol{X})}{\partial X_{k,i}} \cdot X_{k,j} \tag{8.1}$$

N_{all} denotes the total number of feature vectors in the sample; k and i denote the channel and position of a feature vector, respectively. Differing from this, Feature-specific Grad-CAM averages gradients only over similar feature vectors (denoted by "sim") rather than across entire feature map channels:

$$R_j^{\text{Ft. Grad-CAM}} = \frac{1}{N_{\text{sim}}} \sum_{k,i}^{\text{sim}} \frac{\partial f(\boldsymbol{X})}{\partial X_{k,i}} \cdot X_{k,j} \tag{8.2}$$

Using an occlusion-based method, the attribution is determined as shown in Equation 8.3, where N_{occ} denotes the number of feature vectors simultaneously occluded. Assuming the contributions of occluded features are approximately additive, we obtain the expression in Equation 8.4. Here, $\sum_{k,i}^{\text{occ}}$ denotes summation over all occluded positions. By multiplying the term by 1 in the form $\Delta X_{k,i}/\Delta X_{k,i}$, we arrive at Equation 8.5. When features are occluded by setting them to zero, we have $\Delta X_{k,i} = -X_{k,i}$. Since Feature-specific Occlusions involve occluding only similar features simultaneously, we approximate $X_{k,i} \approx X_{k,j}$, leading to Equation 8.6.

$$R_j^{\text{Ft. Occ.}} = -\frac{1}{N_{\text{occ}}} \cdot \Delta f(\boldsymbol{X}, \Delta\boldsymbol{X}) \,, \qquad \Delta f(\boldsymbol{X}, \Delta\boldsymbol{X}) \approx (f(\boldsymbol{X} + \Delta\boldsymbol{X}) - f(\boldsymbol{X})) \tag{8.3}$$

$$\approx -\frac{1}{N_{\text{occ}}} \sum_{k,i}^{\text{occ}} \Delta f(\boldsymbol{X}, \Delta X_{k,i}) \qquad | \cdot \frac{\Delta X_{k,i}}{\Delta X_{k,i}} \tag{8.4}$$

$$= -\frac{1}{N_{\text{occ}}} \sum_{k,i}^{\text{occ}} \frac{\Delta f(\boldsymbol{X}, \Delta X_{k,i})}{\Delta X_{k,i}} \cdot \Delta X_{k,i} \qquad | \, \Delta X_{k,i} = -X_{k,i} \approx -X_{k,j} \tag{8.5}$$

$$\approx \frac{1}{N_{\text{occ}}} \sum_{k,i}^{\text{occ}} \frac{\Delta f(\boldsymbol{X}, \Delta X_{k,i})}{\Delta X_{k,i}} \cdot X_{k,j} \tag{8.6}$$

The resulting equation for Feature-specific Grad-CAM (Equation 8.2) and the approximation for Feature-specific Occlusions (Equation 8.6) take a very similar form, differing in only one aspect: Grad-CAM relies on local gradients, $\partial f / \partial X_{k,i}$, while Feature-specific Occlusions measure global changes relative to a zero baseline, $\Delta f / \Delta X_{k,i}$. For a linear problem, both forms are equivalent apart from a constant, which is irrelevant for the Pearson correlation. The more nonlinear the problem, the more the gradients and the deltas to the zero baseline diverge, and thus the greater the difference between the two attribution methods.

Integrated Gradients, DeepLift, Gradient SHAP, and DeepLift SHAP

We observe a high correlation between Integrated Gradients and DeepLift already prior to harmonization. This is consistent with Ancona et al. (2018), who describe DeepLift as an approximation of Integrated Gradients for most architectures. In our land cover classification experiments, the two methods correlate strongly with each other but show the weakest correlation with the other attribution methods. Although harmonization increases their correlation with other methods, they still exhibit the lowest overall similarities. Additionally, both methods demonstrate comparatively low alignment with the segmentation ground truth. In contrast, Gradient SHAP and DeepLift SHAP exhibit significantly higher correlation with the other methods and greater alignment with the segmentation ground truth. Their primary difference from Integrated Gradients and DeepLift lies in the use of baselines representing the training data rather than a zero baseline.

At first glance, it is striking that a zero baseline is not a good choice, given that all layers are ReLU-activated and zero should therefore serve as a reasonable value for neutrality in a deep layer. Moreover, all other methods (Gradients×Features, Feature Grad-CAM, Grad-CAM, Layer-wise Relevance Propagation, Sliding Window Occlusions, and Feature Occlusions) assume a zero baseline, either explicitly or implicitly (see Section 3.4.3). So why are the explanations from Integrated Gradients and DeepLift less similar to the other methods when a zero baseline is used? We explain this by the fact that these methods compute attributions along a path from the baseline to the feature values. If this path traverses regions of the feature space the model was never trained on, gradients can behave erratically, leading to misleading attributions. With a zero baseline, there is a high likelihood that the path passes through out-of-distribution values, whereas a baseline representing the training data is more likely to result in a path that stays within in-distribution values.

Key Points: Attribution Methods

- Although one of the simplest attribution methods, Gradients×Features is a high performing attribution method after harmonization.
- Layer-wise Relevance Propagation is equivalent to Gradients×Features under the conditions of our experiments.
- Grad-CAM aligns well with the feature vectors already before harmonization. However, explanations fail under certain conditions.
- Feature-specific Grad-CAM successfully addresses the limitation of Grad-CAM.
- Sliding Window Occlusions is limited by resolution, a drawback that harmonization effectively addresses.
- Feature-specific Occlusions successfully addresses the limitation of Sliding Window Occlusions.
- Feature-specific Grad-CAM and Feature-specific Occlusions demonstrate strong similarity, both in theory and evidentially.
- Integrated Gradients and DeepLift fail land cover explanations using a zero baseline.
- Gradient SHAP and DeepLift SHAP successfully address this limitation using baselines representing the training data.

8.5 Recommendations

Based on theory and our results, we recommend the following when computing attributions:

- We recommend computing attributions for a deep layer, as features at this level capture higher-level representations.

- If UH-Net achieves sufficiently high task accuracy, its use is beneficial for generating high-resolution, interpretable attribution maps.

- For deep layers, we recommend using the attribution methods Feature-specific Grad-CAM and Feature-specific Occlusions. Both methods provide reasonable explanations that align well with the feature vectors. In particular, we recommend Feature-specific Occlusions as it better captures the global effect of a feature (see Figure 3.8, page 26). However, depending on the task, Feature-specific Grad-CAM may be more computationally efficient. We do *not* recommend using both methods for input attributions; here, DeepLift SHAP has yielded the best results.

- Harmonization reduces differences between attribution methods, making the choice of method less critical. It also enhances interpretability of the explanations. We therefore recommend its use in all cases.

These recommendations must be reconsidered for each specific use case and architecture, and should not be seen as a one-size-fits-all solution.

Key Points: Recommendations

- When computing attributions, we recommend applying Feature-specific Occlusions to a deep layer and harmonizing the attributions. We recommend UH-Net if it achieves sufficiently high task accuracy.

8.6 Related Insights and Future Directions

8.6.1 Weakly-supervised Segmentation

Training machine learning models in a supervised manner requires large quantities of labels, which are often labor-intensive to obtain, especially in remote sensing applications. Because satellites capture data over vast areas, complete labeling is often unfeasible. As a result, segmentation annotations in remote sensing are rare and often coarse. Weakly-supervised learning methods can assist in addressing this challenge (Zhou, 2018).

Attribution-based Weakly-supervised Segmentation

A common application in weakly-supervised segmentation is deriving pixel-level segmentation from image-level labels. To accomplish this, CNNs are often evaluated using attribution methods (Zhang and Ma, 2021; Kwak et al., 2017; Ahn and Kwak, 2018; Wang et al., 2020b; Chong et al., 2021). Harmonization has the potential to improve the accuracy of most such methods, as it has shown to enhance segmentation performance. It is also applicable across all CNN architectures.

Superpixel-based Weakly-supervised Segmentation

Many weakly-supervised approaches assume that input pixels with similar colors and spatial connectivity are likely to belong to the same class. This assumption helps improve various aspects of segmentation, particularly the delineation of object boundaries when using attribution methods. Building on this idea, Jonnarth and Felsberg (2022) introduce a feature similarity loss; Kwak et al. (2017) propose a superpixel pooling layer that enforces uniform features within each superpixel derived from the input image; and Zeng et al. (2023) present a Global Superpixel

Consistency Module, which ensures that similar superpixels in the input image correspond to similar features in the penultimate feature map.

In contrast to pictures and photographs, where color is a useful feature for distinguishing individual objects, satellite imagery for land-cover analysis is characterized not only by its multispectral information but also strongly by texture and structural patterns. Willbo et al. (2024) evaluate various land cover segmentation models and conclude that texture is even more important than color. Moreover, many land cover classes share similar colors, further limiting the effectiveness of superpixel-based approaches. Harmonizing attributions utilizing the learned feature space offers a promising alternative to relying on the image's spatial domain.

Pseudo-label Training

Often, accuracy can be further improved by reusing predictions from a weakly-supervised model as pseudo-labels to train a new segmentation network from scratch (Ahn and Kwak, 2018; Chen et al., 2020; Zhang et al., 2020; Hanna et al., 2023). However, we do not observe such an improvement in our framework: We train a randomly initialized U-Net using pseudo-labels derived from the harmonized attribution maps of the DFC2020 training data produced by UH-Net. We test two approaches: hard labels (rounded to {0, 1}) and soft labels (normalized to the [0,1] range via min-max normalization). The models are trained to minimize validation loss, with pseudo-labels also used for validation. Evaluation is performed using the ground truth segmentation of the test dataset. Neither approach yields a significant improvement in test accuracy, as the new segmentation model overfits to the pseudo-labels used for validation. Stronger regularization might help, but for the given task there is no alternative reference for computing validation loss besides the pseudo-labels themselves. One possible reason why prior studies, such as those previously listed, report an improvement is that their attribution maps have lower resolution, allowing the newly trained model to enhance spatial precision. In contrast, the attribution maps produced by UH-Net already match the input image resolution.

Independent of the network architecture, harmonizing attributions consistently enhances segmentation accuracy. Thus, when combined with existing attribution-based, weakly-supervised learning methods, it has the potential to provide significant benefits.

Key Points: Weakly-supervised Segmentation

- Harmonizing attributions has the potential to enhance the accuracy of attribution-based, weakly-supervised segmentation.
- It presents a promising alternative to superpixels, particularly for remote sensing tasks.

8.6.2 Vision Transformers

Transformer-based models have been widely applied to sequential data and, since the introduction of the Vision Transformer (Dosovitskiy et al., 2021), to image data as well. They rely on self-attention mechanisms, capturing dependencies between elements. Vision Transformers divide the input image into patches and model dependencies among them. Transformers are popularly investigated in explainable machine learning because their attention mechanisms can provide insights into the importance of individual patches. However, there are ongoing doubts whether attentions truly provide meaningful and interpretable explanations. Jain and Wallace (2019) find in their experiments that attention weights largely do not. Wiegreffe and Pinter (2019) disagree and argue that attentions can indeed serve explanations, although those may not be intuitive or easily interpretable. Darcet et al. (2024) find out that artifacts can emerge within low-information background regions. They introduce additional *register* tokens to the input sequence which fulfill that role, making attention maps considerably more interpretable.

If attentions are interpretable, e.g. through registers, our harmonization approach might also be applicable to Vision Transformers. Images are partitioned into patches, which are then processed by multi-head attention; and attention maps can be derived from the attention weights of a multi-head attention layer. Analogous to the channels in a convolutional layer, the heads in this context could serve to define *attention* vectors (instead of feature vectors). Harmonization would then be carried out in the attention space, following a similar logic. Alternatively, the features of a normalization layer in a transformer's encoder block could be used for harmonization. An interesting research question, which is however beyond the scope of this thesis, is whether harmonization in Vision Transformers leads to similar effects as it does in CNNs.

Key Points: Vision Transformers

- If attentions are interpretable, e.g. through registers (Darcet et al., 2024), harmonization might be applicable to Vision Transformers.

8.6.3 Global Feature-specific Attributions

In Section 5.3, we propose Feature-specific Grad-CAM, in which gradients are averaged over similar feature vectors and then multiplied with those. This is performed separately for each sample, resulting in attributions that require harmonization to ensure full consistency across multiple samples. Alternatively, one could compute average gradient values over similar feature vectors across the entire training dataset. In this approach, k-means clustering would be applied to the full training feature space. Multiplying feature vectors by these "global" gradient vectors $\boldsymbol{\alpha}^c$ would directly yield harmonized attributions.

This idea of having global gradient vectors within the feature space has parallels to the approach of Testing with Concept Activation Vectors (TCAV) introduced by Kim et al. (2018). TCAV identifies human-interpretable concepts such as colors, stripes or styles, on which a neural network bases its decision by locating vectors in the feature space that are sensitive to these concepts. Both TCAV and our idea of global gradient vectors aim to associate interpretable directions in the feature space with model sensitivity. A key difference, however, is that TCAV constructs the feature space using entire feature maps, whereas we construct it using the feature vectors of the feature maps.

Key Points: Global Feature-specific Attributions

- Averaging gradients over the training set for Feature-specific Grad-CAM would already yield harmonized attributions.

Part III

Mapping Naturalness in Fennoscandia

Chapter 9

Related Work and Research Gaps

It is widely recognized that landscapes with a high degree of naturalness offer important ecological and social benefits and therefore warrant preservation. Monitoring these areas can support environmental decision-making and conservation efforts. Although land cover mapping has become a central task in modern environmental research, it faces significant challenges when accurate labeling is difficult. Moreover, given the global impact of human activities, the notion of sites with pristine nature completely unmodified by human actions, especially in Europe, is tenuous at best, limiting the amount of training data.

9.1 Anthropogenic Stressors and Their Impact on Naturalness

From a geophysical perspective, humans have rapidly expanded and strongly influenced Earth's environment within a relatively short timeframe (Steffen et al., 2011). Regions free from human impact on naturalness have significantly diminished and are rare in most parts of the world (Allan et al., 2017). Although urbanization and agriculture have provided numerous benefits, land use has caused substantial ecological impacts. Nevertheless, countless species, including humans, rely on the essential functions of natural ecosystems. Water cycles provide freshwater, forests regulate air quality, moors sequester carbon dioxide, pollinators are essential for the survival of flowering plants and successful harvests, etc. Disturbing natural ecosystems has an impact on biodiversity, pathogen spread, climate, and much more (Foley et al., 2005). Newbold et al. (2015) estimate that land use has resulted in ecological regions losing an average of 13.5 % of species compared to pristine habitats. However, genetic variability can be important for nature's ability to adapt to environmental changes as the model of Lande and Shannon (1996) shows. Concerning these matters, wilderness areas can offer important ecological and social benefits; and there are urgent and pragmatic reasons to

identify where these positive characteristics and ecological functions associated with naturalness are present and able to flourish.

9.2 Traditional Naturalness Mapping

Approaches such as the Human Footprint by Sanderson et al. (2002) provide valuable insights into the global distribution of natural areas. The authors compute a Human Influence index based on eight human pressures across several categories, with most of the data originating from the early 1990s: population density, land transformation, and accessibility indicators such as roads, and electrical power infrastructure. Defined scoring methods produce an index that is globally mapped at a resolution of 1 km. Venter et al. (2016a) apply the same methodology producing an additional Human Footprint map for 2009 alongside the 1993 version. This first set of temporally consistent Human Footprint maps allows analyses of changes over time. Using an improved data basis but a similar approach, Mu et al. (2022) create annual Human Footprint maps between 2000 and 2018. Kennedy et al. (2019) further contribute to the field by creating a global map of Human Modification in 2016, based on 13 anthropogenic stressors

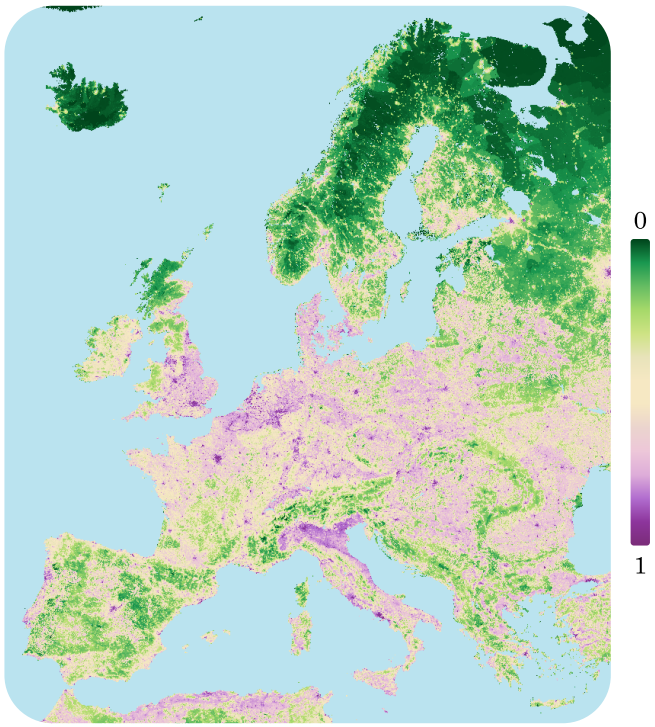


Figure 9.1: The degree of Human Modification by Kennedy et al. (2019) is computed based on 13 anthropogenic stressors resulting in a measure from no modification (value 0) to very high modification (value 1) with a resolution of 1 km. While many regions experience significant human impact, vast natural areas can still be found in places like Fennoscandia (Norway, Sweden, and Finland).

within five categories: human settlement, agriculture, transportation, mining and energy production, and electrical infrastructure. Their map has a resolution of 1 km as well. The Human Modification of Europe is illustrated in Figure 9.1.

While Sanderson et al. (2002) discusses the concept of the "last of the wild", Kennedy et al. (2019) reveal that most of the world's eco-regions exhibit moderate to high levels of modification, with only a few areas remaining untouched by human influence. Venter et al. (2016b) find that the Human Footprint increased by 9% over the 16-year period from 1993 to 2009. These are important conclusions, however, all these traditional methods are restricted to the low spatial and low temporal resolution of their underlying data. By combining low-resolution indicators with high-resolution data, specifically ESA WorldCover (Zanaga et al., 2022) and OpenStreetMap (OpenStreetMap contributors, 2017), Ekim et al. (2021) produce a naturalness map at 10 meter resolution for Bavaria, Lapland, and Scotland. However, traditional approaches rely on human-made assumptions, such as scoring types of human pressure. They also do not assess (bio)physical effects of human pressure, such as vegetation health and composition, which limits the depth of insights they can provide.

9.3 Monitoring the Environment Using Remote Sensing

Utilizing multispectral satellite imagery could address these problems by providing high-resolution, frequent mapping and the ability to include vegetational and geographic assessment. Satellite data is already commonly used for environmental monitoring (Pettorelli et al., 2018) and frequently integrated with machine learning (Alotaibi and Nassif, 2024). Examples of applications include monitoring vegetation phenology, biodiversity, soil composition, habitat extent, water dynamics, air quality, gas emissions, climate, and disturbances like wildfires, floods, and droughts (Pettorelli et al., 2018; Reddy et al., 2021). Many studies focus specifically on forests, utilizing remote sensing data to assess factors such as forest age (Schumacher et al., 2020), treeline ecotones (Nguyen et al., 2022), forest ecosystem condition (Ørka et al., 2022), and forest cover change (Hansen et al., 2013).

While many of these applications also provide insights into assessing the naturalness of an area, their scope is often limited in this regard. For instance, assuming that ecosystems with low biodiversity are inherently dysfunctional is an oversimplification, as some of the largest and oldest ecosystems (boreal forests, heathlands, boglands) have low species diversity (Grime, 1997). Being able to accurately analyze the extent of naturalness would be a valuable tool and could

allow for the tracking of conservation areas, reforestation efforts, etc. As such, it could be utilized as an indicator for policy-making and land use planning, e.g. for the Sustainable Development Goal 15 of the United Nations (2021), which focuses on protecting, restoring, and promoting sustainable land use.

Chapter 10

A Novel Approach for Naturalness Mapping

In this section, we present a novel approach for naturalness mapping, leveraging our framework introduced in Section 5. Hereby, we focus on Fennoscandia, specifically Norway, Sweden, and Finland. This region is an interesting study area as it includes landscapes that are relatively natural by European standards while also showing both historical and ongoing human influence. Traditional naturalness mapping approaches such as the Human Footprint by Sanderson et al. (2002), Venter et al. (2016a), and Mu et al. (2022), as well as the global map of Human Modification by Kennedy et al. (2019), map broad areas within Fennoscandia as regions with relatively minimal (disruptive) anthropogenic influence. Furthermore, all three countries have high environmental standards according to the Environmental Performance Index by Wendling et al. (2020) and there have been longstanding, strict conservation efforts in certain protected regions. On the other hand, over the past 300 years, the landscapes of Fennoscandia, especially forests, have seen anthropogenically-driven changes before regulations were introduced to protect them. This affects the southern regions more than the northern ones. Kouki et al. (2001) and Östlund et al. (1997) give detailed overviews of forest fragmentation in Fennoscandia and the transformation of the boreal forest landscape in Scandinavia, respectively. Clear-cutting remains a predominant forest management practice today, particularly in Sweden and Finland (Lunde et al., 2025).

Fennoscandia

Within Fennoscandia, we consider Norway, Sweden, and Finland.

10.1 Finding Natural and Anthropogenic Regions

10.1.1 Conceptualization

Although truly untouched nature is rare even in Fennoscandia, there are undoubtedly places and conditions where, at the least, certain processes and biophysical conditions are better able to thrive. We do not posit that natural areas are completely free of human presence or intervention; and we are not searching for some idealized or romanticized notion of authentic or true nature as often associated with wilderness (Cronon, 1995). Instead, we are searching for places in which the qualities of naturalness are preserved and promoted. This may even include an active role for humans as managers, preserving natural conditions with influences that are presumably synergistic to nature.

Conversely, there are areas of pervasive and continuous human influence, which are actively and intentionally maintained for specific human purposes or functions (e.g., cities and communities, infrastructures, industrial processes, agriculture, etc.). This creates a categorically different type of human presence and intervention, is destructive to nature, and here denoted as *anthropogenic*.

Naturalness and Anthropogenic Impact

We define naturalness not by the absence of humans, but by the type of human impact. Natural areas may be managed by humans to preserve natural conditions, whereas anthropogenic regions are pervasively shaped to serve human purposes or functions.

10.1.2 Protected Regions

The World Database on Protected Areas (WDPA), managed by the United Nations Environment Programme's World Conservation Monitoring Centre (UNEP-WCMC) and the International Union for Conservation of Nature and Natural Resources (IUCN), provides global polygon data of protected areas. They are classified into seven categories as proposed by Dudley (2008), of which we consider the three most stringent ones:

- **Strict nature reserves (category Ia)** are "set aside to protect biodiversity and also possibly geological/geomorphological features, where human visitation, use and impacts are strictly controlled and limited to ensure protection of the conservation values. Such protected areas can serve as indispensable reference areas for scientific research and monitoring."

- **Wilderness areas (category Ib)** are "usually large unmodified or slightly modified areas, retaining their natural character and influence, without permanent or significant human habitation, which are protected and managed so as to preserve their natural condition."
- **National parks (category II)** are "large natural or near natural areas set aside to protect large-scale ecological processes, along with the complement of species and ecosystems characteristic of the area, which also provide a foundation for environmentally and culturally compatible spiritual, scientific, educational, recreational and visitor opportunities."

We consider terrestrial and coastal areas that have been protected since at least the year 2000 using the database from February 2025.

