UNIVERSITÄT BONN