10.1.3 Anthropogenic Regions

To find anthropogenic areas, we use the Copernicus CORINE Land Cover dataset by the European Environment Agency (2018) and locate areas with the following land cover classes:

- **Artificial surfaces (class 1)**, including the subclasses: urban fabric; industrial, commercial and transport units; mine, dump, and construction sites; artificial, non-agricultural vegetated areas.
- **Agricultural areas (class 2)**, including the subclasses: arable land; permanent crops; pastures; heterogeneous agricultural areas.

We add a 1 km buffer around these areas to include other land cover classes such as forests, shrubland, open spaces, and water bodies, assuming that nearby regions are also largely affected by Human Modification.

10.2 Building a Dataset

Stomberg et al. (2023) published the AnthroProtect dataset, which we improve and extend across multiple years in this thesis. We refer to this enhanced version as *AnthroProtect 2.0*.

10.2.1 Multispectral Sentinel-2 Imagery

We export multispectral image composites from Sentinel-2 for all of Fennoscandia and for each year from 2018 to 2024. We chose Sentinel-2 over Landsat 8/9 due to its higher spatial resolution and the availability of additional red-edge bands. Furthermore, a study by Astola et al. (2019) concludes that Sentinel-2

outperforms Landsat 8/9 in predicting forest parameters in Finland. We export the Sentinel-2 data using Google Earth Engine (Gorelick et al., 2017) with the following procedure:

- We divide Fennoscandia into a grid of 100 × 100 km tiles. For each grid and year, we export a composite.
- Sentinel-2 images are filtered to cover the summer period from June 1st to September 30th.
- We use the atmospheric corrected Sentinel-2 products (Level-2A).
- A mask for clouds, cirrus, and cloud shadows is created for each image using the Quality-60 m band (QA60) and scene classification map (SCL) provided by Copernicus Sentinel-2. Only images with a mask fraction of less than 10% within the grid tile are taken to build the composite. If the resulting composite contains gaps or cloud-related artifacts, a 20% filter is applied on a case-by-case basis. If the composite remains unsuitable, a 50% filter is used as a last resort.
- The following twelve bands are exported: B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, B12. Bands with a 20-meter resolution are upsampled using nearest-neighbor interpolation to match the 10-meter resolution of the higher-resolution bands.
- The composite is computed using the 25th percentile, which we find produces fewer artifacts than median or mosaic compositing — a finding also reported by Corbane et al. (2020). Masked areas are dilated with a radius of 100 meters to prevent artifacts at the transition of clouds and are excluded from the calculation.

10.2.2 Training, Validation and Test Samples

We intend for a convolutional neural network to distinguish between natural and anthropogenic landscapes, thereby learning meaningful feature representations from the corresponding patterns. Since labeling satellite images with conventional methods — namely visual inspection by humans — is not feasible for the task at hand, we instead assign labels based on whether the images fall within protected or anthropogenic regions, as described in Section 10.1. To this end, we divide Fennoscandia into a grid of 2560 × 2560 meter tiles:

- Tiles containing at least 90% protected areas are assigned label 1 (*protected*). They are included in the train, validation and test datasets.

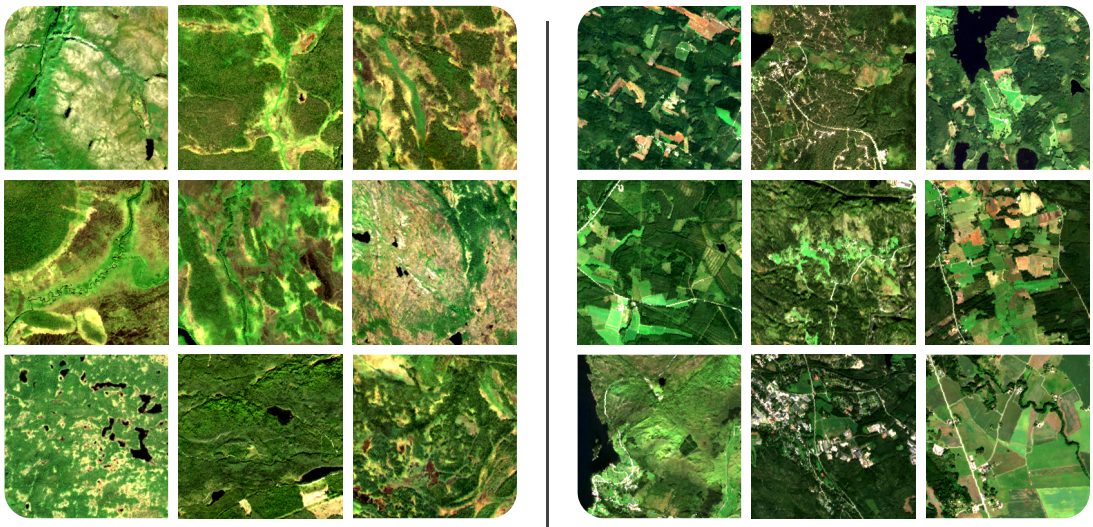


Figure 10.1: Training samples from protected and anthropogenic regions in 2024.

- Tiles containing at least 90% anthropogenic areas are assigned label 0 (*anthropogenic*). They are included in the train, validation and test datasets.
- Tiles containing both categories in any proportion are excluded.
- Tiles that contain neither protected nor anthropogenic areas are added to the test dataset and called *other*.

Since natural features can exist outside protected regions and anthropogenic impact may occur within them — e.g., roads within national parks — the annotation may be incomplete and prone to noise.

In Fennoscanida, protected and anthropogenic regions are unevenly distributed: Protected regions tend to be more northern and inland, while anthropogenic regions are generally located further south and closer to the coasts. To reduce this imbalance, we divide Fennoscandia into a 100 × 100 km grid. From each grid cell, we randomly select 100 tiles from protected areas, 30 tiles from anthropogenic areas, and 10 tiles from other regions. This results in a reduced dataset that favors less densely represented tiles within each category, leading to a more geographically balanced distribution — effective for anthropogenic regions, but still limited for protected ones. Examples of Sentinel-2 images are provided in Figure 10.1 *(left)* and the locations of all samples are shown in Figure 10.2.

Data Split

We split the protected and anthropogenic locations separately for both categories into training, validation, and test subsets using a 60 % / 20 % / 20 % ratio. To ensure spatial independence, we priorly build spatial clusters as follows: locations

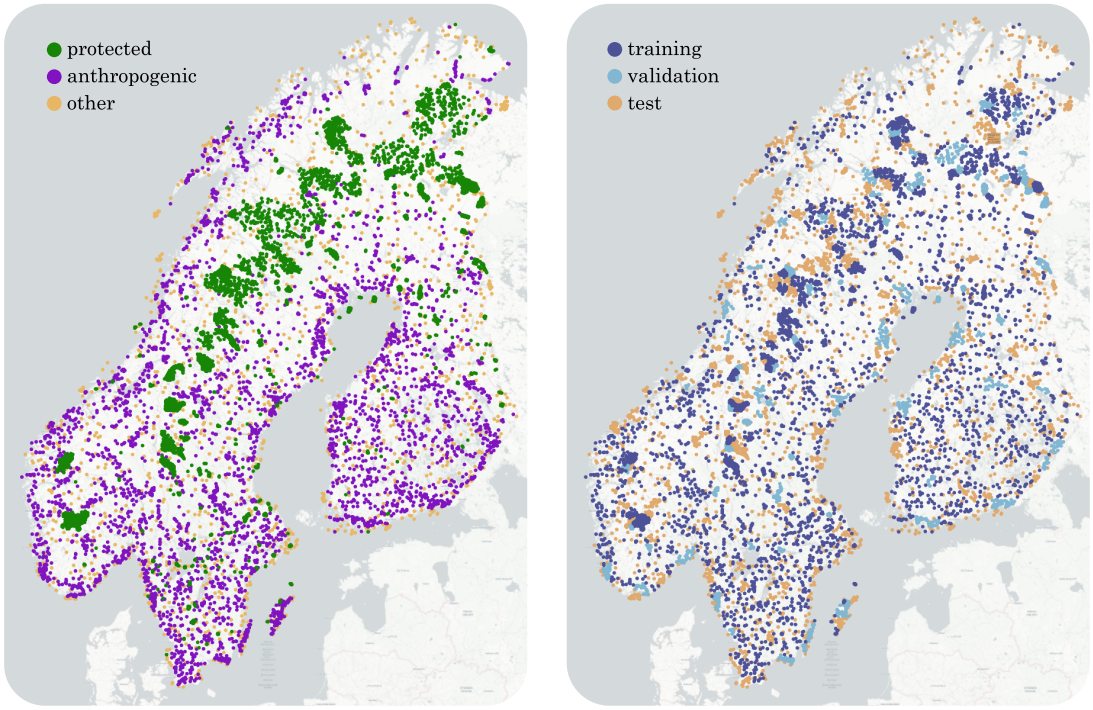


Figure 10.2: Locations of samples within our dataset. Left: Colored by category. Right: Colored by data split. For each of the 8,830 locations, one Sentinel-2 image per year is included, covering the period from 2018 to 2024. (Created using Plotly Technologies Inc. (2015); map copyright holders are © Carto and OpenStreetMap contributors.)

Table 10.1: Number of samples in the dataset separated by categories and subsets.

Category / Subset	Training	Validation	Test	Total
Protected	15,449	5,159	5,145	25,753
Anthropogenic	15,449	5,145	5,159	25,753
Other	0	0	10,304	10,304
Total	30,898	10,304	20,608	61,810

are grouped using DBSCAN (Ester et al., 1996), where points are considered part of the same cluster if they lie within 10 km of one another. Large clusters are further subdivided using the k-means algorithm by Lloyd (1982). All samples within a given cluster are assigned to the same dataset, with small clusters (< 5) assigned to the training set. Both clustering algorithms are implemented using scikit-learn by Pedregosa et al. (2011). Compared to a random split, our procedure reduces the incidence of spatially close samples appearing in different subsets. The resulting data split is visualized in Figure 10.2 *(right)*, and sample counts are summarized in Table 10.1.

10.3 Technique and Experimental Setup

We train UH-Net, introduced in Section 5.1, to classify between protected and anthropogenic regions using our dataset presented in the previous section (Section 10.2). Subsequently, we generate high-resolution attribution maps using Feature-specific Occlusions as proposed in Section 5.3. To enable large-scale mapping, the attributions are harmonized as advised in Section 5.2.

10.3.1 Model Architecture and Training

The attribution maps of UH-Net show strong alignment with the segmentation ground truth for land cover classification (Section 6.2.4), making it a promising candidate for naturalness mapping as well. We omit batch normalization at the intermediate layer and train the model using a final sigmoid activation, binary cross entropy loss, the AdamW optimizer, a batch size of 32, a learning rate of 10^{-4}, a weight decay of 10^{-4}, and a dropout probability of 0.3. The learning rate is linearly warmed up over 1 epoch and reduced by a factor of 10 if the validation loss stagnates for 3 epochs. The model state with the lowest validation loss is selected. We train 5 models with different initializations and randomizations. The results are averaged. Training is performed using PyTorch (Paszke et al., 2019) on an NVIDIA A100 40 GB PCIe GPU.

Mixing Augmentation

During training, we apply mixing augmentation methods, specifically MixUp (Zhang et al., 2018) and CutMix (Yun et al., 2019). Both are data-driven regularization techniques that generate softer targets, which helps smooth decision boundaries in neural networks. They reduce the model's sensitivity to small input variations, improve generalization and decrease overfitting (Zhang et al., 2021). Applying mixing augmentations, we find that our models produce more meaningful attribution maps, as it distributes attention more evenly across the entire image rather than concentrating on just a few regions. Both augmentation methods are illustrated in Figure 10.3.

MixUp generates new training data by blending pairs of input images, $\boldsymbol{X}_0^{(1)}$ and $\boldsymbol{X}_0^{(2)}$, and their corresponding targets, $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(2)}$, through linear interpolation:

$$\boldsymbol{X}_0^{(\text{mix})} = \lambda \boldsymbol{X}_0^{(1)} + (1-\lambda)\boldsymbol{X}_0^{(2)} \tag{10.1}$$

$$\boldsymbol{y}^{(\text{mix})} = \lambda \boldsymbol{y}^{(1)} + (1-\lambda)\boldsymbol{y}^{(2)} \tag{10.2}$$

CutMix combines two images through a cut-and-paste approach. The targets are adjusted with a mixing coefficient λ corresponding to the relative area of the cut-mixed patch as shown in Equation 10.2. We apply both MixUp and CutMix

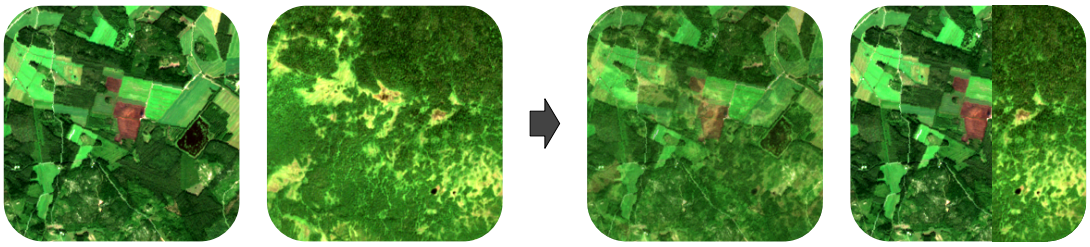


Figure 10.3: Mixing augmentation. The Sentinel-2 images on the left show an anthropogenic area and a protected area, respectively. On the right, MixUp and CutMix have been applied to these images, incorporating 40% of the protected area, resulting in a label of 0.4.

to each sample by randomly drawing another sample from the training data and randomly selecting the mixing coefficient $\lambda \in [0.5, 1]$, resulting in continuous targets between 0 and 1.

Test Accuracy

When evaluated on the non-augmented test images from the categories *anthropogenic* and *protected* (excluding *others*), the five models achieve a classification accuracy of 99.37 ± 0.04 %.

10.3.2 Harmonized, Feature-specific Attributions

To obtain pixel-wise naturalness maps from our image-wise trained model, we use Feature-specific Occlusions, as proposed in Section 5.3. For UH-Net, Feature-specific Occlusions proves to be one of the most accurate attribution methods for land cover classification, as demonstrated in Section 6.2.4. It is further a more global attribution method than Feature-specific Grad-CAM and should therefore better capture the global effect of a feature as discussed in Section 8.4.

When computing original attributions for large-scale maps, the input must be divided into patches that match the training data size, which is 256×256 pixels. Other image sizes are not supported because of the final fully connected layers and missing adaptive pooling in UH-Net. However, as shown in Figure 10.4, original attributions are not well comparable across patches. This inconsistency arises from interactions and mutual influences between multiple patterns within each patch, which are processed jointly by the fully connected layers. In contrast, harmonized attributions are derived directly from the features of the layer of interest, ignoring the final fully connected layers. As a result, they remain consistent across patches and can be computed for images of any size. For this reason, harmonizing attributions is essential for generating consistent large-scale maps.

Figure 10.4: To compute large-scale original attribution maps for our UH-Net, the input must be split into 256 × 256 patches and the attribution values are inconsistent across patches. In contrast, harmonized attributions remain consistent and can be computed for images of any size. The region shown is the area surrounding the Letsi hydroelectric power plant in northern Sweden. The illustrated attribution maps are generated from one of the five models; ■ negative, ■ zero, ■ positive.

To produce harmonized attribution maps, we first compute original Feature-specific Occlusions for the UH-Net's layer of interest using the mixing-augmented training data. We select a representative random subset of approximately 100,000 training feature vectors and their corresponding attributions for performance reasons. Harmonized attributions are then computed using the k-nearest neighbors regressor from the RAPIDS cuML Python library by Raschka et al. (2020), with $k = 100$ and an Euclidean distance metric. We choose the Euclidean metric instead of cosine similarity because of the binary classification problem with opposing classes. Here, not only the angle but also the magnitude of the vectors may influence the strength of the attributions. However, in line with the findings in Section 7.1, we do not observe significant differences between the two distance metrics.

10.3.3 Large-scale Mapping

Since the entire Fennoscandia region is too large to be processed at once, we divide it into tiles of size 1024 × 1024 pixels. To mitigate potential border effects, caused by filters not fully traversing the patch edges, we add a padding of 256 pixels around each tile. This padding consists of the original surrounding data and is removed from the predictions afterward.

In this way, we compute the harmonized attributions for all of Fennoscandia and for each year from 2018 to 2024, separately for each of the five models. The resulting attribution maps are then averaged across the models. This results in one naturalness map for each year from 2018 to 2024 which are illustrated in Appendix A.2, Figure A.5 (page 143). These yearly maps are used as a time series for change detection, as described in Section 10.3.5. To create a naturalness

base map, we average the maps over the seven years to produce a composite map representing the midpoint year, 2021, shown in Figure 10.6 (page 100). We denote a naturalness score by Ψ.

10.3.4 Quantile Intersection Threshold

Figure 10.5 shows the distributions of harmonized attributions in our training regions, grouped by the labels *anthropogenic* and *protected*. The differing spread of the distributions indicate that our models assess natural characteristics differently from anthropogenic ones: in *protected* regions, the values fall within a relatively narrow range, whereas in *anthropogenic* regions, they span a much broader range. Also, within *protected* regions, 99 % of the attributions are positive, while within *anthropogenic* regions only 76 % are negative.

For this reason, using a zero-threshold may not be sufficient to determine whether attributions rather represent natural or anthropogenic patterns. To address this, we define the quantile intersection threshold Ψ_Q: For attribution values in *anthropogenic*-labeled images $\mathcal{X}^{(y=0)}$, the q-quantile should equal the $(1-q)$-quantile of attribution values in *protected*-labeled images $\mathcal{X}^{(y=1)}$:

$$\Psi(\mathcal{X}^{(y=0)}, q) \stackrel{!}{=} \Psi(\mathcal{X}^{(y=1)}, 1-q) \tag{10.3}$$

Using the training data, we find this threshold at $\Psi_Q = 2.02e-5$ with $q = 96$ % and use it for accuracy and change detection computations.

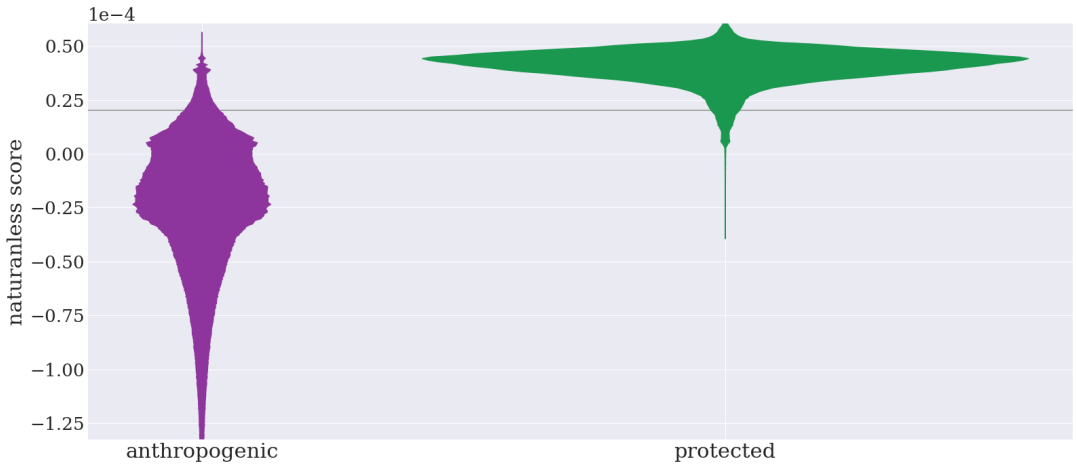


Figure 10.5: Distributions of our naturalness predictions (y-axis) on the training regions, grouped by label (x-axis). They look similar for the test regions. The violins are scaled to have equal area and the grey horizontal line indicates the quantile intersection threshold $\Psi_Q = 2.02e-5$.

Quantile Intersection Threshold Ψ_Q

The quantile intersection threshold is used to distinguish between natural- and anthropogenic-correlated attribution values. It is observed at $\Psi_Q = 2.02e-5$, corresponding to the 96^{th} percentile ($q = 96$ %).

10.3.5 Change Detection

We assess whether anthropogenic changes occur in regions initially classified as natural by applying an offline change detection test for a shift in the mean (Basseville and Nikiforov, 1993). We allow potential change points τ between the years 2019/2020, 2020/2021, 2021/2022, and 2022/2023. For each candidate change point, we compute the mean naturalness score and its standard deviation before ($-$) and after ($+$) the change. The difference between the two mean scores yields the change:

$$\Delta\Psi(\tau) = \Psi_+(\tau) - \Psi_-(\tau) = \text{mean}(\{\Psi_t : t > \tau\}) - \text{mean}(\{\Psi_t : t < \tau\}) \quad (10.4)$$

From the standard deviations, we compute the pooled standard deviation, which represents a weighted estimate of the common variability:

$$\sigma(\tau) = \sqrt{\frac{(n_- - 1)\sigma_-^2 + (n_+ - 1)\sigma_+^2}{n_- + n_+ - 2}} \quad (10.5)$$

The quantities σ_-, σ_+, n_-, and n_+ all depend on τ. Here, n_- and n_+ denote the number of years before and after the change point, respectively. We subject the change to a t-test (Student, 1908) and compute the t-value as:

$$t = \frac{\Delta\Psi(\tau)}{\sigma(\tau)\sqrt{\frac{1}{n_-} + \frac{1}{n_+}}} \quad (10.6)$$

The denominator represents the standard error of the change. To account for year-to-year fluctuations at the 95 % confidence level, we require:

$$t \overset{!}{>} 2.015 \quad (10.7)$$

This value is obtained from the 95^{th} percentile of the Student's t-distribution for a one-sided test with 5 degrees of freedom (corresponding to the 7 years).

To reduce the impact of local fluctuations at the native 10-meter resolution, we report changes at a coarser 100-meter grid. This is implemented by first averaging the annual naturalness maps to 100-meter resolution and then applying the change detection procedure described above. To find anthropogenic changes

in regions initially classified as natural, we use the quantile intersection threshold of $\Psi_Q = 2.02e-5$ identified in Section 10.3.4. It must be exceeded before the change point and fall below it afterwards:

$$\Psi_- \overset{!}{>} \Psi_Q \quad \text{and} \quad \Psi_+ \overset{!}{<} \Psi_Q \tag{10.8}$$

We do not consider changes that do not exceed the required t-value. Figure 10.7 (page 101) presents two example events in close-up. We do not display the complete map because the detected changes usually occur in areas too small to be visible at the scale of Fennoscandia.

Resulting Products

Our resulting products are:

- A **naturalness map** representing the midpoint year, 2021, at a spatial resolution of 10 meters (Figure 10.6, page 100).
- A **change map** indicating anthropogenic changes between 2018 and 2024 in regions initially classified as natural, at spatial resolution of 100 meters. Close-up views are shown in Figure 10.7 (page 101).

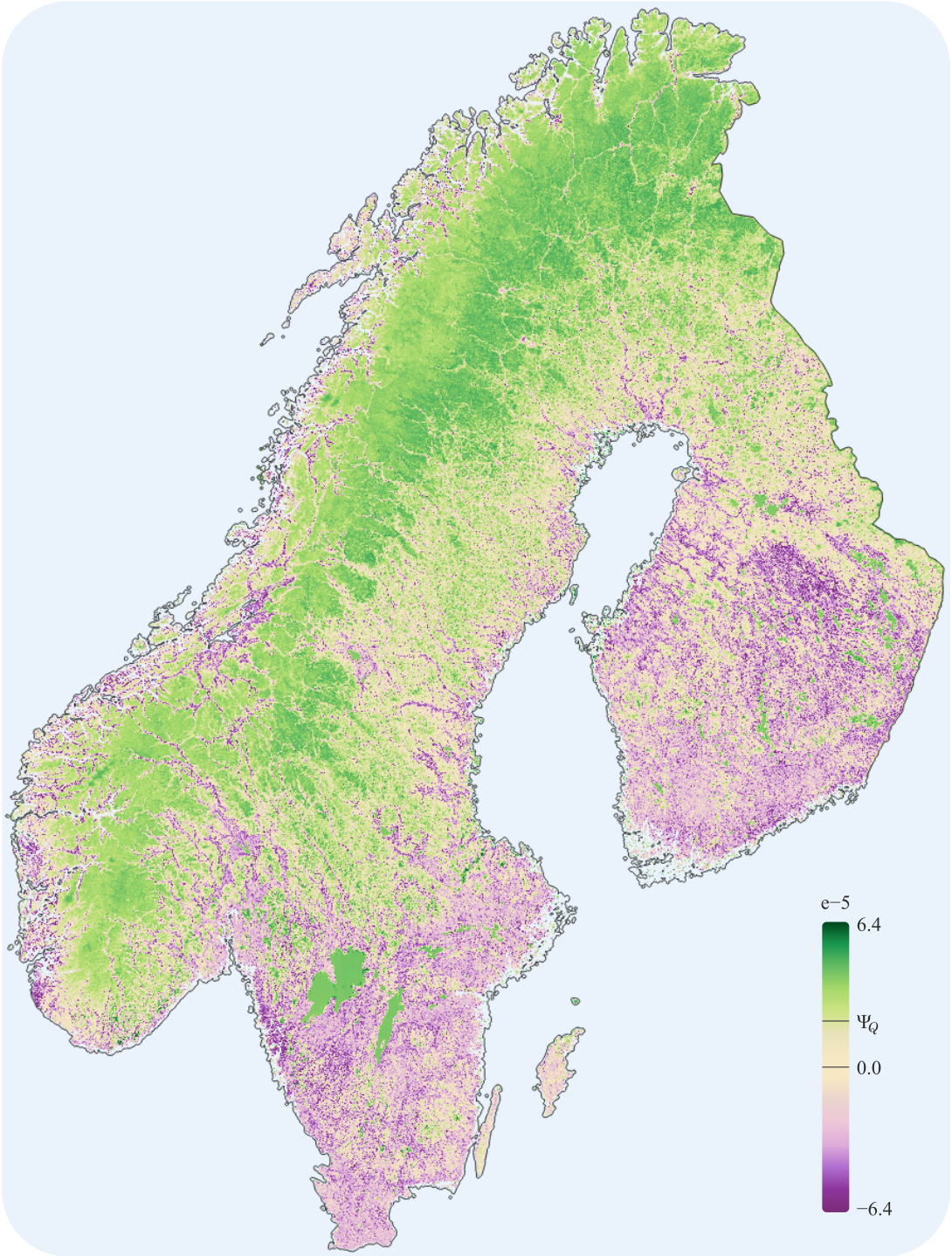
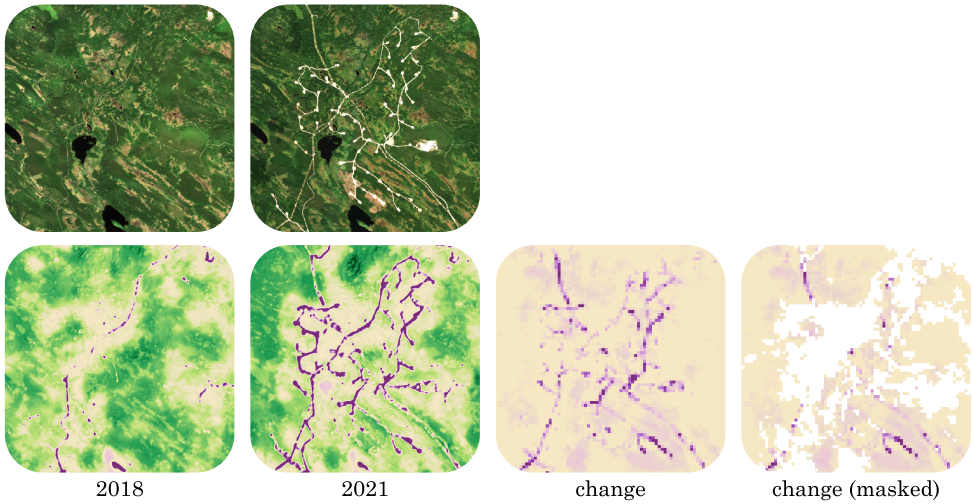
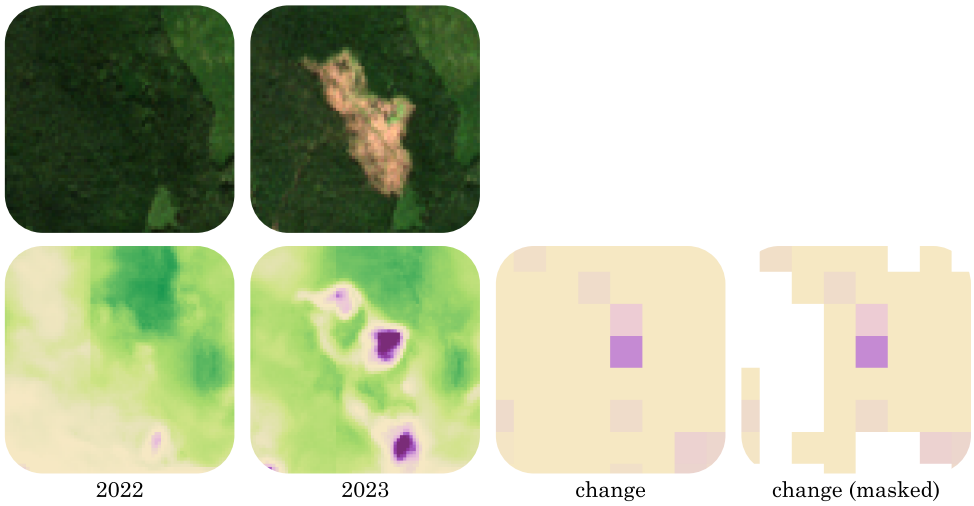


Figure 10.6: Our predicted naturalness map of Fennoscandia, representing the midpoint year, 2021.



(a) The Markbygden Wind Farm in Sweden is the largest wind farm in Europe. Parts of it were under construction between 2018 and 2021. The area shown spans about 8,000 × 8,000 meters.



(b) Deforestation took place in the observed area between 2022 and 2023, spanning approximately 500 × 150 meters. The anthropogenic impact is reflected with a particular emphasis on the edges, as investigated in Section 11.1.2.

Figure 10.7: Two anthropogenic events: (a) the construction of a wind farm and (b) a deforestation area. Shown are the Sentinel-2 images and the corresponding naturalness maps before and after the events, along with the change detection results. The masked change detection on the right excludes regions that were not initially classified as natural. ■ negative, ■ zero, ■ positive, □ masked.

Chapter 11

Results and Evaluation

In this section, we evaluate our naturalness map of Fennoscandia and the temporal changes between 2018 and 2024. We further contextualize our dataset presented in Section 10.2 and explain specific behaviors of the models. For this, we use additional data including CORINE Land Cover (European Environment Agency, 2018), Human Modification (Kennedy et al., 2019), and Dynamic World (Brown et al., 2022).

11.1 Evaluation of the Naturalness Map

11.1.1 Distributions in Protected and Anthropogenic Regions

Figure 10.5 (page 97) shows the distributions of naturalness scores in our *training* regions, separated by the labels *anthropogenic* and *protected*. The distributions in the *test* regions look very similar, so we omit an additional figure. As in the training regions, 96 % of the naturalness scores in *anthropogenic* test regions fall below Ψ_Q, while 96 % of those in *protected* test regions exceed it. This corresponds to a per-class accuracy of 96 % for the test regions.

Figure 11.1 illustrates distributions of the naturalness scores within protected areas, grouped by IUCN categories: strict nature reserves (Ia), wilderness areas (Ib), national parks (II), habitat or species management area (IV), and protected landscape or seascape (V). The amount of anthropogenic management allowed increases with increasing category numbers. As expected, we observe a decreasing trend in naturalness scores as the category number increases. However, there is considerable overlap between the distributions. Also, categories IV and V show roughly similar distributions; and category Ia contains a relatively high number of negative naturalness scores, likely due to its often very small area sizes, which make it more susceptible to surrounding anthropogenic influences.

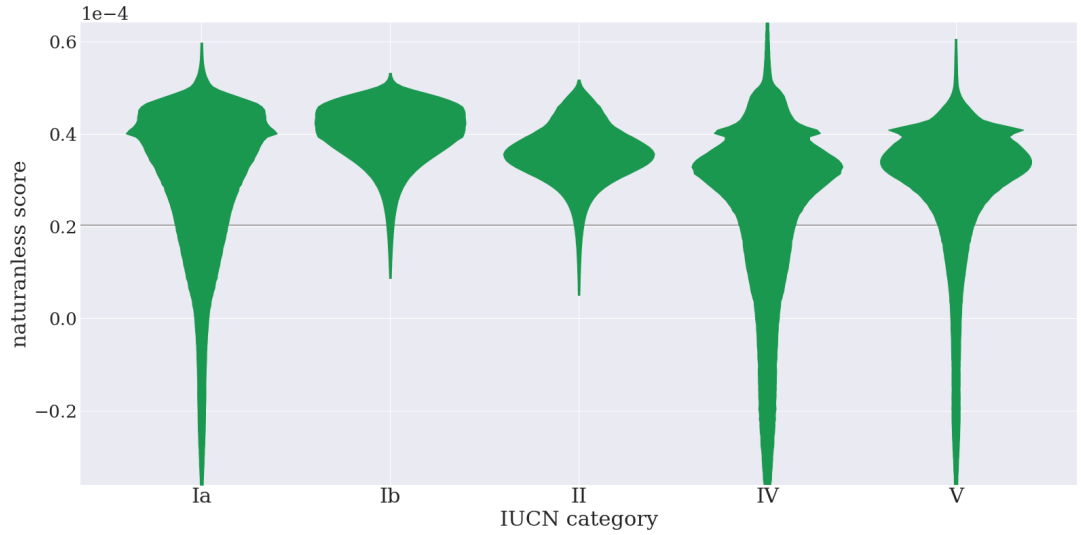


Figure 11.1: Distributions of our naturalness predictions (y-axis) grouped by IUCN categories (x-axis). The violins are scaled to have equal width and the grey horizontal line indicates $\Psi_Q = 2.02e - 5$.

Key Points: Protected and Anthropogenic Regions

- We observe a per-class accuracy of 96 % evaluating the attribution values within *anthropogenic* and *protected* test regions.
- We observe a decreasing trend of naturalness scores with increasing IUCN category numbers, which indicate a higher allowance for anthropogenic management.

11.1.2 Behavior of the Models

Attribution maps illustrate a model's decision-making process and therefore do not directly measure naturalness or anthropogenic presence. We observe some related behaviors.

Attribution Patterns

The attribution patterns of our five models vary by class. In natural regions, the models tend to focus on spatially broad structures, resulting in smooth and uniform attributions. In contrast, anthropogenic regions exhibit more heterogeneous attributions, with strong contributions from edges that lead to pronounced fluctuations between negative values and values near zero. This behavior is illustrated in Figure 11.2.

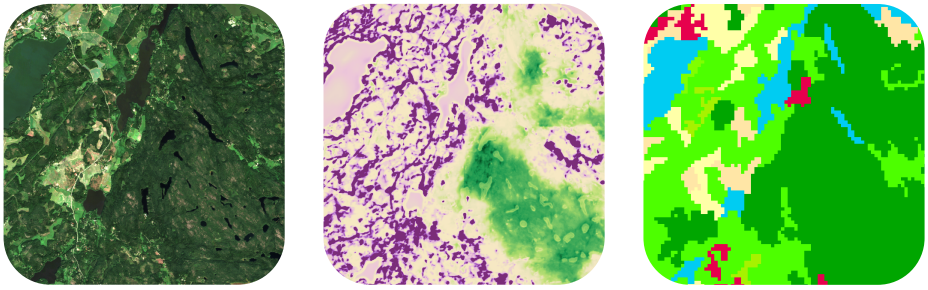


Figure 11.2: Our naturalness map tends to exhibit homogeneous values within natural areas, but emphasizes edges in anthropogenic regions. This behavior is exemplified by the displayed region near Helsinki in southern Finland, which includes parts of Nuuksio National Park. Naturalness map: ■ negative, ■ zero, ■ positive. CORINE Land Cover: ■ Urban fabric, ■ ■ Agricultural areas, ■ ■ Forests, ■ Water bodies.

Receptive Field

Another model-related characteristic is the influence of anthropogenic areas on surrounding pixels, extending 60 to 70 pixels in each direction, equivalent to 600 to 700 meters. This effect arises from the receptive field of the models — the spatial extent of the input image that a neuron is sensitive to and integrates information from. In UH-Net, the receptive field mainly results from the four encoding steps, each consisting of a pooling layer and two convolutions with a kernel size of 3. It becomes most apparent where natural and anthropogenic regions meet, for example, along streets (see Figure 11.3). The transitions are typically marked by areas with relatively neutral values.

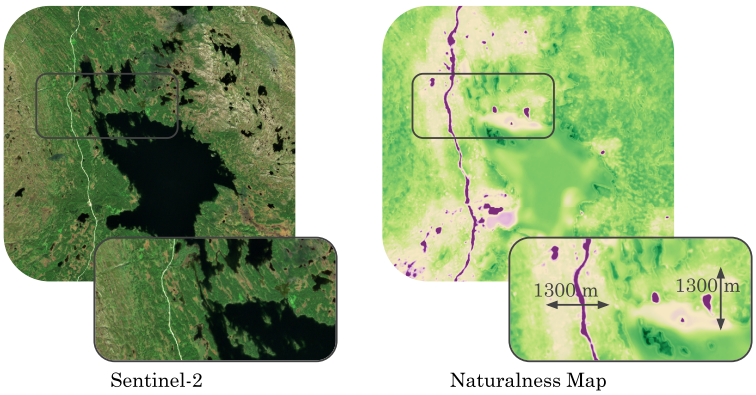


Figure 11.3: The receptive field of the models is apparent at around 130 pixels (1,300 meters), particularly where natural and anthropogenic regions meet, for example, along roads. This figure shows Route 705 close to the Skardsfjella and Hyllingsdalen conservation area in Norway. ■ negative, ■ zero, ■ positive.

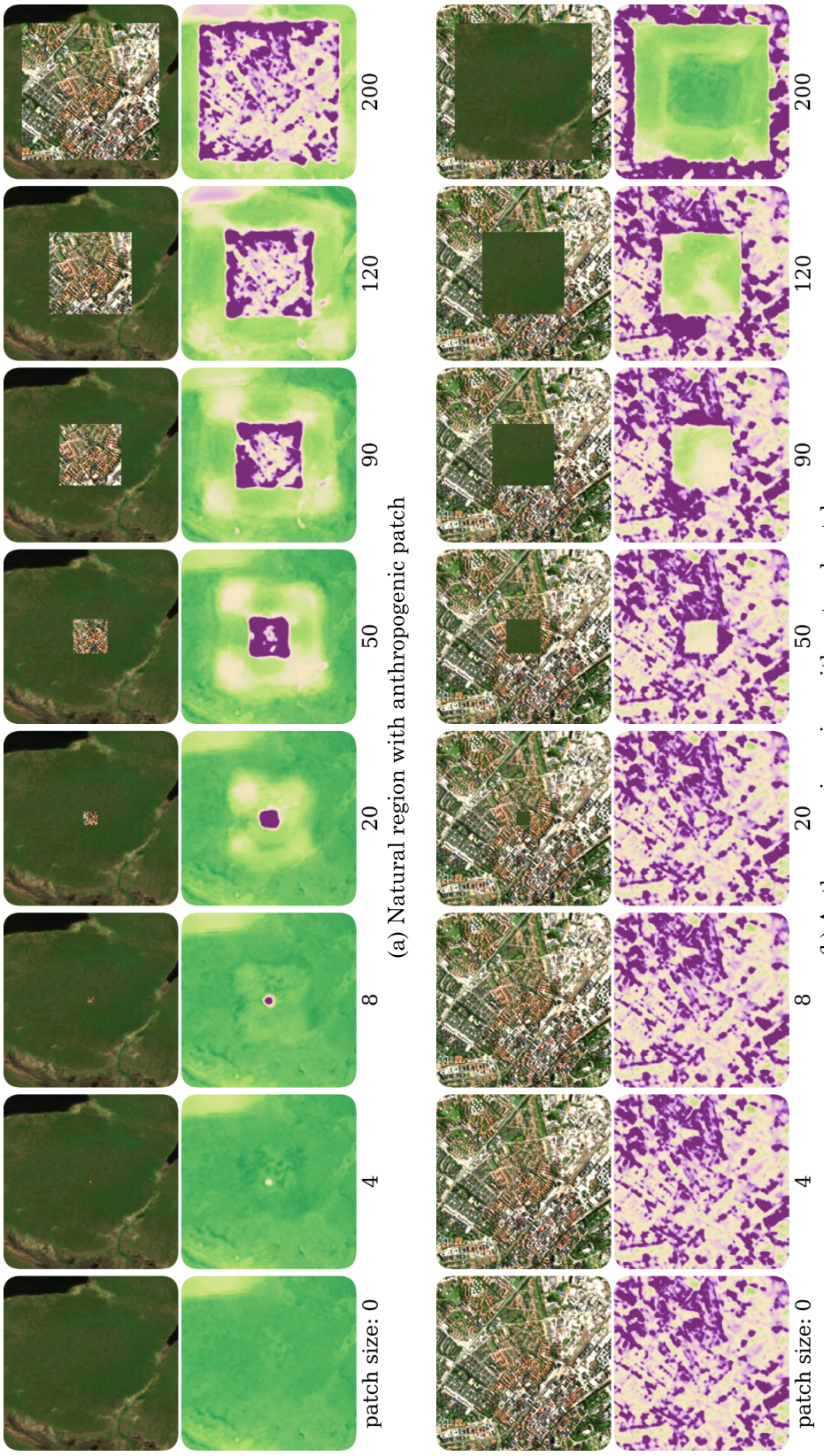


Figure 11.4: Inserting (a) an anthropogenic patch into a natural region and (b) a natural patch into an anthropogenic region. The anthropogenic patch is detected at a size of 4×4 to 8×8 pixels, while the natural patch is detected with positive values at approximately 90×90 pixels. ■ negative, ■ zero, ■ positive.

Model Responses to Inserted Patches

We investigate how the models respond to small influences by selecting a natural broad-leaved forest in Swedish Lapland and an urban environment in the city of Uppsala. In each landscape, we insert a patch of the other landscape, varying in size. The perturbed images are evaluated using the averaged harmonized attributions of all five models. The results are shown in Figure 11.4.

The anthropogenic patch embedded in the natural region becomes detectable at approximately 4×4 pixels, with a pronounced negative response emerging from around 8×8 pixels onward. At a size of 20×20 pixels, the receptive field becomes clearly discernible. Beyond 50 pixels, the models increasingly emphasize the edges within the anthropogenic patch. On the other hand, a natural patch placed within the urban environment begins to affect the attribution map at around 20×20 pixels, showing neutral values. Only at approximately 90×90 pixels does the natural region cause significant positive responses.

Key Points: Behavior of the Models

- Natural regions exhibit smooth and uniform attributions, whereas anthropogenic regions show heterogeneous attributions with contributions from edges.
- The receptive field influences approximately 60 to 70 pixels in each direction of a location, corresponding to 600 to 700 meters.
- Anthropogenic features can be detected at sizes starting from approximately 4×4 to 8×8 pixels. Natural features within anthropogenic regions are detected at sizes of about 90×90 pixels.

11.1.3 Distributions across Land Cover Classes

We evaluate our naturalness map with respect to land cover classes consulting the CORINE Land Cover dataset. It comprises 44 land cover classes with a minimum mapping width of 100 meters and is released every six years for Europe as part of the Copernicus program (European Environment Agency, 2018). We use the 2018 version; the 2024 version is scheduled for publication in 2026.[1] The dataset provides a three-level classification hierarchy: the coarsest level includes 5 categories, the middle level 15 categories, and the detailed level 44 categories. We use a combination of these levels, as it is best suited to the requirements of our naturalness application. We exclude the categories *Mine, dump and construction sites* (13) and *Burnt areas* (334), as these land cover types are likely to have unfamiliarly changed during our study period.

[1] https://land.copernicus.eu/en/products/corine-land-cover?tab=roadmap (accessed August 1, 2025)

Table 11.1: Customized CORINE Land Cover classes and their relative areas across all of Fennoscandia and within our training dataset, grouped by *protected* (label 1) and *anthropogenic* (label 0) categories. Values above 1 % are bold.

Customized CORINE Land Cover Class			Fennoscandia	Training Data	
				Protected	Anthropogenic
■	11	Urban fabric	0.8 %	0.3 %	**8.7** %
■	12	Industrial, commercial, and transport units	0.2 %	0.4 %	0.5 %
■	14	Artificial, non-agricultural vegetated areas	0.2 %	0.1 %	**1.9** %
■	21	Arable land	**4.4** %	0.3 %	**7.5** %
■	22	Permanent crops	0.002 %	0 %	0.1 %
■	23	Pastures	0.2 %	0.4 %	**2.4** %
■	24	Heterogeneous agricultural areas	**2.5** %	0.02 %	**7.0** %
■	311	Broad-leaved forest	**6.0** %	**14.3** %	**9.2** %
■	312	Coniferous forest	**35.5** %	**38.3** %	**33.9** %
■	313	Mixed forest	**8.0** %	**3.5** %	**8.2** %
■	321	Natural grasslands	0.2 %	**4.7** %	0.3 %
■	322	Moors and heathland	**6.9** %	**7.5** %	**1.8** %
■	324	Transitional woodland-shrub	**5.4** %	**2.1** %	**1.9** %
■	331	Beaches, dunes, sands	0.007 %	0.9 %	0.1 %
■	332	Bare rocks	**2.3** %	**1.8** %	0.1 %
■	333	Sparsely vegetated areas	**7.5** %	**2.7** %	0.5 %
■	335	Glaciers and perpetual snow	0.3 %	**9.9** %	0 %
■	4	Wetlands	**6.2** %	**6.1** %	**1.7** %
■	5	Water bodies	**13.4** %	**6.7** %	**14.3** %

Table 11.1 presents our customized version of the CORINE Land Cover classes including their relative areas across Fennoscandia and within our training dataset. The CORINE Land Cover map for Fennoscandia is provided in Appendix A.2, Figure A.6 (page 144).

Distributions

For evaluation, we reduce the scale of our naturalness map from 10 to 100 meters computing the mean value, so that our map has the same scale as the CORINE Land Cover map. Figure 11.5 illustrates the distributions of our predictions for each land cover class. Some classes can be clearly assigned an anthropogenic

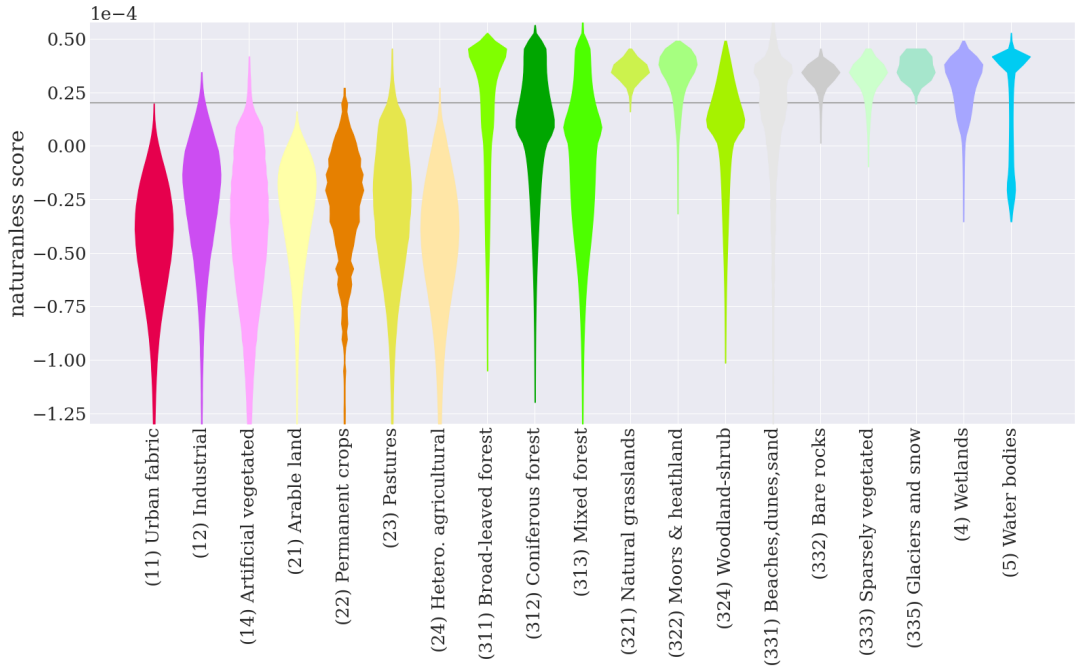


Figure 11.5: Distributions of our naturalness prediction (y-axis) grouped by CORINE Land Cover (x-axis). The violins are scaled to have the same maximum width and the grey horizontal line indicates $\Psi_Q = 2.02e - 5$.

designation, specifically classes 11, 12, 14, 21, 22, 23, and 24. At the locations corresponding to these classes, 99.7 % of the values in our naturalness map are below Ψ_Q, indicating anthropogenic influence. In contrast, classes 321, 322, 332, 335, and 4 are inherently natural or clearly not associated with human habitation. For these, 81 % of our predictions are above Ψ_Q, reflecting natural conditions. This corresponds to a per-class accuracy (natural vs. anthropogenic) of 91 %. For the three forest classes, as well as woodland-shrub and water bodies, both anthropogenic- and natural-correlated values occur regularly.

Counterintuitively, *Industrial* yields higher values than *Urban fabric* or the agricultural classes. We observe that areas with extensive asphalt coverage, large-roofed buildings, airports, and certain open-pit mines often even exhibit positive values. It is likely that the models treat such regions similarly to bare land. Another explanation is that industrial areas are also found to a significant amount within the protected regions of our training data (see Table 11.1).

Key Points: CORINE Land Cover

- We observe a per-class accuracy of 91 % evaluating the attribution values for clearly anthropogenic and natural land cover classes.

11.1.4 Comparison with Human Modification

Kennedy et al. (2019) provide a global map of Human Modification for the year 2016, which is modeled using 13 anthropogenic stressors grouped into five categories:

- **Human settlement:** population density, built-up areas
- **Agriculture:** cropland, livestock
- **Transportation:** major roads, minor roads, two tracks, railroads
- **Mining and energy production:** mining, oil wells, wind turbines
- **Elecrical infrastructure:** powerlines, nighttime lights

The map has a spatial resolution of 1 km, with Human Modification values ranging from 0 (low) to 1 (very high). The authors define four categories of Human Modification M_H based on the following value ranges:

- **Low** ($0.0 \leq M_H \leq 0.1$):
  This range is defined to include 50 % of the global land area.
- **Moderate** ($0.1 < M_H \leq 0.4$)
- **High** ($0.4 < M_H \leq 0.7$):
  The threshold of 0.4 corresponds to observed species responses to habitat loss and aligns with areas of low-intensity agriculture.
- **Very high** ($0.7 < M_H \leq 1.0$)

A map of Human Modification for Fennoscanida is shown in Appendix A.2, Figure A.7 (page 145). In comparison to our naturalness scores, the following differences emerge:

- Human Modification and naturalness represent opposing concepts, which results in an inverse relationship. For clarity, we invert the axes for Human Modification in our visualizations.
- The Human Modification index has a clearly defined range from 0 to 1, whereas our naturalness scores are attribution-based and do not exhibit a fixed numerical range.
- The Human Modification index and our naturalness scores stem from fundamentally different data sources and methodologies. While some correlation is to be expected, so too are differences.

Human Modification in Protected and Anthropogenic Regions

Figure 11.6 illustrates Human Modification's distributions in our test regions, grouped by the labels *anthropogenic* and *protected*. They look similar to the distributions of our naturalness scores (Figure 10.5) with values in protected regions falling within a narrow range, while those in anthropogenic regions span a much broader range. 92 % of the Human Modification values within protected regions fall within the *low* Human Modification range (0 to 0.1). 9 % of the values

within anthropogenic regions fall within the *very high* Human Modification range (0.7 to 1). Less than 1 % of the values are higher than 0.8.

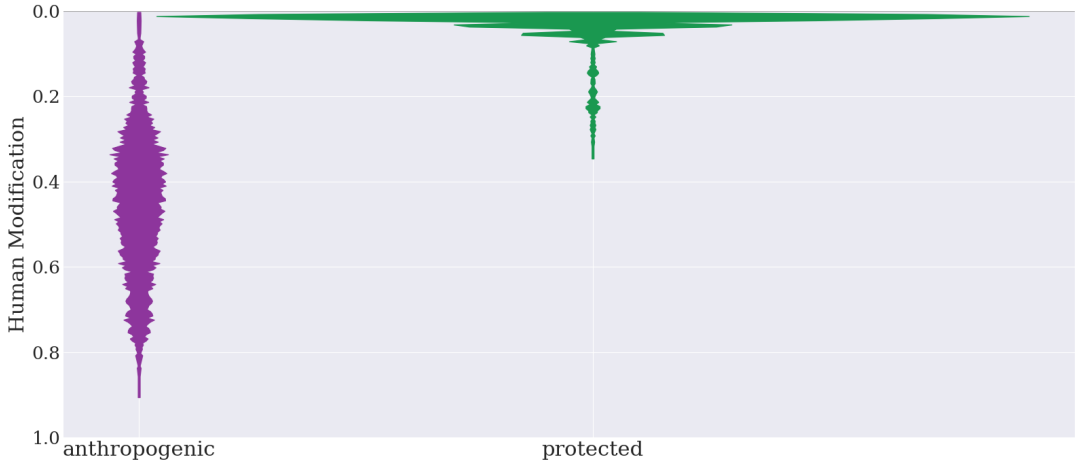


Figure 11.6: Distributions of Human Modification (y-axis) on the test data, grouped by label (x-axis). The violins are scaled to have equal area.

Human Modification and CORINE Land Cover

Human Modification's distributions grouped by the CORINE Land Cover classes are largely comparable to those of our naturalness scores (Figure 11.7 vs. Fig-

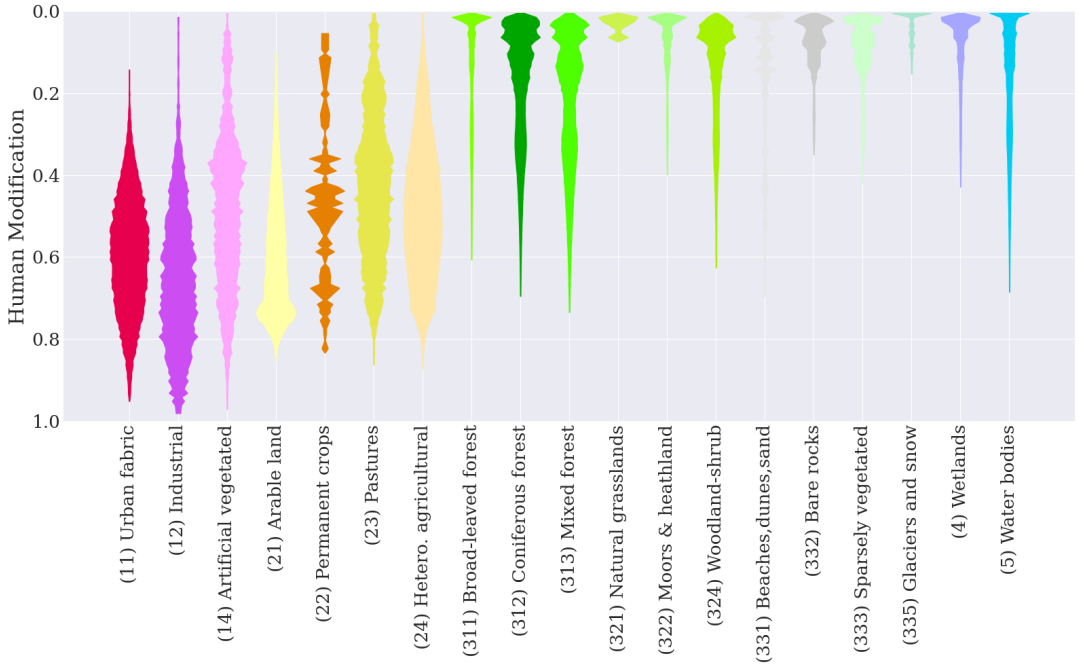


Figure 11.7: Distributions of Human Modification (y-axis, inversed) grouped by CORINE Land Cover (x-axis). The violins are scaled to have the same maximum width.

ure 11.5). We observe differences for *Industrial*, which is more anthropogenic than *Urban fabric* according to Human Modification. Similarly, *Arable land* is considered significantly more anthropogenic according to the Human Modification index. *Permanent crops* are rare in Fennoscandia (0.002 %), suggesting that the natural ratings assigned by Human Modification may result from its coarse spatial resolution of 1 km. Forests and *Woodland-shrub* are generally assessed as more natural compared to our predictions.

Correlation

We evaluate the correlation between our naturalness scores and the Human Modification index. Since the two are derived from fundamentally different methods, we do not expect a linear relationship and therefore use the Spearman rank correlation (Spearman, 1904), which measures the monotonic relationship between two variables by comparing their rank orders. Unlike the Pearson correlation, it does not assume linearity. The Spearman rank correlation ranges from -1 (perfect negative monotonic relationship) to 1 (perfect positive monotonic relationship). For this evaluation, we rescale our naturalness map from 10 meters to 1 km by computing the mean value, ensuring that it matches the resolution of the Human Modification map. We observe a Spearman rank correlation of -0.75 between our naturalness scores and the Human Modification index. Their relation is visualized in Figure 11.8.

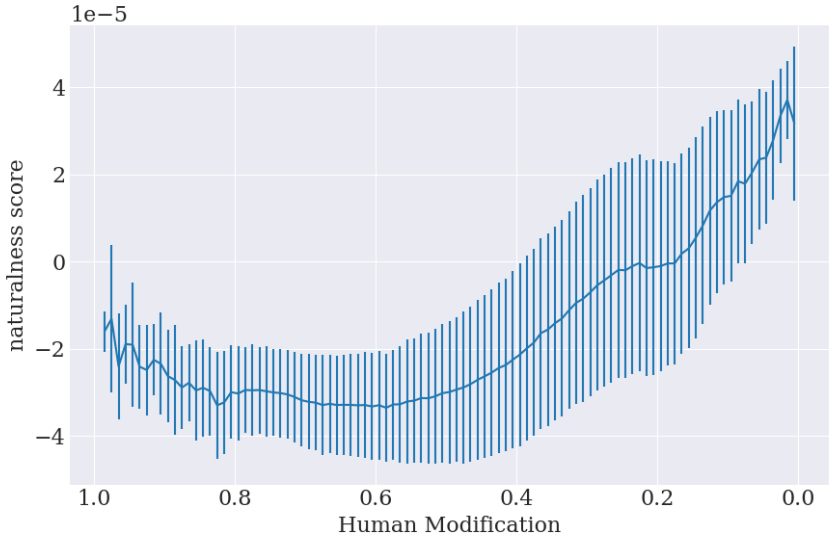


Figure 11.8: Correlation between our naturalness scores (y-axis) and the Human Modification index (x-axis, inversed). The data is binned in steps of 0.01 for Human Modification and the plot shows the mean values of our predictions with their standard deviations.

- For low to moderate Human Modification (approximately $0.0 < M_H < 0.5$), which includes 91 % of the Fennoscandian area, the correlation is approximately linear. On average, $\Psi_Q = 2.02e-5$ corresponds to a Human Modification index of about 0.1.
- For high Human Modification (approximately $0.5 < M_H < 0.8$) there is no correlation. This range covers 9 % of the Fennoscandian area and primarily includes arable land as well as outer urban areas and their adjacent forests.
- For very high Human Modification (approximately $0.8 < M_H < 1.0$), the correlation is positive, which is counterintuitive as the concepts are opposing. It mainly concerns the city centers of major cities such as Oslo, Stockholm, Gothenburg, Malmö, and Helsinki and represents just 0.2 % of the Fennoscandian area. One reason for this behavior are industrial regions, as described in Section 11.1.3.

Key Points: Human Modification

- Human Modification and our naturalness scores have a Spearman rank correlation of -0.75.
- For low to moderate Human Modification, the correlation is approximately linear. This covers 91 % of the Fennoscandian area.

11.2 Evaluation of the Change Detection

As described in Section 10.3.5, our change detection focuses on identifying anthropogenic influences in areas initially classified as natural. Figure 10.7 (page 101) illustrates two of such events. The first one shows a wind farm area that was constructed during the investigated time period. The other one shows a deforestation area. The models emphasize the *edges* of the deforested region — a behavior described in Section 11.1.2. We account for this behavior by also considering a change detection to be correct when our models identify areas adjacent to the actual event. Conversely, if no event has occurred and adjacent areas are still detected, this is classified as an incorrect detection. We exclude water areas from evaluation.

To estimate the accuracy of our change detection, we examine various randomly selected locations using different sampling strategies. We manually inspect the corresponding satellite images to assess whether anthropogenic influences can be identified. These influences primarily include deforestation, forestation, and the construction of built-ups like roads, wind farms, or buildings. The resulting confusion matrices and metrics are listed in Tables 11.2 and 11.3 on page 114.

Random Sampling of Locations

We randomly select 100 locations within regions that are classified as natural. In 85 of these, we do not observe any human influence upon inspecting the satellite images. In the remaining 15 cases, we identify anthropogenic events, specifically, 7 instances of deforestation and 8 of forestation. The performance of our change detection on these regions is illustrated in Table 11.2a. For the *no change* category, it achieves a precision of 97 % and a recall of 80 %. For the *change* category, precision is 43 % and recall is 87 %.

Due to the strong imbalance between the two categories, this random sampling strategy results in many *no change* samples. To obtain more reliable estimates for the metrics related to the *change* category, we employ additional sampling strategies.

Random Sampling of Model-Detected Changes

We randomly select 100 locations from regions where changes have been detected by our modelling. Among these, 42 show evidence of anthropogenic activity, specifically, 21 instances of deforestation, 14 of forestation, and 7 of built-up development. This yields a precision of 42 % for the *change* category. The confusion matrix is shown in Table 11.2b.

We also sample 100 random locations from regions with strong detected changes, defined as those where the naturalness score falls below the median of anthropogenic training regions: $\Psi_+ < \Psi(\mathcal{X}^{(y=0)}, q = 0.5)$. Here, we observe a precision of 69 % with a notably higher number of built-up events: 29 instances of deforestation, 7 of forestation, and 33 of built-up development. The confusion matrix is shown in Table 11.2c.

Random Sampling of Dynamic World-Detected Changes

The Dynamic World land cover product (Brown et al., 2022) provides global 10-meter resolution maps with 9 classes, at the temporal resolution of Sentinel-2 imagery, which ranges from 2 to 5 days. The included classes are *Water*, *Trees*, *Grass*, *Flooded vegetation*, *Crops*, *Shrub & scrub*, *Built area*, *Bare ground*, and *Snow & ice*. We use this product to identify regions where anthropogenic events occurred during our study period. To do so, we aggregate the land cover maps over the summer months (June to September) for both 2018 and 2024. These aggregated maps are then used to detect specific land cover changes within the study period.

Using Dynamic World, we identify deforestation as a land cover change from *Trees* to one of the classes *Shrub & scrub*, *Crops*, or *Grass*, and randomly select 100 such locations. Comparing these with satellite imagery, we find that deforestation

Table 11.2: Confusion matrices illustrating the performance of our change detection. Reported are the numbers of samples.

(a) Random sampling of 100 locations.

		Detection	
		change	*no change*
Ground truth	*change*	**13**	2
	no change	17	**68**

(b) Random sampling of 100 changes detected by the models.

		Detection	
		change	*no change*
Ground truth	*change*	**42**	-
	no change	58	**-**

(c) Random sampling of 100 strong changes detected by the models.

		Detection	
		change	*no change*
Ground truth	*change*	**69**	-
	no change	31	**-**

(d) Random sampling of deforestation areas identified using Dynamic World.

		Detection	
		change	*no change*
Ground truth	*change*	**38**	10
	no change	-	**-**

(e) Random sampling of built-up development areas identified using Dynamic World.

		Detection	
		change	*no change*
Ground truth	*change*	**58**	1
	no change	-	**-**

Table 11.3: Metrics demonstrating the performance of our change detection, reported in percentages.

	Precision	Recall	F1-Score
change			
Random	43	87	65
Model-detected changes	42	-	-
Model-detected strong changes	69	-	-
Dynamic World-detected deforestation	-	79	-
Dynamic World-detected built-up	-	98	-
no change			
Random	97	80	91

actually occurred at 48 locations. Of the 48 confirmed deforestation regions, our model detects 38 changes, corresponding to a recall of 79 % for the *change* category. The confusion matrix is shown in Table 11.2d. For the 52 locations where no change occurred, our model correctly identifies no change in 41 cases. We do not include this number in the confusion matrix because the selection of *no change* samples is biased: these regions often represent a mixture of several classes, causing Dynamic World to falsely detect a change even when clearly none occurred.

We further identify built-up development as a land cover change from *Trees*, *Flooded vegetation*, or *Shrub & scrub* to *Built area*. Of the 100 randomly selected locations, 59 truly show anthropogenic impact. Our models detect 58 of these, corresponding to a recall of 98 % for the *change* category. The confusion matrix is shown in Table 11.2e. Among the 41 locations without change, our model correctly detects *no change* in 24 cases. For the same reasons as above, we exclude these values from the confusion matrix.

Key Points: Change Detection

- For the *no change* category, we estimate a precision of 97 % and a recall of 80 %. This indicates that when *no change* is detected, it is likely to be correct.
- For the *change* category, we estimate a precision of 42 % and a recall greater than 79 %. This suggests that about 6 out of 10 detected changes are false positives; and that roughly a fifth of actual changes go undetected.

11.3 Land Cover Composition and Naturalness in Fennoscandia

Land Cover Composition

The CORINE Land Cover map for Fennoscandia is provided in Appendix A.2, Figure A.6 (page 144) and the corresponding amounts of land cover classes are listed in Table 11.1 (page 107). Coniferous forest is the most common class in Fennoscandia, particularly present in Sweden and Finland. In the Scandinavian Mountains, which primarily cover Norway and parts of Sweden, coniferous forests give way to broad-leaved and mixed forests. Other frequent landscapes include moors, heathland, wetlands, woodland-shrub areas, and, especially at higher elevations, sparsely vegetated areas. Anthropogenic land cover including built-up and arable land occur mainly in southern regions, as well as in valleys and along the fjords.

Natural and Anthropogenic Regions

Our naturalness map for Fennoscandia is presented in Figure 10.6 (page 100). Natural regions tend to be located inland and farther north, as well as in the southern parts of Norway. There is substantial overlap with the Scandinavian Mountains. Anthropogenic regions are concentrated particularly in southern Sweden, southern Finland, and along the coasts. Especially in southern Sweden and southern Finland, there are extensive areas with almost no naturalness.

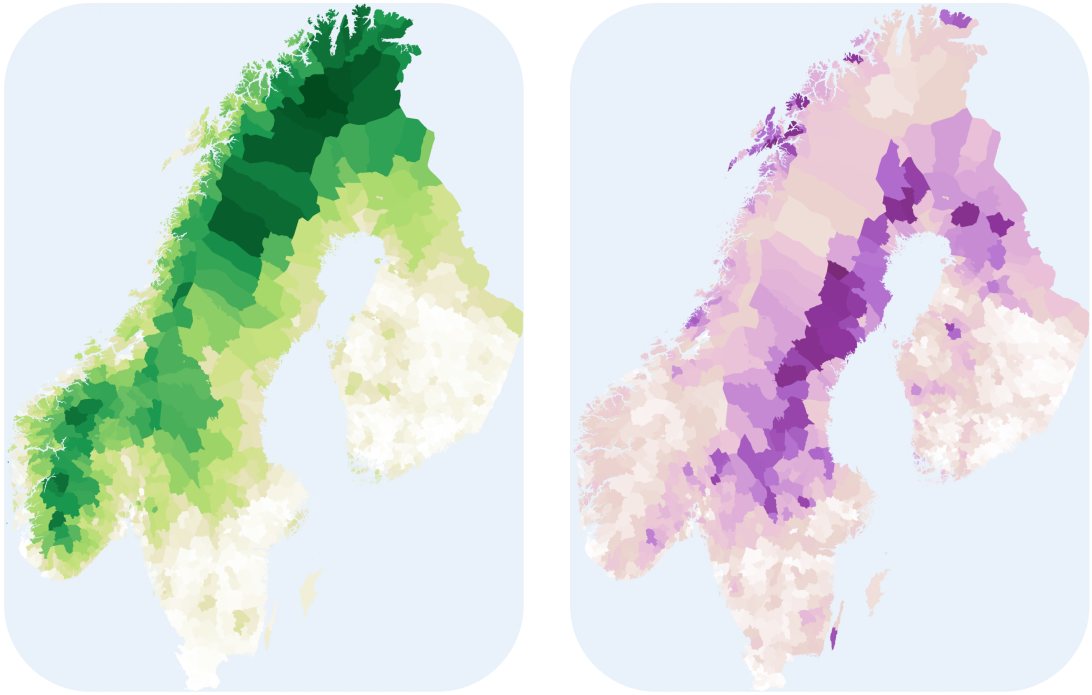
Excluding water bodies, we find that 44 % of the Fennoscandian landscapes are classified as natural ($\Psi > \Psi_Q$), while 56 % fall below this threshold. According to Human Modification, 42 % of the Fennoscandian landscapes experience *low* human modification ($M_H \leq 0.1$). Figure 11.9a subdivides Norway, Sweden, and Finland into their municipalities and illustrates the proportion of natural landscapes in each region. Four adjacent municipalities in the far north, though not along the coast, exhibit the highest proportions of naturalness, exceeding 95 %: Karasjok and Kautokeino in Norway, and Utsjoki and Enontekiö in Finland.

Anthropogenic Changes in Natural Regions

We also quantify the amount of anthropogenic changes in natural regions relative to the total land area for each municipality during the study period from 2018 to 2024, as shown in Figure 11.9b. The highest proportions are found in municipalities of eastern Sweden at intermediate latitudes, with values exhibiting double-digit percentages. These regions overlap with areas of intensive forest management, as revealed by Hansen et al. (2013). Regions with an already high degree of anthropogenic influence (e.g., southern Sweden) show low proportions, since there are few natural areas to begin with. Across all of Fennoscandia, we detect anthropogenic changes in 6 % of the total land area.

Key Points: Naturalness in Fennoscandia

- Our models classify 44 % of the Fennoscandian landscape as natural.
- Anthropogenic changes that degrade natural regions are detected in 6 % of the total Fennoscandian landscape.



(a) Proportions of natural regions

(b) Proportions of anthropogenic impact on natural regions

Figure 11.9: Proportions of natural area and anthropogenically changed area for the municipalities of Norway, Sweden and Finland. Water bodies are excluded from the analysis.

Chapter 12

Discussion

We presented a framework to assess naturalness directly from satellite imagery rather than relying on indirect, human-defined indicators used in traditional naturalness mapping approaches. In doing so, we build the AnthroProtect dataset that contains multispectral Sentinel-2 composites of protected and anthropogenic regions in Fennoscandia. Using this dataset, we train UH-Net to distinguish between natural and anthropogenic landscapes and compute high-resolution attribution maps. Our novel harmonization approach enables consistent, large-scale attribution mapping. Based on this framework, we generate a high-resolution naturalness map for Fennoscandia and detect anthropogenic changes occurring between 2018 and 2024.

12.1 Methodology

Deriving pixel-level estimates from image-level labels is conceptually related to weakly-supervised segmentation, as discussed in Section 8.6.1. However, instead of predicting discrete classes, we evaluate a continuous score and aim to rediscover patterns associated with protected or anthropogenic landscapes. The resulting maps are therefore not direct pixel-wise regression outputs with precise per-pixel predictions, but rather reflect broader spatial patterns and relative degrees of naturalness. In this sense, we focus more on uncovering learned patterns through explainable machine learning than on performing exact weakly-supervised regression; however, the distinction between the two is not clearly defined.

Harmonization

Original attribution methods are computed for fixed image sizes, where interactions and mutual influences among multiple patterns due to fully connected layers result in non-comparable patches (see Figure 10.4, page 96). Our harmonization method introduces a previously unavailable capability: attribution scores become

comparable both within large scenes and across different scenes. It establishes a direct link between features and attribution values while bypassing the fully connected layers. Because convolutional neural networks preserve consistent relationships between inputs and learned features, the harmonized attribution maps inherit this consistency and thus remain comparable.

UH-Net Architecture

The use of our UH-Net architecture enables the generation of high-resolution attribution maps. While the receptive field is negligible in the multi-label tasks of Chapter II, it becomes evident in this context (Figure 11.3, page 104). Nevertheless, our approach produces sharply defined and narrow attributions, such as the distinctly negative scores assigned to roads. This level of detail is made possible by the skip connections within the U-Net architecture, which preserve fine-grained spatial information throughout the network.

We experiment with several modified architectures and training procedures, including a UH-Net with a simple prediction head, similar to the variant described in Section 7.2, as well as a UH-Net with a mean-based head, in which the final prediction is obtained by averaging the features of the last U-Net layer. For the latter, we test multiple configurations that vary the order of activation, averaging, and loss computation. Across these variants, we observe similarly large receptive fields; however, the resulting attribution maps exhibit less level of detail. Overall, the outputs of many tested modified architectures resemble a similar effect of applying a mean filter to our naturalness map presented.

Change Detection and Trend Analysis

We focus on offline change detection of anthropogenic events within regions previously classified as natural. Applying trend analysis — reducing the naturalness maps across the seven years to a regression-based intercept and temporal trend — we do not reliably detect subtle changes of naturalness. This limitation is further discussed in Section 12.3. In fact, we find that the results of trend analysis are largely consistent with those produced by our change detection approach.

There are more sophisticated methods for trend analysis and change detection (Mudelsee, 2019; Asokan and Anitha, 2019), which could potentially improve the accuracy of our results. However, time-series analysis is not the focus of this work; therefore, we apply relatively simple tools to provide a basic demonstration of what is feasible with our proposed attribution-based mapping.

Key Points: Methodology

- Harmonizing attributions enables the previously untouched capability of making attribution scores comparable within large scenes and across different scenes.
- Despite the receptive field, UH-Net allows for distinct, narrow attributions.

12.2 Performance

Naturalness Map

Across the test data, our naturalness map achieves a per-class accuracy of 96 %. When evaluated against CORINE Land Cover classes that can be clearly assigned to either category, we obtain a per-class accuracy of 91 %. Furthermore, we observe a Spearman rank correlation of -0.75 with the Human Modification index by Kennedy et al. (2019). The negative sign arises because naturalness and human modification are opposing concepts. For low to moderate Human Modification, which covers 91 % of Fennoscandia, the relationship is approximately linear; at higher modification levels, the correlation breaks down. For protected regions across different IUCN categories, we observe expected trends in the distributions of naturalness scores.

The alignment of our predictions with multiple external datasets strongly suggests that they are plausible and meaningfully reflect underlying landscape patterns. However, evaluation performance is limited by the absence of exact ground truth. This issue begins with the training data: protected regions can contain infrastructure such as roads, while anthropogenic regions may include patches of natural vegetation. Categorical land cover classes cannot express degrees of naturalness, and classes such as forest may occur in both natural and anthropogenic contexts. Although the Human Modification product provides continuous values, it is derived from traditional indicator-based modeling and cannot capture the structural and (bio)physiological characteristics of landscapes visible in satellite imagery. Moreover, it is limited to a coarse 1 km resolution. Similarly, differences across IUCN categories offer only indirect signals of naturalness and cannot serve as precise targets. Because no dataset provides exact, high-resolution, continuous ground truth for naturalness, the true performance of our naturalness map cannot be precisely quantified. Nonetheless, there are strong indicators that our naturalness scores are plausible and meaningful.

Change Detection

Our change detection identifies anthropogenic disturbances within areas previously predicted as natural. Due to the lack of reliable ground truth for anthropogenic change, we perform manual verification at random locations using satellite imagery. For the *no change* category, we observe a precision of 97 % and a recall of 80 %. The precision for detected changes is estimated at 42 %, increasing to 69 % for strong detected changes, while recall is estimated to exceed 79 %. Thus, our change detection tends to overestimate changes. This is primarily due to shadows cast by clouds or terrain, as well as vegetative phenology. These limitations are further discussed in Section 12.3. Although the precision of 42 % for detected changes may seem modest, it can be considered relatively good given the extreme class imbalance: the number of unchanged regions vastly exceeds that of changed regions. Overestimation of changes also occurs when searching for disturbances using the land cover product Dynamic World by Brown et al. (2022). Here, precision is 48 % for deforestation-related land cover changes and 59 % for construction-related changes. For this product, the main reason for overestimation is that the predicted classes fluctuate in regions containing a mixture of multiple land cover types.

Key Points: Performance

- Our naturalness scores align with multiple external datasets, strongly indicating their plausibility.
- Our change detection tends to overestimate changes.

12.3 Limitations

Anthropogenic Attribution Patterns

We observe that the attributions in anthropogenic regions are narrow and concentrated along edges, whereas those in natural areas tend to be smoother and more extensive (Figure 11.2, page 104). This suggests that specific edges are important indicators of anthropogenic landscapes for the models, while expanse and uniformity are factors associated with naturalness. While this is a valid and comprehensible behavior for convolutional neural networks, it complicates and restricts the evaluation of anthropogenic landscapes. For example, it hinders the ability to scale the degree of anthropogenic impact. Thereby it also complicates comparisons across years, making trend analysis or change detection more difficult. Further, in small anthropogenic areas, such as deforestation sites, it becomes challenging to estimate the actual size of affected regions.

Comparability across Years and Change Detection

As outlined previously, the focus on edges as indicators of anthropogenic landscapes complicates year-to-year comparisons of such regions. Additionally, several other limitations hinder the fine-grained assessment of naturalness needed to detect subtle temporal trends. Interannual climatic variability, fluctuating snow cover, artifacts introduced by clouds, and topographic shading caused by mountains or uneven terrain reduce the temporal consistency of the satellite scenes. Similarly, variable illumination conditions can cause features such as forest roads to appear more or less prominent. A further source of variability arises from the use of satellite composites. Because cloud-free composites draw on observations from different dates each year, the dominant temporal window between June and September may differ between years, and thus the phenology of the vegetation. As a result, the harmonized attribution maps exhibit fluctuations that cannot be attributed to true land-surface processes. Our results demonstrate that hard detection tasks, specifically, offline change detection, can still be performed reliably for abrupt changes. In such cases, the magnitude of the signal is strong enough to overcome the noise inherent in the data. However, more gradual or subtle trends — such as an area becoming progressively more natural — remain statistically insignificant for our models, as the interannual variability in the imagery is often too large.

Limited Visibility of Certain Processes

Not all forms of ecological processes are detectable in satellite imagery. For instance, wildlife typically leaves no visible signature unless it directly alters vegetation structure. Likewise, understory vegetation or soil conditions are often obscured by forest canopies. Additionally, the spatial resolution of Sentinel-2 imagery inherently limits the detectability of fine-scale features. By contrast, aerial photography provides sub-meter resolution and could capture subtle patterns, offering a level of detail that is unattainable with Sentinel-2 data.

Key Points: Limitations

- Comparability across years is limited due to fluctuations in the satellite imagery and the anthropogenic attribution patterns.
- The spatial assessment of small-scale impacts is limited by edge-focussed attributions.
- Satellite imagery offers only limited visibility of certain ecological processes.

12.4 Comparison with Traditional Naturalness Mapping

Traditional naturalness mapping approaches, such as Human Footprint (Sanderson et al., 2002), its temporal extensions (Venter et al., 2016a; Mu et al., 2022), and Human Modification (Kennedy et al., 2019), rely on indicator-derived indices. These include, among others, population density, land transformation, agriculture, transportation, and electrical infrastructure. Resulting spatial resolutions are typically around 1 km. In contrast, our method derives naturalness from multispectral satellite imagery, which represents a completely different data basis. It enables a spatial resolution of 10 meters, and in theory, a temporal resolution that matches the Sentinel-2 revisit frequency (every 2 to 5 days), accounting for limitations due to cloud cover.

Indicator-Based vs. Appearance-Based Measurement of Naturalness

Traditional mapping approaches reflect human activity itself which is valuable for understanding socio-economic pressures. However, our approach enables a completely different perspective. It reflects the (bio)physical outcomes of human activity and can assess vegetation health and composition. This means our modelling can detect anthropogenic influence even when indicators are absent, and may classify regions as natural despite proximity to human activity if vegetation remains intact.

Interpretation of the Results

Traditional naturalness maps have a scale that is relatively straightforward to interpret and thresholds can be defined, for example, as having *low* or *high* anthropogenic impact. In contrast, our scores are derived from attribution values that indicate whether the models' internal representations align more closely with protected or anthropogenic training samples. This can lead to effects that require initial investigation and interpretation. Examples include the models' focus on edges in anthropogenic regions, the influence of the receptive field, and responses to natural patterns within anthropogenic areas and vice versa (Section 11.1.2). For instance, we observe that anthropogenic patterns within natural regions are detected at scales from 40 $\times$ 40 meters, while the models remain robust to natural "islands" within urban areas up to approximately 900 $\times$ 900 meters in size. These behaviors are meaningful and valid, but since they are not predefined, they must first be discovered through analysis.

Key Points: Comparison with Traditional Naturalness Mapping

- Our naturalness map reflects the (bio)physical outcomes of human activity as seen by satellites while traditional maps reflect human activity itself.
- Scores from traditional maps are straightforward to interpret while our approach leads to behaviors that require initial investigation.
- Our resolution is 10 meters, while traditional approaches typically have resolutions of 1 km. Furthermore, our approach has the potential for frequent temporal updates aligned with the Sentinel-2 revisit interval.

12.5 Naturalness in Fennoscandia

Regions free from human impact on naturalness are rare in most parts of the world. In Europe, Fennoscandia stands out for its relatively natural landscapes. Norway, Sweden, and Finland maintain high environmental regulations (Wendling et al., 2020) and invest in conservation through large protected areas. At the same time, clear-cutting remains the predominant forest management practice in coniferous forests, particularly in Sweden and Finland (Lunde et al., 2025).

The Scandinavian Mountains contain large areas with minimal human modification, which explains the high naturalness scores observed there. Similarly, remote northern municipalities contain vast forest and tundra areas with little human presence, resulting in naturalness proportions exceeding 95 %. In contrast, coastal and southern municipalities include fragmented and modified landscapes. Southern Sweden and southern Finland show the highest levels of population density, agriculture, and infrastructure. In total, our modelling classifies 44 % of the Fennoscandien landscape to be natural. This proportion aligns with the Human Modification map, where 42 % is categorized as having *low* human modification.

The municipalities with the highest proportion of detected anthropogenic changes overlap with forested areas managed through clear-cutting practices. Under clear-cutting, most or all trees within a designated stand are harvested simultaneously, after which the area is typically replanted. This practice remains controversial from a nature conservation perspective (Lunde et al., 2025). Protected areas are generally unaffected by these practices, which explains why inland regions are less strongly impacted. Across Fennoscandia we identify anthropogenically driven changes in 6 % of the total area. However, it should be noted that our change detection's precision is higher than the recall, which suggests that such estimates are likely overestimated. Due to the receptive field, anthropogenic changes affect the surrounding area within a radius of approximately 600 to 700 meters, which also contributes to an overestimation. On the

other hand, the emphasis on edges may lead to an underestimation. The proportion of anthropogenic changes is therefore subject to considerable uncertainty and should not be interpreted in isolation, but rather be used for comparison between regions, as shown in Figure 11.9b (page 117). The order of magnitude, however, appears to be reasonable: according to Global Forest Change by Hansen et al. (2013), 4.4 % of the Fennoscandian landscape experienced deforestation between 2018 and 2024.[1]

Key Points: Naturalness in Fennoscandia

- We predict that 44 % of the Fennoscandian landscape is natural, which appears to be a reasonable estimate.
- We detect anthropogenically driven changes across 6 % of the Fennoscandian landscape. While the magnitude seems plausible, this value is subject to considerable uncertainty.

12.6 Future Directions

We provide a valuable tool for naturalness mapping using satellite imagery which is, however, subject to some limitations. These relate not only to the methodology but also to the quality of the available data. In the following, we present ideas that might mitigate these limitations.

Attribution Patterns and Scale

We observe a naturalness score with a scale that is partly difficult to interpret. Employing a proxy index, such as Human Modification, to generate continuous targets might yield a more interpretable scale, although the resolution of such indices is typically low. Efforts in a similar direction are made by Ekim and Schmitt (2024). They combine low-resolution indicators with high-resolution data, specifically ESA WorldCover (Zanaga et al., 2022) and OpenStreetMap (OpenStreetMap contributors, 2017), to generate naturalness indices at 10-meter resolution as described in Ekim et al. (2021), and use them for training and testing a regression task. However, the authors do not conduct large-scale mapping.

With regard to the emphasis on edges, certain constraints during model training could help reduce this effect. For example, a simple smoothing filter could help; however, it has the disadvantage that isolated roads etc. may no longer be accurately detected. As an alternative, a loss function could be introduced that penalizes attributions focusing on edges. Combined with our harmonization

[1] Global Forest Watch. Tree cover loss in Norway, Sweden, and Finland. www.globalforestwatch.org (accessed November 26, 2025)

method, it may not even be necessary to compute attributions during training; instead, edge-based features could be directly penalized. With appropriate weighting, these features might be suppressed while still being preserved where they add value, such as isolated roads.

Comparability across Years

Several factors in satellite imagery contribute to fluctuations in naturalness scores between years. These include fluctuating snow cover, artifacts introduced by clouds, topographic shading, variable illumination conditions, interannual weather variations, as well as image composites with differing dominant temporal windows. To ensure that the harmonized attributions still remain consistent across years, the feature vectors from the same location should be similar for different years. We enforce this by applying a contrastive loss following the one used in Siamese-networks (Hadsell et al., 2006). We construct positive pairs from identical locations across different years and negative pairs from random location–year combinations. While we observe only a minor improvement, the method introduces artifacts in certain land cover classes. However, another contrastive loss function might be found without this disadvantage.

Using time series instead of composites could significantly improve accuracy, but this would require substantial changes to both the data and the overall methodology. Including weather information could also help the model align vegetation appearance with corresponding weather conditions. An alternative and highly current approach might be to use data embeddings derived from foundation models as input data. For example, AlphaEarth (Brown et al., 2025) provides global 64-dimensional embeddings at a 10×10 meter resolution, generated from a wide range of data sources. These embeddings may overcome the limitations of Sentinel-2 composites and capture information on naturalness more consistently, with less year-to-year fluctuations and variability.

Global Implementation

Machine learning models generally show a low generalization ability if the test data has a significantly different distribution compared to the training data. Thus, outside of Fennoscandia, our trained models are likely not applicable — particularly if the region has distinctly different landscapes or ecosystems such as savannas or tropical forests. One could set up a training dataset similar to ours for the biogeographic region(s) of interest. However, the low density of strictly protected, natural areas in many biogeographic regions makes it challenging to create a suitable dataset. We have contributed to a global dataset, MapInWild, which follows a similar concept to AnthroProtect (Ekim et al., 2023). However,

we have not predicted or evaluated a global map of naturalness, for the reasons outlined above.

Application

The promotion of natural places is an important and critical task as stated by the Sustainable Development Goals (SDGs) of the United Nations (2021). SDG 15 focuses on protecting, restoring, and promoting sustainable land use, including forest management and biodiversity conservation. Being able to analyze the extent and characteristics of natural areas, even if approximate, is a valuable tool. Analyses over time can allow for the monitoring and tracking of conservation and renaturation efforts. As such, our approach could be utilized for policy-making and land use planning.

Compared to traditional naturalness mapping, our method offers both advantages and limitations. However, the key difference lies in what both methods measure: traditional indices capture human presence and infrastructure, which is valuable for understanding socio-economic pressures; on the other hand, our method measures the (bio)physical expression of these pressures in vegetation and land surface structure. Combining both approaches could offer a more comprehensive view of ecosystem conditions at both large and small scales. Moreover, discrepancies between the two methods may reveal interesting cases — for example, visually intact landscapes that are nevertheless subject to significant anthropogenic pressures.

Key Points: Future Directions

- Employing a proxy index to generate continuous targets might yield a more interpretable scale.
- A loss penalizing edge-based features might suppress the observed attribution patterns in anthropogenic regions.
- Using embeddings derived from foundation models may overcome the limitations of Sentinel-2 composites.
- The low density of natural areas in many biogeographic regions makes it challenging to create a global dataset.
- Our tool could be utilized for policy-making and land-use planning.
- Combining our approach and traditional naturalness mapping could offer a more comprehensive view of ecosystem conditions.

Part IV

Conclusion

Conclusion

This thesis explores methods to improve the interpretability and consistency of explanations for convolutional neural networks (CNNs), with a focus on land cover classification. To this end, experiments are conducted using three CNN architectures, three datasets — two dedicated to land cover classification — and ten attribution methods. Finally, a novel approach for mapping naturalness from satellite imagery is presented, building on the developed methods.

Improving Explanations of CNNs

Learned feature vectors play an important role in the decision-making process of CNNs, as they represent the patterns the network has learned to solve a given task. Similar feature vectors indicate similar semantic content, and deeper layers capture increasingly complex semantic representations. To our knowledge, no previous study has systematically compared multiple attribution methods at both the input and deeper layers. This gap likely exists because some methods were originally designed for input-level analysis, while others are intended for deeper layers. We address this gap and our experiments show that attributions from deeper layers generally provide more intuitive explanations across all methods.

Due to the low spatial resolution in deeper layers of common CNN architectures, we introduce a novel architecture, UH-Net. It combines a U-Net with a prediction head, enabling an intermediate layer that produces high-resolution, semantically rich attribution maps. In our experiments, UH-Net yields the most interpretable feature maps for human understanding. However, this may come at the cost of reduced classification accuracy, as we observe with the Caltech 100 dataset.

We further address the challenge that different attribution methods often yield inconsistent explanations by proposing a novel harmonization technique. It builds on insights from theory of mechanistic interpretability — specifically, the idea that similar feature vectors encode similar semantics. Accordingly, we compute the harmonized attribution of an (unseen) feature vector as the local average of attributions within the feature space of the training data. This approach results in more global explanations that are easier for humans to interpret. By ap-

plying harmonization, we achieve greater consistency across attribution methods, making the choice of method less critical. The effectiveness of our harmonization approach is constrained by the degree of entanglement in the learned feature space. For this reason, it is not effective at the input level.

Although Gradients×Features is one of the simplest attribution methods, we find it performs well after harmonization. Grad-CAM already aligns closely with feature vectors even before harmonization; however, its explanations can fail under certain conditions. Our proposed method, Feature-specific Grad-CAM, successfully addresses this limitation. Sliding Window Occlusion suffers from limited resolution, a drawback effectively mitigated by harmonization. Our method, Feature-specific Occlusions, overcomes this limitation even prior to harmonization. Gradient SHAP and DeepLift SHAP improve upon Integrated Gradients and DeepLift, respectively, by using baselines that reflect the training data rather than zero. Layer-wise Relevance Propagation is equivalent to Gradients×Features under the z-Rule, and in our experiments, it yields nearly identical results also for the Epsilon Rule.

In future work, our harmonization technique could be successfully integrated into weakly-supervised segmentation approaches that rely on attribution methods. It may also be applicable to Vision Transformers, provided that the attention mechanisms are interpretable, for example, through the use of registers (Darcet et al., 2024).

Mapping Naturalness in Fennoscandia

Territorial protection can provide important ecological and social benefits, and there are urgent, pragmatic reasons to identify where the positive characteristics and ecological functions associated with naturalness are present and able to thrive. Traditional naturalness mapping approaches rely on indicators of anthropogenic stressors such as settlement, agriculture, transportation, and electrical infrastructure. However, they are often limited by low spatial resolution and do not account for vegetation health or composition, which restricts the depth of ecological insight they can provide.

We therefore propose a naturalness mapping approach based on satellite imagery. To this end, we introduce the AnthroProtect dataset, which contains Sentinel-2 imagery of both protected and anthropogenically influenced regions in Fennoscandia. Training UH-Net on this classification task enables the generation of high-resolution attribution maps. By harmonizing these maps, we achieve consistent large-scale naturalness mapping across the Fennoscandian region. When comparing our naturalness map to test regions and land cover classes, we observe per-class accuracies exceeding 90 %. Additionally, we find a correlation of -0.75 with the Human Modification index (Kennedy et al., 2019).

We also detect anthropogenic changes between 2018 and 2024 within natural regions and evaluate the results at randomly selected locations. While these results are comprehensible, comparability across years is limited by fluctuations in satellite composites and certain behaviors of the models. We identify additional limitations related to the (continuous) scale of naturalness scores, which supports only coarse comparisons. Several approaches that could help address these limitations are discussed, including the use of proxy indices as targets, the incorporation of additional loss functions, and the use of embeddings from foundation models.

Our modelling estimates that approximately 44 % of Fennoscandia consists of natural landscapes. Anthropogenically driven changes are detected in 6 % of the total area, primarily corresponding to regions affected by clear-cutting practices. The ability to analyze the extent and characteristics of non-anthropogenic areas, even approximately, is a valuable tool for tracking conservation areas, informing policy-making, and supporting land-use planning. Combining our approach with traditional methods could provide a comprehensive view of ecosystem conditions.

Part V

Appendix

A.1 Improving Explanations of CNNs

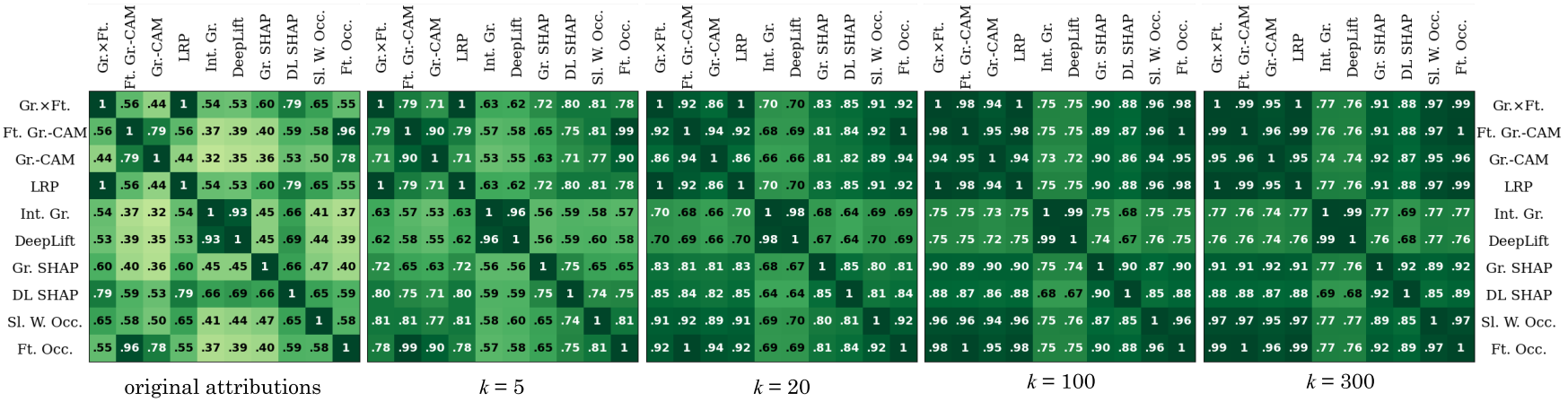


Figure A.1: Pearson correlations between attribution methods for DFC2020 and one of the ten UH-Net models (deep layer) for **varying numbers of nearest neighbors**. The parameter is quite robust for values above $k = 20$. For our experiments we have chosen $k = 100$.

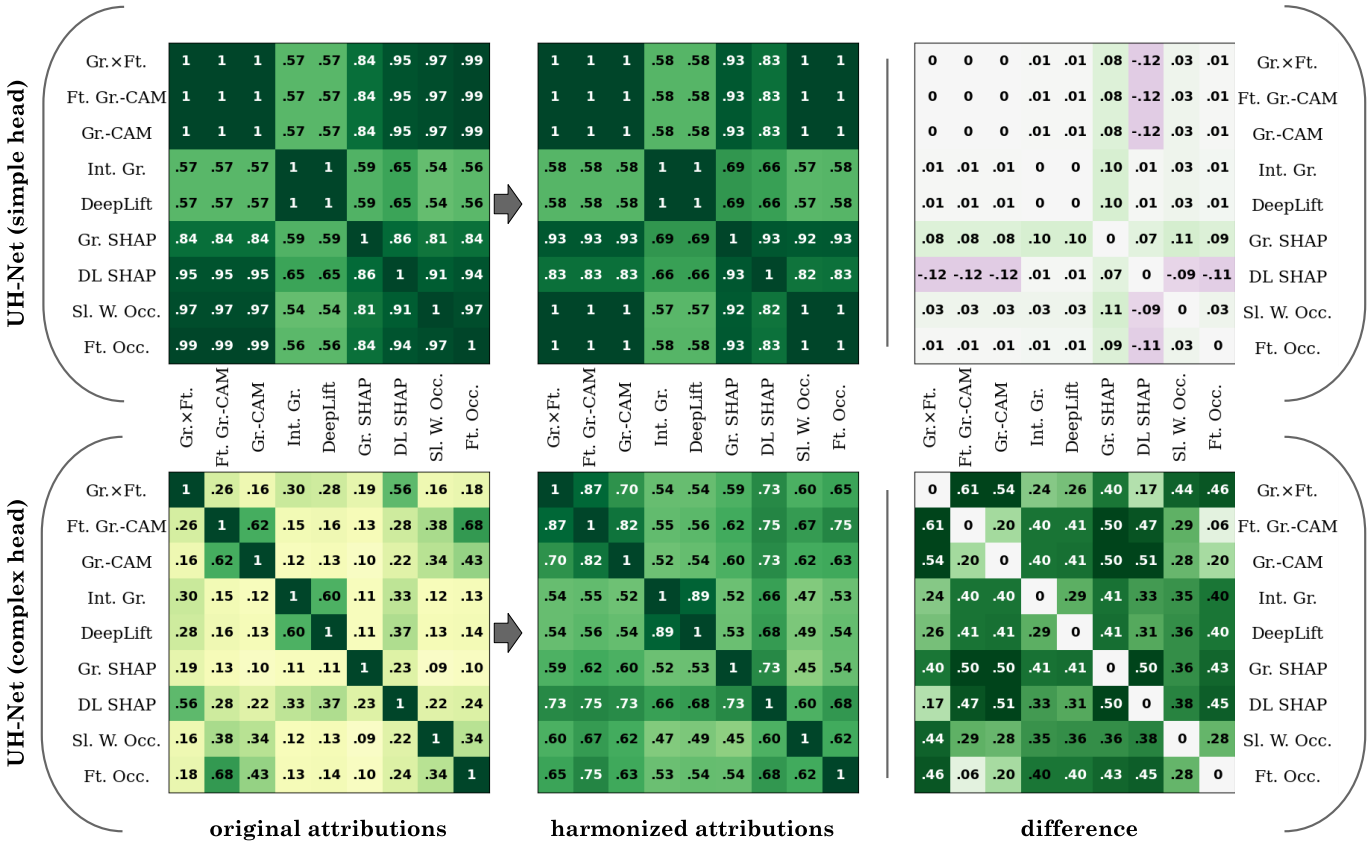


Figure A.2: Pearson correlations between attribution methods for DFC2020 and the deep layers of two **UH-Net variants**: Simple head (top) and complex ResNet-18 head (bottom). The values are averaged over ten models. From left to right: coefficients for the original attributions, the harmonized attributions, and their difference.

Table A.1: Metrics comparing the predominant class attributions with the segmentation ground truth for DFC2020 and the **UH-Net variants** (deep layers). Values are averaged over the ten models and reported as percentages, including the standard deviation. For clarity, the numbers are rounded to whole integers. The best values in a dataset column, along with those up to 2% lower, are highlighted in bold. Deviations in the differences may result from rounding.

CNN	Attribution Method	Accuracy			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.
Simple UH-Net	Gradients×Features	**75** ± 1	**75** ± 1	0 ± 0	**65** ± 1	**64** ± 1	0 ± 0
	Ft. Grad-CAM	68 ± 3	**73** ± 3	5 ± 2	54 ± 3	57 ± 3	3 ± 2
	Grad-CAM	**75** ± 1	**75** ± 1	0 ± 0	**65** ± 1	**64** ± 1	0 ± 0
	Integrated Gradients	56 ± 7	57 ± 6	1 ± 1	47 ± 6	47 ± 6	1 ± 1
	DeepLift	56 ± 7	57 ± 6	1 ± 1	47 ± 6	47 ± 6	1 ± 1
	Gradient SHAP	69 ± 1	72 ± 1	3 ± 1	59 ± 1	**63** ± 1	4 ± 1
	DeepLift SHAP	**73** ± 1	69 ± 1	-4 ± 1	**65** ± 1	61 ± 0	-3 ± 1
	Sl. W. Occlusions	**73** ± 0	**75** ± 0	2 ± 0	**63** ± 1	**64** ± 1	1 ± 0
	Ft. Occlusions	**75** ± 1	**75** ± 1	0 ± 0	**64** ± 1	**64** ± 1	0 ± 0
Complex UH-Net	Gradients×Features	32 ± 2	66 ± 3	34 ± 3	26 ± 1	50 ± 3	24 ± 3
	Ft. Grad-CAM	**75** ± 1	**75** ± 1	0 ± 0	**65** ± 1	**64** ± 1	0 ± 0
	Grad-CAM	69 ± 3	**74** ± 3	4 ± 1	56 ± 2	59 ± 3	3 ± 1
	Integrated Gradients	27 ± 3	63 ± 10	36 ± 8	23 ± 2	53 ± 7	30 ± 6
	DeepLift	30 ± 4	65 ± 11	35 ± 8	26 ± 3	54 ± 8	28 ± 6
	Gradient SHAP	21 ± 1	63 ± 6	42 ± 5	19 ± 1	52 ± 4	34 ± 4
	DeepLift SHAP	36 ± 3	**75** ± 2	40 ± 2	30 ± 2	**64** ± 2	33 ± 2
	Sl. W. Occlusions	46 ± 4	64 ± 5	18 ± 4	36 ± 3	48 ± 5	12 ± 3
	Ft. Occlusions	59 ± 8	67 ± 9	8 ± 2	47 ± 6	54 ± 7	6 ± 2

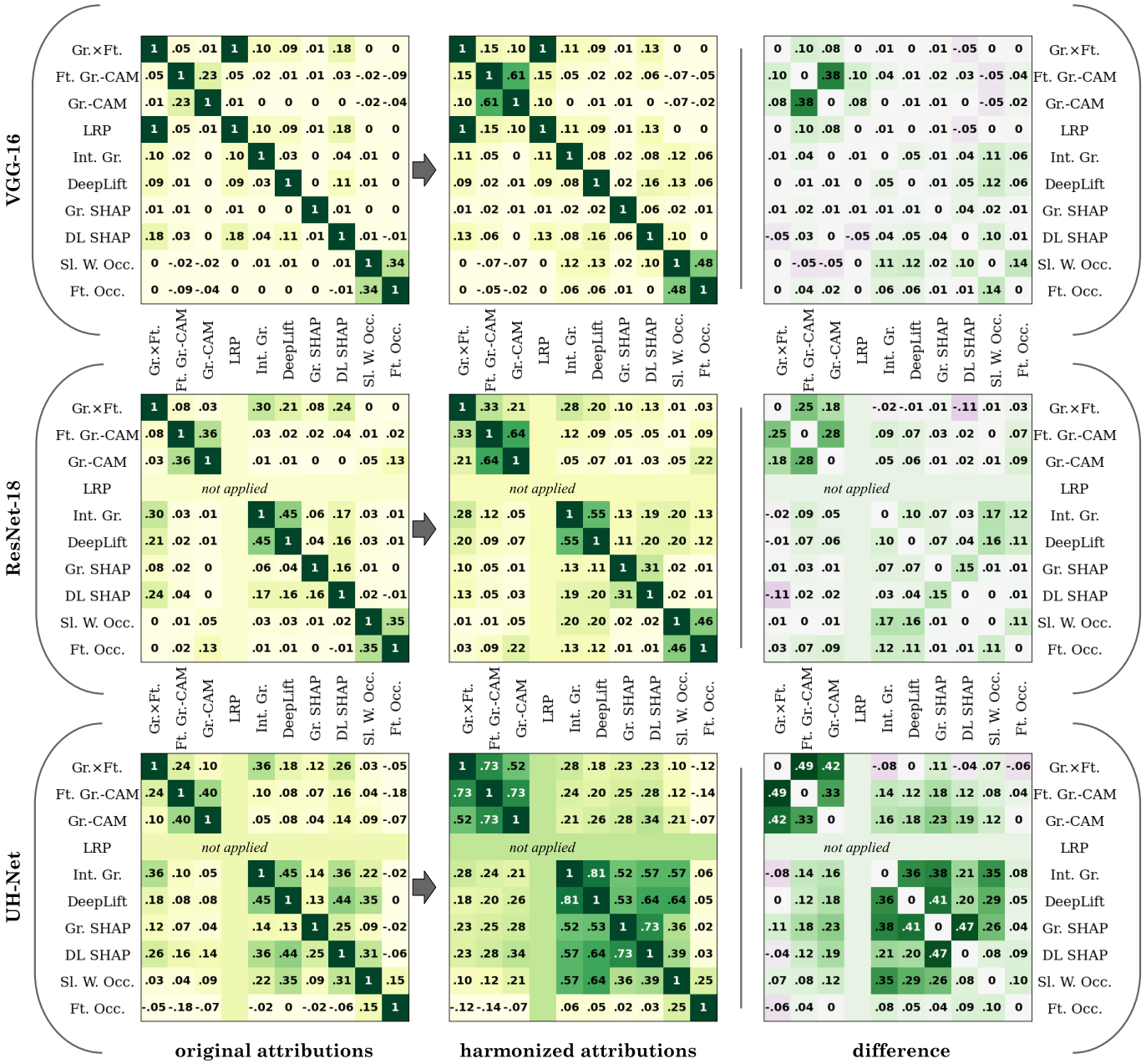


Figure A.3: Pearson correlations between attribution methods for DFC2020 and the **inputs** of the three architectures: VGG-16 (top), ResNet-18 (middle), and UH-Net (bottom). The values are averaged over ten models. From left to right: coefficients for the original attributions, the harmonized attributions, and their difference.

Table A.2: Metrics comparing the predominant class attributions with the segmentation ground truth for DFC2020 and the **inputs** of the three CNNs. Values are averaged over the ten models and reported as percentages, including the standard deviation. For clarity, the numbers are rounded to whole integers. The best values in a dataset column, along with those up to 2% lower, are highlighted in bold. Deviations in the differences may result from rounding.

CNN	Attribution Method	Accuracy			F1 (macro)		
		orig.	harm.	diff.	orig.	harm.	diff.
VGG-16	Gradients×Features	18 ± 0	19 ± 1	1 ± 1	15 ± 0	16 ± 1	1 ± 1
	Ft. Grad-CAM	21 ± 1	24 ± 2	3 ± 1	17 ± 1	19 ± 1	2 ± 1
	Grad-CAM	23 ± 2	22 ± 2	-1 ± 1	19 ± 2	18 ± 2	-1 ± 1
	LRP	18 ± 0	19 ± 1	1 ± 1	15 ± 0	16 ± 1	1 ± 1
	Integrated Gradients	17 ± 0	19 ± 2	2 ± 2	14 ± 0	16 ± 2	2 ± 1
	DeepLift	18 ± 0	22 ± 2	4 ± 2	15 ± 0	18 ± 1	3 ± 1
	Gradient SHAP	14 ± 0	19 ± 1	5 ± 1	12 ± 0	16 ± 1	4 ± 1
	DeepLift SHAP	19 ± 1	**35** ± 2	16 ± 2	16 ± 0	**29** ± 2	13 ± 1
	Sl. W. Occlusions	23 ± 3	23 ± 5	0 ± 5	20 ± 3	17 ± 6	-4 ± 5
	Ft. Occlusions	**26** ± 5	21 ± 6	-5 ± 3	**23** ± 3	18 ± 4	-5 ± 2
ResNet-18	Gradients×Features	**23** ± 1	21 ± 2	-2 ± 2	**20** ± 1	18 ± 1	-2 ± 1
	Ft. Grad-CAM	**23** ± 4	23 ± 6	-1 ± 3	**20** ± 3	20 ± 4	0 ± 2
	Grad-CAM	19 ± 4	17 ± 7	-2 ± 3	17 ± 3	15 ± 5	-3 ± 3
	Integrated Gradients	21 ± 1	25 ± 5	4 ± 4	17 ± 1	21 ± 3	4 ± 2
	DeepLift	**22** ± 1	27 ± 6	5 ± 5	18 ± 1	23 ± 4	5 ± 3
	Gradient SHAP	17 ± 0	31 ± 5	14 ± 4	15 ± 0	26 ± 4	12 ± 3
	DeepLift SHAP	**24** ± 1	**50** ± 6	26 ± 5	**20** ± 1	**41** ± 5	21 ± 4
	Sl. W. Occlusions	19 ± 3	15 ± 4	-4 ± 3	18 ± 2	11 ± 2	-7 ± 2
	Ft. Occlusions	**24** ± 4	20 ± 4	-4 ± 2	**21** ± 3	16 ± 4	-5 ± 2
UH-Net	Gradients×Features	23 ± 2	37 ± 6	14 ± 5	19 ± 2	29 ± 4	9 ± 3
	Ft. Grad-CAM	37 ± 5	42 ± 6	5 ± 2	29 ± 4	33 ± 5	4 ± 2
	Grad-CAM	37 ± 5	40 ± 6	3 ± 2	30 ± 4	32 ± 4	2 ± 1
	Integrated Gradients	28 ± 3	56 ± 9	29 ± 6	23 ± 2	46 ± 5	23 ± 3
	DeepLift	36 ± 5	67 ± 9	31 ± 6	30 ± 4	56 ± 6	26 ± 4
	Gradient SHAP	20 ± 1	62 ± 2	42 ± 2	17 ± 1	51 ± 2	34 ± 2
	DeepLift SHAP	**49** ± 3	**72** ± 1	23 ± 3	**40** ± 2	**61** ± 1	21 ± 3
	Sl. W. Occlusions	33 ± 5	48 ± 11	15 ± 8	28 ± 3	38 ± 8	10 ± 6
	Ft. Occlusions	35 ± 7	30 ± 8	-5 ± 3	29 ± 6	24 ± 7	-5 ± 2

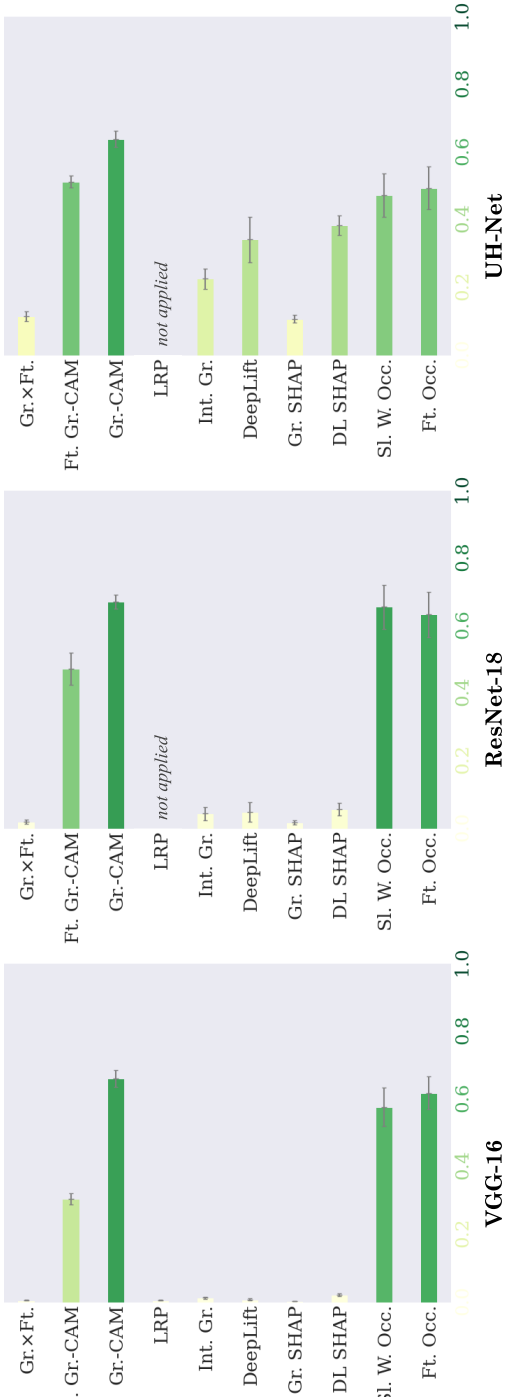


Figure A.4: Similarities between original and harmonized **input** attributions for DFC2020, the three architectures (left to right) and all attribution methods (top to bottom within each plot). The bars represent the Pearson correlations. The color intensity highlights the bar values, with darker shades indicating higher ones. The values are averaged across ten models, and grey lines represent the standard deviations.

A.2 Mapping Naturalness in Fennoscandia

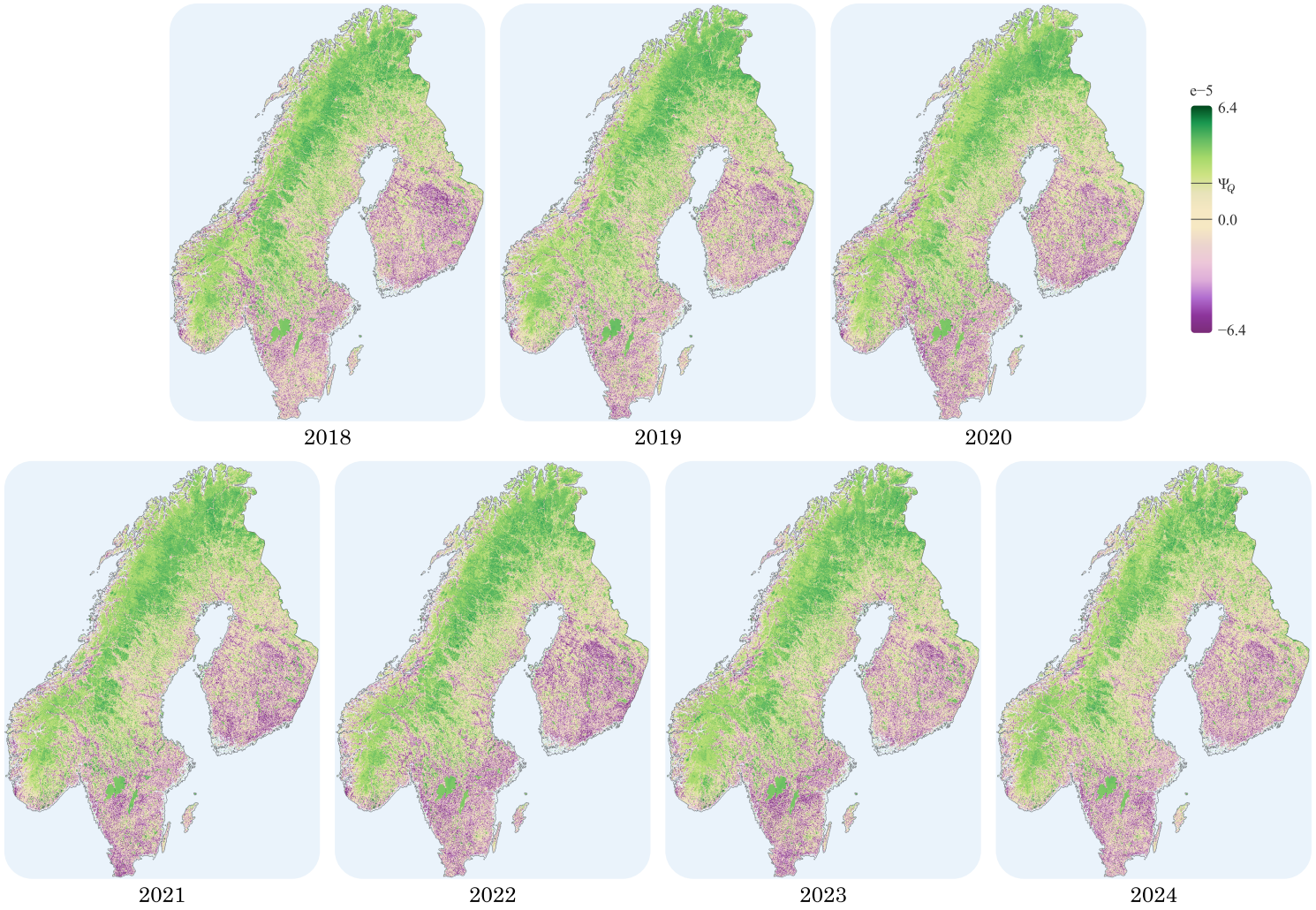


Figure A.5: Our predicted naturalness maps of Fennoscandia for the years 2018 to 2024.

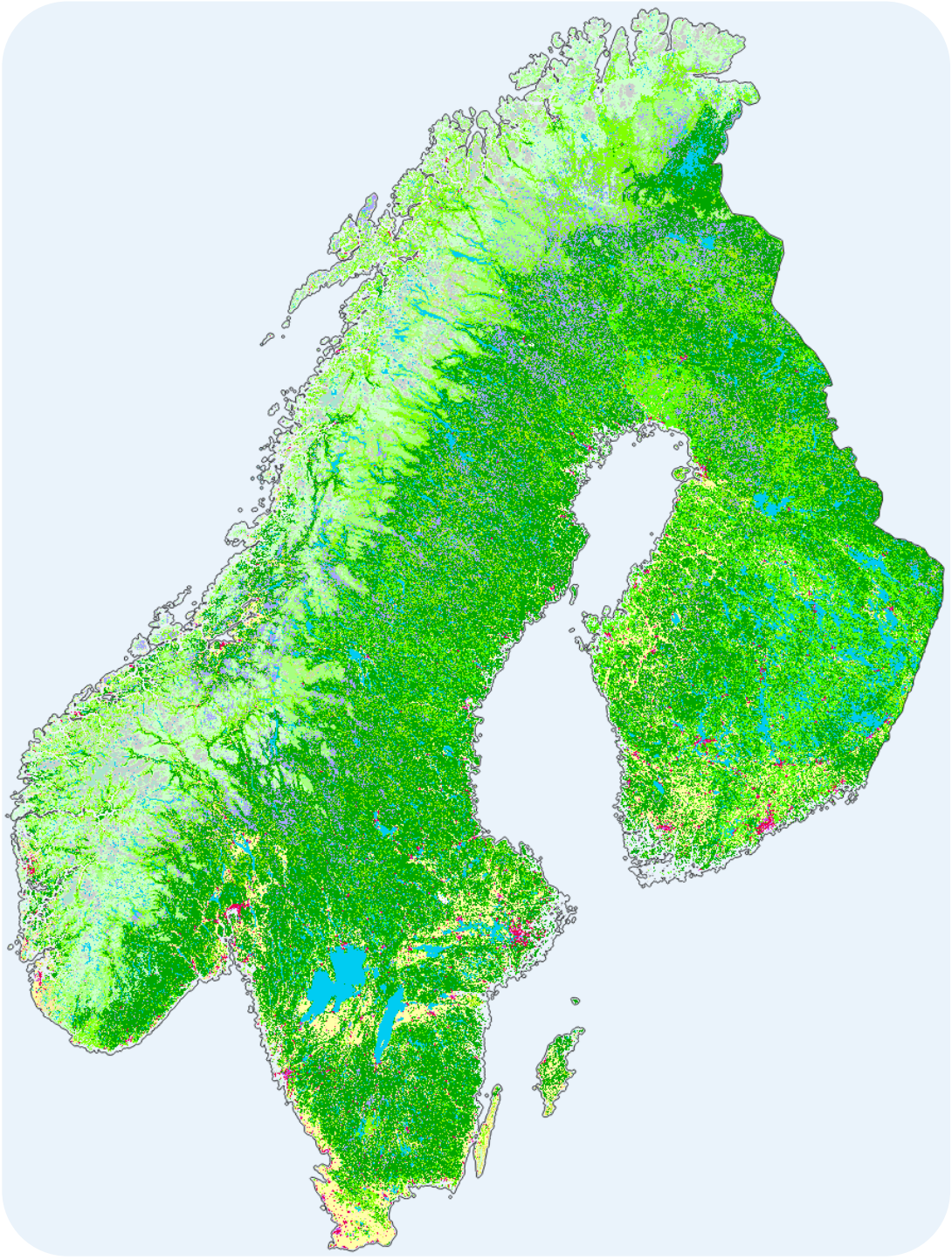


Figure A.6: The CORINE Land Cover map of Fennoscandia (European Environment Agency, 2018). Urban fabric, Arable land, Broad-leaved forest, Coniferous forest, Mixed forest, Moors and heathland, Transitional woodland-shrub, Bare rocks, Sparsely vegetated areas, Wetlands, Water bodies. The full color legend is provided in Table 11.1 (page 107).

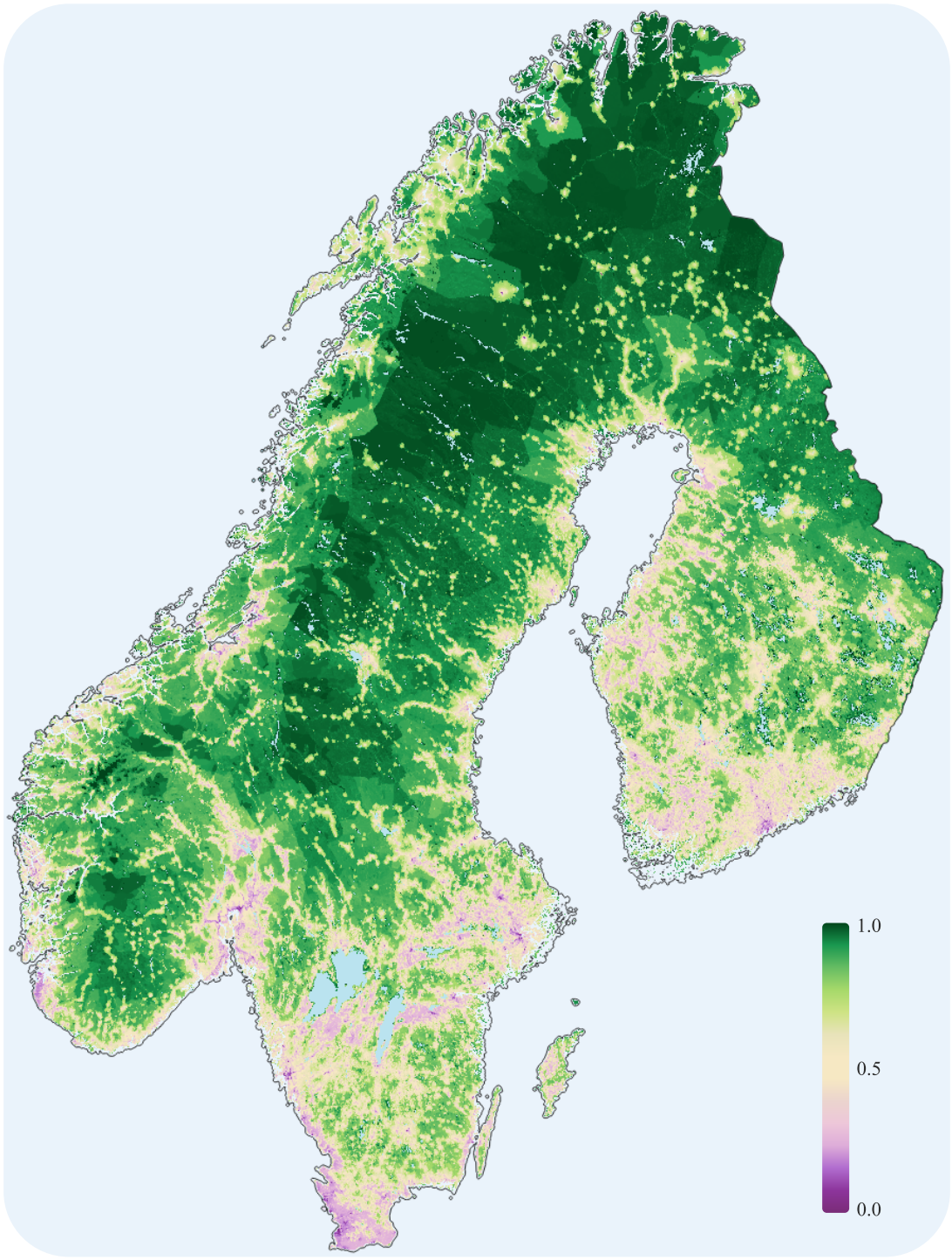


Figure A.7: The Human Modification map of Fennoscandia (Kennedy et al., 2019).

Bibliography

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). "Sanity Checks for Saliency Maps". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc.

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). "On the Surprising Behavior of Distance Metrics in High Dimensional Space". In: *Database Theory — ICDT 2001*. Ed. by G. Goos, J. Hartmanis, J. Van Leeuwen, J. Van Den Bussche, and V. Vianu. Vol. 1973. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 420–434. ISBN: 978-3-540-41456-8. DOI: 10.1007/3-540-44503-X_27.

Ahn, J. and Kwak, S. (2018). "Learning Pixel-Level Semantic Affinity with Image-Level Supervision for Weakly Supervised Semantic Segmentation". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, pp. 4981–4990. ISBN: 978-1-5386-6420-9. DOI: 10.1109/CVPR.2018.00523.

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., and Khan, F. S. (2023). "Transformers in Remote Sensing: A Survey". In: *Remote Sensing* 15.7, p. 1860. ISSN: 2072-4292. DOI: 10.3390/rs15071860.

Allan, J. R., Venter, O., and Watson, J. E. (2017). "Temporally inter-comparable maps of terrestrial wilderness and the Last of the Wild". In: *Scientific Data* 4.1, p. 170187. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.187.

Alotaibi, E. and Nassif, N. (2024). "Artificial intelligence in environmental monitoring: in-depth analysis". In: *Discover Artificial Intelligence* 4.1, p. 84. ISSN: 2731-0809. DOI: 10.1007/s44163-024-00198-1.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2018). *Towards better understanding of gradient-based attribution methods for Deep Neural Networks*.

Asokan, A. and Anitha, J. (2019). "Change detection techniques for remote sensing applications: a survey". In: *Earth Science Informatics* 12.2, pp. 143–160. ISSN: 1865-0481. DOI: 10.1007/s12145-019-00380-5.

Astola, H., Häme, T., Sirro, L., Molinier, M., and Kilpi, J. (2019). "Comparison of Sentinel-2 and Landsat 8 imagery for forest variable prediction in boreal region". In: *Remote Sensing of Environment* 223, pp. 257–273. ISSN: 0034-4257. DOI: 10.1016/j.rse.2019.01.019.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLOS ONE* 10.7. Ed. by O. D. Suarez, e0130140. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130140.

Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Change Theory and Application*. Vol. 15. ISBN: 978-0-13-126780-0.

Bello, O. M. and Aina, Y. A. (2014). "Satellite Remote Sensing as a Tool in Disaster Management and Sustainable Development: Towards a Synergistic Approach". In: *Procedia - Social and Behavioral Sciences* 120, pp. 365–373. ISSN: 18770428. DOI: 10.1016/j.sbspro.2014.02.114.

Bengio, Y., Simard, P., and Frasconi, P. (1994). "Learning long-term dependencies with gradient descent is difficult". In: *IEEE Transactions on Neural Networks* 5.2, pp. 157–166. ISSN: 1941-0093. DOI: 10.1109/72.279181.

Betancourt, C., Stomberg, T., Roscher, R., Schultz, M. G., and Stadtler, S. (2021). "AQ-Bench: a benchmark dataset for machine learning on global air quality metrics". In: *Earth System Science Data* 13.6, pp. 3013–3033. ISSN: 1866-3508. DOI: 10.5194/essd-13-3013-2021.

Betancourt, C., Stomberg, T. T., Edrich, A.-K., Patnala, A., Schultz, M. G., Roscher, R., Kowalski, J., and Stadtler, S. (2022). "Global, high-resolution mapping of tropospheric ozone – explainable machine learning and impact of uncertainties". In: *Geoscientific Model Development* 15.11, pp. 4331–4354. ISSN: 1991-9603. DOI: 10.5194/gmd-15-4331-2022.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information science and statistics. New York: Springer. ISBN: 978-0-387-31073-2.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*. COLT '92. New York, NY, USA: Association for Computing Machinery, pp. 144–152. ISBN: 978-0-89791-497-0. DOI: 10.1145/130385.130401.

Breiman, L. (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.

Brendel, W. and Bethge, M. (2018). "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet". In: International Conference on Learning Representations.

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., and Tait, A. M. (2022). "Dynamic World, Near real-time global 10 m land use land cover mapping". In: *Scientific Data* 9.1, p. 251. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01307-4.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., and Kohli, P. (2025). *AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data*. DOI: 10.48550/ARXIV.2507.22291.

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., and Olah, C. (2019). "Activation Atlas". In: *Distill* 4.3, e15. ISSN: 2476-0757. DOI: 10.23915/distill.00015.

Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. (2019). *Explaining Image Classifiers by Counterfactual Generation*. DOI: 10.48550/arXiv.1807.08024.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847. DOI: 10.1109/WACV.2018.00097.

Chen, L., Wu, W., Fu, C., Han, X., and Zhang, Y. (2020). "Weakly Supervised Semantic Segmentation with Boundary Exploration". In: *Lecture Notes in Computer Science*. Cham: Springer International Publishing, pp. 347–362. ISBN: 978-3-030-58574-7. DOI: 10.1007/978-3-030-58574-7_21.

Chong, Y., Chen, X., Tao, Y., and Pan, S. (2021). "Erase then grow: Generating correct class activation maps for weakly-supervised semantic segmentation". In: *Neurocomputing* 453, pp. 97–108. ISSN: 09252312. DOI: 10.1016/j.neucom.2021.04.103.

Chughtai, A. H., Abbasi, H., and Karas, I. R. (2021). "A review on change detection method and accuracy assessment for land use land cover". In: *Remote Sensing Applications: Society and Environment* 22, p. 100482. ISSN: 23529385. DOI: 10.1016/j.rsase.2021.100482.

Corbane, C., Politis, P., Kempeneers, P., Simonetti, D., Soille, P., Burger, A., Pesaresi, M., Sabo, F., Syrris, V., and Kemper, T. (2020). "A global cloud free pixel- based image composite from Sentinel-2 data". In: *Data in Brief* 31, p. 105737. ISSN: 2352-3409. DOI: 10.1016/j.dib.2020.105737.

Cronon, W. (1995). "The Trouble with Wilderness; or, Getting Back to the Wrong Nature". In: *Uncommon Ground: Rethinking the Human Place in Nature*. New York: W.W. Norton & Co., pp. 69–90.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. (2024). *Vision Transformers Need Registers*. DOI: 10.48550/arXiv.2309.16588.

Desai, S. and Ramaswamy, H. G. (2020). "Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization". In: *2020 IEEE*

Winter Conference on Applications of Computer Vision (WACV). Snowmass Village, CO, USA: IEEE. DOI: 10.1109/wacv45572.2020.9093360.

Dhore, V., Bhat, A., Nerlekar, V., Chavhan, K., and Umare, A. (2024). *Enhancing Explainable AI: A Hybrid Approach Combining GradCAM and LRP for CNN Interpretability*. DOI: 10.48550/arXiv.2405.12175.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *arXiv:2010.11929 [cs]*.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., and Bargellini, P. (2012). "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services". In: *Remote Sensing of Environment* 120, pp. 25–36. ISSN: 00344257. DOI: 10.1016/j.rse.2011.11.026.

Dudley, N. (2008). *Guidelines for applying protected area management categories*. IUCN. ISBN: 978-2-8317-1086-0. DOI: 10.2305/IUCN.CH.2008.PAPS.2.en.

Ekim, B., Dong, Z., Rashkovetsky, D., and Schmitt, M. (2021). "The naturalness index for the identification of natural areas on regional scale". In: *International Journal of Applied Earth Observation and Geoinformation* 105, p. 102622. ISSN: 03032434. DOI: 10.1016/j.jag.2021.102622.

Ekim, B. and Schmitt, M. (2024). *Mapping Land Naturalness from Sentinel-2 using Deep Contextual and Geographical Priors*. DOI: 10.48550/arXiv.2406.19302.

Ekim, B., Stomberg, T. T., Roscher, R., and Schmitt, M. (2023). "MapInWild: A remote sensing dataset to address the question of what makes nature wild". In: *IEEE Geoscience and Remote Sensing Magazine* 11.1, pp. 103–114. ISSN: 2168-6831, 2473-2397, 2373-7468. DOI: 10.1109/MGRS.2022.3226525.

Emam, A., Stomberg, T. T., and Roscher, R. (2024). "Leveraging activation maximization and generative adversarial training to recognize and explain patterns in natural areas in satellite imagery". In: *IEEE Geoscience and Remote Sensing Letters* 21, pp. 1–5. ISSN: 1545-598X, 1558-0571. DOI: 10.1109/LGRS.2023.3335473.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. Vol. 96. AAAI Press, pp. 226–231.

European Environment Agency (2018). *Corine Land Cover (CLC) 2018, Version 2020_20u1*.

European Environment Agency (2021). *Urban Atlas Land Cover/Land Use 2018 (vector), Europe, 6-yearly, Jul. 2021.* DOI: 10.2909/fb4dffa1-6ceb-4cc0-8372-1ed354c285e6.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., Helkowski, J. H., Holloway, T., Howard, E. A., Kucharik, C. J., Monfreda, C., Patz, J. A., Prentice, I. C., Ramankutty, N., and Snyder, P. K. (2005). "Global Consequences of Land Use". In: *Science* 309.5734, pp. 570–574. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1111772.

Fong, R. C. and Vedaldi, A. (2017). "Interpretable Explanations of Black Boxes by Meaningful Perturbation". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE. DOI: 10.1109/iccv.2017.371.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). "Google Earth Engine: Planetary-scale geospatial analysis for everyone". In: *Remote Sensing of Environment* 202, pp. 18–27. ISSN: 0034-4257. DOI: 10.1016/j.rse.2017.06.031.

Grime, J. P. (1997). "Biodiversity and Ecosystem Function: The Debate Deepens". In: *Science* 277.5330, pp. 1260–1261. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.277.5330.1260.

Gulum, M. A., Trombley, C. M., and Kantardzic, M. (2021). "Improved Deep Learning Explanations for Prostate Lesion Classification through Grad-CAM and Saliency Map Fusion". In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. Aveiro, Portugal: IEEE, pp. 498–502. ISBN: 978-1-6654-4121-6. DOI: 10.1109/CBMS52027.2021.00099.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). "Dimensionality Reduction by Learning an Invariant Mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*. Vol. 2. New York, NY, USA: IEEE, pp. 1735–1742. ISBN: 978-0-7695-2597-6. DOI: 10.1109/CVPR.2006.100.

Hanna, J., Mommert, M., and Borth, D. (2023). "Sparse Multimodal Vision Transformer for Weakly Supervised Semantic Segmentation". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Vancouver, BC, Canada: IEEE, pp. 2145–2154. ISBN: 9798350302493. DOI: 10.1109/CVPRW59228.2023.00208.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R. G. (2013). "High-Resolution Global Maps of 21st-Century Forest Cover Change". In: *Science* 342.6160, pp. 850–853. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1244693.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, pp. 770–778. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.90.

Helber, P., Bischke, B., Dengel, A., and Borth, D. (2019). "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.7, pp. 2217–2226. ISSN: 1939-1404, 2151-1535. DOI: 10.1109/JSTARS.2019.2918242.

Höhl, A., Obadic, I., Fernández-Torres, M.-A., Najjar, H., Oliveira, D. A. B., Akata, Z., Dengel, A., and Zhu, X. X. (2024). "Opening the Black Box: A systematic review on explainable artificial intelligence in remote sensing". In: *IEEE Geoscience and Remote Sensing Magazine* 12.4, pp. 261–304. ISSN: 2168-6831. DOI: 10.1109/MGRS.2024.3467001.

Hsu, C.-Y. and Li, W. (2023). "Explainable GeoAI: can saliency maps help interpret artificial intelligence's learning process? An empirical study on natural feature detection". In: *International Journal of Geographical Information Science* 37.5, pp. 963–987. ISSN: 1365-8816, 1362-3087. DOI: 10.1080/13658816.2023.2191256.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Ienco, D., Interdonato, R., Gaetano, R., and Ho Tong Minh, D. (2019). "Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 158, pp. 11–22. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2019.09.016.

Iqbal, H. (2018). *HarisIqbal88/PlotNeuralNet v1.0.0*. DOI: 10.5281/zenodo.2526396.

Jain, P., Ienco, D., Interdonato, R., Berchoux, T., and Marcos, D. (2024). *SenCLIP: Enhancing zero-shot land-use mapping for Sentinel-2 with ground-level prompting*. DOI: 10.48550/arXiv.2412.08536.

Jain, S. and Wallace, B. C. (2019). *Attention is not Explanation*. DOI: 10.48550/arXiv.1902.10186.

Jonnarth, A. and Felsberg, M. (2022). "Importance Sampling Cams For Weakly-Supervised Segmentation". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2639–2643. DOI: 10.1109/ICASSP43922.2022.9746641.

Kakogeorgiou, I. and Karantzalos, K. (2021). "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote

sensing". In: *International Journal of Applied Earth Observation and Geoinformation* 103, p. 102520. ISSN: 0303-2434. DOI: 10.1016/j.jag.2021.102520.

Kattenborn, T., Leitloff, J., Schiefer, F., and Hinz, S. (2021). "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 173, pp. 24–49. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2020.12.010.

Kennedy, C. M., Oakleaf, J. R., Theobald, D. M., Baruch-Mordo, S., and Kiesecker, J. (2019). "Managing the middle: A shift in conservation priorities based on the global human modification gradient". In: *Global Change Biology* 25.3, pp. 811–826. ISSN: 1365-2486. DOI: 10.1111/gcb.14549.

Kierdorf, J., Stomberg, T. T., Drees, L., Rascher, U., and Roscher, R. (2024). "Investigating the contribution of image time series observations to cauliflower harvest-readiness prediction". In: *Frontiers in Artificial Intelligence* 7, p. 1416323. ISSN: 2624-8212. DOI: 10.3389/frai.2024.1416323.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018). "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)". In: p. 10.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). "Concept Bottleneck Models". In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, pp. 5338–5348.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, O. (2020). *Captum: A unified and generic model interpretability library for PyTorch*.

Kouki, J., Löfman, S., Martikainen, P., Rouvinen, S., and Uotila, A. (2001). "Forest Fragmentation in Fennoscandia: Linking Habitat Requirements of Wood-associated Threatened Species to Landscape and Habitat Changes". In: *Scandinavian Journal of Forest Research* 16.sup003, pp. 27–37. ISSN: 0282-7581, 1651-1891. DOI: 10.1080/028275801300090564.

Kwak, S., Hong, S., and Han, B. (2017). "Weakly Supervised Semantic Segmentation Using Superpixel Pooling Network". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1. ISSN: 2374-3468, 2159-5399. DOI: 10.1609/aaai.v31i1.11213.

Lande, R. and Shannon, S. (1996). "The Role of Genetic Variation in Adaptation and Population Persistence in a Changing Environment". In: *Evolution* 50.1, pp. 434–437. ISSN: 0014-3820. DOI: 10.2307/2410812.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). "Unmasking Clever Hans predictors and assessing what machines really learn". In: *Nature Communications* 10.1, p. 1096. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08987-4.

Levering, A., Marcos, D., Lobry, S., and Tuia, D. (2020). "Interpretable Scenicness from Sentinel-2 Imagery". In: *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. Waikoloa, HI, USA: IEEE, pp. 3983–3986. ISBN: 978-1-72816-374-1. DOI: 10.1109/IGARSS39084.2020.9323706.

Li, F.-F., Andreeto, M., Ranzato, M., and Perona, P. (2022). *Caltech 101*.

Lloyd, S. (1982). "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137. ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489.

Loveland, T. R. and Belward, A. S. (1997). "The International Geosphere Biosphere Programme Data and Information System global land cover data set (DISCover)". In: *Acta Astronautica*. Developing Business 41.4, pp. 681–689. ISSN: 0094-5765. DOI: 10.1016/S0094-5765(98)00050-2.

Lundberg, S. M. and Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

Lunde, L. F., Birkemoe, T., Sverdrup-Thygeson, A., Asplund, J., Halvorsen, R., Kjønaas, O. J., Nordén, J., Maurice, S., Skrede, I., Nybakken, L., and Kauserud, H. (2025). "Towards repeated clear-cutting of boreal forests – a tipping point for biodiversity?" In: *Biological Reviews* 100.3, pp. 1181–1205. ISSN: 1469-185X. DOI: 10.1111/brv.13180.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., and Johnson, B. A. (2019). "Deep learning in remote sensing applications: A meta-analysis and review". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 152, pp. 166–177. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2019.04.015.

Machichi, M. A., l. E. mansouri loubna, y. imani yasmina, Bourja, O., Lahlou, O., Zennayi, Y., Bourzeix, F., Houmma, I. H., and Hadria, R. (2023). "Crop mapping using supervised machine learning and deep learning: a systematic literature review". In: *International Journal of Remote Sensing*. ISSN: 0143-1161.

Marcos, D., Fong, R., Lobry, S., Flamary, R., Courty, N., and Tuia, D. (2020). "Contextual Semantic Interpretability". In: Proceedings of the Asian Conference on Computer Vision.

Mashala, M. J., Dube, T., Mudereri, B. T., Ayisi, K. K., and Ramudzuli, M. R. (2023). "A Systematic Review on Advancements in Remote Sensing for Assessing and Monitoring Land Use and Land Cover Changes Impacts on Surface Water Resources in Semi-Arid Tropical Environments". In: *Remote Sensing* 15.16, p. 3926. ISSN: 2072-4292. DOI: 10.3390/rs15163926.

McInnes, L., Healy, J., and Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*.

Misra, G., Cawkwell, F., and Wingler, A. (2020). "Status of Phenological Research Using Sentinel-2 Data: A Review". In: *Remote Sensing* 12.17, p. 2760. ISSN: 2072-4292. DOI: 10.3390/rs12172760.

Mohan, A. and Peeples, J. (2023). "Quantitative Analysis of Primary Attribution Explainable Artificial Intelligence Methods for Remote Sensing Image Classification". In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 950–953. DOI: 10.1109/IGARSS52108.2023.10281981.

Mommert, M., Kesseli, N., Hanna, J., Scheibenreif, L., Borth, D., and Demir, B. (2023). "Ben-Ge: Extending Bigearthnet with Geographical and Environmental Data". In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. Pasadena, CA, USA: IEEE, pp. 1016–1019. ISBN: 9798350320107. DOI: 10.1109/IGARSS52108.2023.10282767.

Mu, H., Li, X., Wen, Y., Huang, J., Du, P., Su, W., Miao, S., and Geng, M. (2022). "A global record of annual terrestrial Human Footprint dataset from 2000 to 2018". In: *Scientific Data* 9.1, p. 176. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01284-8.

Mudelsee, M. (2019). "Trend analysis of climate time series: A review of methods". In: *Earth-Science Reviews* 190, pp. 310–322. ISSN: 0012-8252. DOI: 10.1016/j.earscirev.2018.12.005.

Muhammad, M. B. and Yeasin, M. (2020). "Eigen-CAM: Class Activation Map using Principal Components". In: pp. 1–7. DOI: 10.1109/IJCNN48605.2020.9206626.

Newbold, T., Hudson, L. N., Hill, S. L. L., Contu, S., Lysenko, I., Senior, R. A., Börger, L., Bennett, D. J., Choimes, A., Collen, B., Day, J., De Palma, A., Díaz, S., Echeverria-Londoño, S., Edgar, M. J., Feldman, A., Garon, M., Harrison, M. L. K., Alhusseini, T., Ingram, D. J., Itescu, Y., Kattge, J., Kemp, V., Kirkpatrick, L., Kleyer, M., Correia, D. L. P., Martin, C. D., Meiri, S., Novosolov, M., Pan, Y., Phillips, H. R. P., Purves, D. W., Robinson, A., Simpson, J., Tuck, S. L., Weiher, E., White, H. J., Ewers, R. M., Mace, G. M., Scharlemann, J. P. W., and Purvis, A. (2015). "Global effects of land use on local terrestrial biodiversity". In: *Nature* 520.7545, pp. 45–50. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14324.

Nguyen, T.-A., Kellenberger, B., and Tuia, D. (2022). "Mapping forest in the Swiss Alps treeline ecotone with explainable deep learning". In: *Remote Sensing of Environment* 281, p. 113217. ISSN: 00344257. DOI: 10.1016/j.rse.2022.113217.

Nieradzik, L., Stephani, H., Sieburg-Rockel, J., Helmling, S., Olbrich, A., and Keuper, J. (2024). *Challenging the Black Box: A Comprehensive Evaluation of Attribution Maps of CNN Applications in Agriculture and Forestry.*

Odena, A., Dumoulin, V., and Olah, C. (2016). "Deconvolution and Checkerboard Artifacts". In: *Distill*. ISSN: 2476-0757. DOI: 10.23915/distill.00003.

Olah, C., Mordvintsev, A., and Schubert, L. (2017). "Feature Visualization". In: *Distill* 2.11, e7. ISSN: 2476-0757. DOI: 10.23915/distill.00007.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). "The Building Blocks of Interpretability". In: *Distill* 3.3, e10. ISSN: 2476-0757. DOI: 10.23915/distill.00010.

OpenStreetMap contributors (2017). *Planet dump retrieved from https://planet.osm.org*.

Ørka, H. O., Jutras-Perreault, M.-C., Næsset, E., and Gobakken, T. (2022). "A framework for a forest ecological base map – An example from Norway". In: *Ecological Indicators* 136, p. 108636. ISSN: 1470160X. DOI: 10.1016/j.ecolind.2022.108636.

Östlund, L., Zackrisson, O., and Axelsson, A. .-. (1997). "The history and transformation of a Scandinavian boreal forest landscape since the 19th century". In: *Canadian Journal of Forest Research* 27.8, pp. 1198–1206. DOI: 10.1139/x97-070.

Otsuki, S., Iida, T., Doublet, F., Hirakawa, T., Yamashita, T., Fujiyoshi, H., and Sugiura, K. (2025). "Layer-Wise Relevance Propagation with Conservation Property for ResNet". In: *Computer Vision – ECCV 2024*. Ed. by A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol. Cham: Springer Nature Switzerland, pp. 349–364. ISBN: 978-3-031-72775-7. DOI: 10.1007/978-3-031-72775-7_20.

Pal, M. and Mather, P. M. (2003). "An assessment of the effectiveness of decision tree methods for land cover classification". In: *Remote Sensing of Environment* 86.4, pp. 554–565. ISSN: 00344257. DOI: 10.1016/S0034-4257(03)00132-9.

Papoutsis, I., Bountos, N.-I., Zavras, A., Michail, D., and Tryfonopoulos, C. (2022). *Benchmarking and scaling of deep learning models for land cover image classification*. DOI: 10.48550/arXiv.2111.09451.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011).

"Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830.

Petsiuk, V., Das, A., and Saenko, K. (2018). "RISE: Randomized Input Sampling for Explanation of Black-box Models". In: *British Machine Vision Conference (BMVC)*.

Pettorelli, N., Schulte To Bühne, H., Tulloch, A., Dubois, G., Macinnis-Ng, C., Queirós, A. M., Keith, D. A., Wegmann, M., Schrodt, F., Stellmes, M., Sonnenschein, R., Geller, G. N., Roy, S., Somers, B., Murray, N., Bland, L., Geijzendorffer, I., Kerr, J. T., Broszeit, S., Leitão, P. J., Duncan, C., El Serafy, G., He, K. S., Blanchard, J. L., Lucas, R., Mairota, P., Webb, T. J., and Nicholson, E. (2018). "Satellite remote sensing of ecosystem functions: opportunities, challenges and way forward". In: *Remote Sensing in Ecology and Conservation* 4.2. Ed. by M. Rowcliffe and M. Disney, pp. 71–93. ISSN: 20563485. DOI: 10.1002/rse2.59.

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V. R., Murayama, Y., and Ranagalage, M. (2020). "Sentinel-2 Data for Land Cover/Use Mapping: A Review". In: *Remote Sensing* 12.14, p. 2291. ISSN: 2072-4292. DOI: 10.3390/rs12142291.

Plotly Technologies Inc. (2015). *Collaborative data science*. Montreal, QC.

Raschka, S., Patterson, J., and Nolet, C. (2020). *Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence*. DOI: 10.48550/arXiv.2002.04803.

Reddy, C. S., Kurian, A., Srivastava, G., Singhal, J., Varghese, A. O., Padalia, H., Ayyappan, N., Rajashekar, G., Jha, C. S., and Rao, P. V. N. (2021). "Remote sensing enabled essential biodiversity variables for biodiversity assessment and monitoring: technological advancement and potentials". In: *Biodiversity and Conservation* 30.1, pp. 1–14. ISSN: 1572-9710. DOI: 10.1007/s10531-020-02073-8.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: Association for Computing Machinery, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.

Richards, J. A. (2022). *Remote Sensing Digital Image Analysis*. Cham: Springer International Publishing. ISBN: 978-3-030-82326-9. DOI: 10.1007/978-3-030-82327-6.

Robinson, C., Malkin, K., Jojic, N., Chen, H., Qin, R., Xiao, C., Schmitt, M., Ghamisi, P., Hansch, R., and Yokoya, N. (2021). "Global Land-Cover Mapping With Weak Supervision: Outcome of the 2020 IEEE GRSS Data Fusion Contest". In: *IEEE Journal of Selected Topics in Applied Earth Observa-*

tions and Remote Sensing 14, pp. 3185–3199. ISSN: 1939-1404, 2151-1535. DOI: 10.1109/JSTARS.2021.3063849.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Cham: Springer International Publishing, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020a). "EXPLAIN IT TO ME – FACING REMOTE SENSING CHALLENGES IN THE BIO- AND GEOSCIENCES WITH EXPLAINABLE MACHINE LEARNING". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-3-2020, pp. 817–824. ISSN: 2194-9050. DOI: 10.5194/isprs-annals-V-3-2020-817-2020.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020b). "Explainable Machine Learning for Scientific Insights and Discoveries". In: *IEEE Access* 8, pp. 42200–42216. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2976199.

Roy, P. S., Ramachandran, R. M., Paul, O., Thakur, P. K., Ravan, S., Behera, M. D., Sarangi, C., and Kanawade, V. P. (2022). "Anthropogenic Land Use and Land Cover Changes—A Review on Its Environmental Consequences and Climate Change". In: *Journal of the Indian Society of Remote Sensing* 50.8, pp. 1615–1640. ISSN: 0974-3006. DOI: 10.1007/s12524-022-01569-w.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. (2017). "Evaluating the Visualization of What a Deep Neural Network Has Learned". In: *IEEE Transactions on Neural Networks and Learning Systems* 28.11, pp. 2660–2673. ISSN: 2162-2388. DOI: 10.1109/TNNLS.2016.2599820.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Muller, K.-R. (2021). "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications". In: *Proceedings of the IEEE* 109.3, pp. 247–278. ISSN: 0018-9219, 1558-2256. DOI: 10.1109/JPROC.2021.3060483.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Múller, K.-R. (2020). "Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond". In: *arXiv:2003.07631*.

Sanderson, E. W., Jaiteh, M., Levy, M. A., Redford, K. H., Wannebo, A. V., and Woolmer, G. (2002). "The Human Footprint and the Last of the Wild". In: *BioScience* 52.10, pp. 891–904. ISSN: 0006-3568. DOI: 10.1641/0006-3568(2002)052[0891:THFATL]2.0.CO;2.

Saphra, N. and Wiegreffe, S. (2024). *Mechanistic?* DOI: 10.48550/arXiv.2410.09087.

Schmitt, M., Hughes, L. H., Qiu, C., and Zhu, X. X. (2019a). "SEN12MS – A Curated Dataset Of Georeferenced Multi-Spectral Sentinel-1/2 Imagery For Deep Learning And Data Fusion". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W7, pp. 153–160. ISSN: 2194-9050. DOI: 10.5194/isprs-annals-IV-2-W7-153-2019.

Schmitt, M., Hughes, L., Ghamisi, P., Yokoya, N., and Hänsch, R. (2019b). *2020 IEEE GRSS Data Fusion Contest*. DOI: 10.21227/rha7-m332.

Schumacher, J., Hauglin, M., Astrup, R., and Breidenbach, J. (2020). "Mapping forest age using National Forest Inventory, airborne laser scanning, and Sentinel-2 data". In: *Forest Ecosystems* 7.1, p. 60. ISSN: 2197-5620. DOI: 10.1186/s40663-020-00274-9.

Segarra, J., Buchaillot, M. L., Araus, J. L., and Kefauver, S. C. (2020). "Remote Sensing for Precision Agriculture: Sentinel-2 Improved Features and Applications". In: *Agronomy* 10.5, p. 641. ISSN: 2073-4395. DOI: 10.3390/agronomy10050641.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2, pp. 336–359. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-019-01228-7.

Shapley, L. (1953). "A Value for n-Person Games." In: *Classics in Game Theory*. Princeton University Press, pp. 307–317. ISBN: 978-1-4008-2915-6.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). "Learning Important Features Through Propagating Activation Differences". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, pp. 3145–3153.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *arXiv:1312.6034 [cs]*.

Simonyan, K. and Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. DOI: 10.48550/arXiv.1409.1556.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). *SmoothGrad: removing noise by adding noise*.

Song, J., Shaohua, G., Yunqiang, Z., and Ma, C. (2019). "A survey of remote sensing image classification based on CNNs". In: *Big Earth Data* 3.3, pp. 232–254. ISSN: 2096-4471. DOI: 10.1080/20964471.2019.1657720.

Spearman, C. (1904). "The Proof and Measurement of Association between Two Things". In: *The American Journal of Psychology* 15.1, pp. 72–101. ISSN: 0002-9556. DOI: 10.2307/1412159.

Steffen, W., Persson, A., Deutsch, L., Zalasiewicz, J., Williams, M., Richardson, K., Crumley, C., Crutzen, P., Folke, C., Gordon, L., Molina, M., Ramanathan,

V., Rockström, J., Scheffer, M., Schellnhuber, H. J., and Svedin, U. (2011). "The Anthropocene: From Global Change to Planetary Stewardship". In: *AMBIO* 40.7, pp. 739–761. ISSN: 1654-7209. DOI: 10.1007/s13280-011-0185-x.

Stomberg, T., Weber, I., Schmitt, M., and Roscher, R. (2021). "jUngle-Net: Using explainable machine learning to gain new insights into the appearance of wilderness in satellite imagery". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-3-2021, pp. 317–324. ISSN: 2194-9042. DOI: 10.5194/isprs-annals-V-3-2021-317-2021.

Stomberg, T. T., Leonhardt, J., Weber, I., and Roscher, R. (2023). "Recognizing protected and anthropogenic patterns in landscapes using interpretable machine learning and satellite imagery". In: *Frontiers in Artificial Intelligence* 6, p. 1278118. ISSN: 2624-8212. DOI: 10.3389/frai.2023.1278118.

Stomberg, T. T., Reißner, L. A., Schultz, M. G., and Roscher, R. (2025). "Building consistency in explanations: Harmonizing CNN attributions for satellite-based land cover classification". In: *Machine Learning with Applications* 20, p. 100653. ISSN: 26668270. DOI: 10.1016/j.mlwa.2025.100653.

Stomberg, T. T., Stone, T., Leonhardt, J., Weber, I., and Roscher, R. (2022). "Exploring wilderness characteristics using explainable machine learning in satellite imagery". In: *arXiv (cs)*. DOI: 10.48550/arXiv.2203.00379.

Student (1908). "The Probable Error of a Mean". In: *Biometrika* 6.1, pp. 1–25. ISSN: 0006-3444. DOI: 10.2307/2331554.

Sumbul, G., Wall, A. de, Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., and Markl, V. (2021). "BigEarthNet-MM: A Large Scale Multi-Modal Multi-Label Benchmark Archive for Remote Sensing Image Classification and Retrieval". In: *IEEE Geoscience and Remote Sensing Magazine* 9.3, pp. 174–180. ISSN: 2168-6831, 2473-2397, 2373-7468. DOI: 10.1109/MGRS.2021.3089174.

Sundararajan, M., Taly, A., and Yan, Q. (2017). "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 3319–3328.

Thackway, R., Lymburner, L., and Guerschman, J. P. (2013). "Dynamic land cover information: bridging the gap between remote sensing and natural resource management". In: *Ecology and Society* 18.1. ISSN: 1708-3087.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I. N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, M., and Rostan, F. (2012). "GMES Sentinel-1 mission". In: *Remote Sensing of Environment* 120, pp. 9–24. ISSN: 00344257. DOI: 10.1016/j.rse.2011.05.028.

UNEP-WCMC and IUCN (2025). *Protected Planet: The World Database on Protected Areas (WDPA) [Online], [February 2025]*.

United Nations (2021). *The Sustainable Development Goals Report*. ISBN: 978-92-1-101439-6.

Venter, O., Sanderson, E. W., Magrach, A., Allan, J. R., Beher, J., Jones, K. R., Possingham, H. P., Laurance, W. F., Wood, P., Fekete, B. M., Levy, M. A., and Watson, J. E. M. (2016a). "Global terrestrial Human Footprint maps for 1993 and 2009". In: *Scientific Data* 3.1, p. 160067. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.67.

Venter, O., Sanderson, E. W., Magrach, A., Allan, J. R., Beher, J., Jones, K. R., Possingham, H. P., Laurance, W. F., Wood, P., Fekete, B. M., Levy, M. A., and Watson, J. E. M. (2016b). "Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation". In: *Nature Communications* 7.1, p. 12558. ISSN: 2041-1723. DOI: 10.1038/ncomms12558.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2020a). "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE. DOI: 10.1109/cvprw50498.2020.00020.

Wang, Y., Zhang, J., Kan, M., Shan, S., and Chen, X. (2020b). "Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, pp. 12272–12281. ISBN: 978-1-72817-168-5. DOI: 10.1109/CVPR42600.2020.01229.

Wendling, Z. A., Emerson, J. W., Sherbinin, A. de, Esty, D. C., and et. al. (2020). *2020 Environmental Performance Index*. New Haven, CT: Yale Center for Environmental Law & Policy.

Wiegreffe, S. and Pinter, Y. (2019). *Attention is not not Explanation*. DOI: 10.48550/arXiv.1908.04626.

Willbo, M., Pirinen, A., Martinsson, J., Zec, E. L., Mogren, O., and Nilsson, M. (2024). *Impacts of Color and Texture Distortions on Earth Observation Data in Deep Learning*.

Wulder, M. A., Roy, D. P., Radeloff, V. C., Loveland, T. R., Anderson, M. C., Johnson, D. M., Healey, S., Zhu, Z., Scambos, T. A., Pahlevan, N., Hansen, M., Gorelick, N., Crawford, C. J., Masek, J. G., Hermosilla, T., White, J. C., Belward, A. S., Schaaf, C., Woodcock, C. E., Huntington, J. L., Lymburner, L., Hostert, P., Gao, F., Lyapustin, A., Pekel, J.-F., Strobl, P., and Cook, B. D. (2022). "Fifty years of Landsat science and impacts". In: *Remote Sensing of Environment* 280, p. 113195. ISSN: 00344257. DOI: 10.1016/j.rse.2022.113195.

Yang, M. and Kim, B. (2019). *Benchmarking Attribution Methods with Relative Feature Importance*.

Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., and Choe, J. (2019). "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, pp. 6022–6031. ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00612.

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., and Arino, O. (2022). *ESA WorldCover 10 m 2021 v200*. DOI: 10.5281/zenodo.7254221.

Zeiler, M. D. and Fergus, R. (2014). "Visualizing and Understanding Convolutional Networks". In: *Computer Vision – ECCV 2014*. Ed. by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 818–833. ISBN: 978-3-319-10590-1. DOI: 10.1007/978-3-319-10590-1_53.

Zeng, X., Wang, T., Dong, Z., Zhang, X., and Gu, Y. (2023). "Superpixel Consistency Saliency Map Generation for Weakly Supervised Semantic Segmentation of Remote Sensing Images". In: *IEEE Transactions on Geoscience and Remote Sensing* 61, pp. 1–16. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2023.3264232.

Zhang, D., Zhang, H., Tang, J., Hua, X.-S., and Sun, Q. (2020). "Causal Intervention for Weakly-Supervised Semantic Segmentation". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 655–666.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). *mixup: Beyond Empirical Risk Minimization*.

Zhang, L., Zhang, L., and Du, B. (2016). "Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art". In: *IEEE Geoscience and Remote Sensing Magazine* 4.2, pp. 22–40. ISSN: 2168-6831, 2473-2397. DOI: 10.1109/MGRS.2016.2540798.

Zhang, L. and Ma, J. (2021). "Salient Object Detection Based on Progressively Supervised Learning for Remote Sensing Images". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.11, pp. 9682–9696. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2020.3045708.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. (2021). *How Does Mixup Help With Robustness and Generalization?* DOI: 10.48550/arXiv.2010.04819.

Zhou, B., Bau, D., Oliva, A., and Torralba, A. (2019). "Interpreting Deep Visual Representations via Network Dissection". In: *IEEE Transactions on Pattern*

Analysis and Machine Intelligence 41.9, pp. 2131–2145. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2018.2858759.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). "Learning Deep Features for Discriminative Localization". In: pp. 2921–2929.

Zhou, Z.-H. (2018). "A brief introduction to weakly supervised learning". In: *National Science Review* 5.1, pp. 44–53. ISSN: 2095-5138, 2053-714X. DOI: 10.1093/nsr/nwx106.

Abbreviations

CAM	Class Activation Mapping
CNN	Convolutional neural network
DL	DeepLift
ESA	European Space Agency
Ft.	Features
Gr.	Gradients
Int.	Integrated
IUCN	International Union for Conservation of Nature and Natural Resources
***k*-m.**	*k*-means
LR	Learning rate
LRP	Layer-wise Relevance Propagation
M	Million
NASA	National Aeronautics and Space Administration
Occ.	Occlusions
Params	Parameters
SDG	Sustainable Development Goals
Sl. W.	Sliding Window
USGS	United States Geological Survey
WD	Weight decay
WDPA	World Database on Protected Areas

Notation

Scalars, such as a and A, as well as functions $f(\cdot)$ and $F(\cdot)$, are denoted using regular italic letters.

Tensors are written in boldface. Vectors, such as $\boldsymbol{a} = (a_1, ..., a_k)^T \in \mathbb{R}^k$, are denoted using bold lowercase letters, while higher-dimensional tensors $\boldsymbol{A} \in \mathbb{R}^{k_1 \times ... \times k_n}$ are written in bold uppercase letters.

Sets $\mathcal{A} = \{a_1, ..., a_n\}$ are written in uppercase calligraphic letters.

List of Figures

List of Tables

Acknowledgements

Danksagung

Während der vergangenen fünf Jahre war ich Teil einer großartigen Arbeitsgruppe mit wunderbaren Kolleg:innen und einer engagierten Betreuerin. Liebe Ribana, in diesen Jahren warst du jederzeit erreichbar, hast dir viel Zeit für Diskussionen und Feedback genommen und mir zugleich die Freiheit gelassen, eigene Forschungsschwerpunkte zu wählen. Der Austausch mit dir war immer offen und herzlich. Auch bin ich dankbar für meine tollen Kolleg:innen Lukas, Jana, Eike, Johannes, Felix, Immanuel, Mohamed, Ahmed, Lydia und Genc. Wir haben uns nicht nur im Büro und auf Konferenzen bestens verstanden, sondern auch viele schöne Erlebnisse geteilt — von gemeinsamen Radtouren über Spieleabende bis hin zu gnadenlosen Lasertag-Runden.

Lieber Martin, als zweiter Betreuer hast du mir regelmäßig wertvolles Feedback gegeben und mich ohne Zögern für das erste halbe Jahr in Jülich aufgenommen. Lieber Michael, ich denke gerne an lustige Abende auf Konferenzen zurück und danke dir für den motivierenden Forschungsaufenthalt in St. Gallen. Danken möchte ich außerdem allen, mit denen ich gemeinsam an Publikationen gearbeitet und dabei wertvolle Erfahrungen gesammelt habe. Neben meinen Betreuern und Kolleg:innen möchte ich hier besonders Taylor, Burak, Michael, Lennart, Clara, Scarlet, Ann-Kathrin und Ankit hervorheben. Auch den Projektpartnern von KI:STE, PhenoRob und TrAgS sowie der Graduiertenschule HDS-LEE danke ich für viele bereichernde Meetings und wertvolle Impulse.

Zuletzt und von ganzem Herzen danke ich meiner Familie, meinen Freunden und meiner Partnerin Jana. Eure Unterstützung, euer Verständnis und euer Zuspruch haben mir geholfen, diese oft herausfordernde Zeit gut zu meistern.

The work presented in this thesis is partially supported by the German Federal Ministry for the Environment, Nature Conservation, and Nuclear Safety (67KI2043, KISTE); the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; EXC-2070—390732324—PhenoRob, and RO 4839/7-1 | STO 1087/2-1); and the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE). Many of the figures in this work are created using Matplotlib by Hunter, 2007.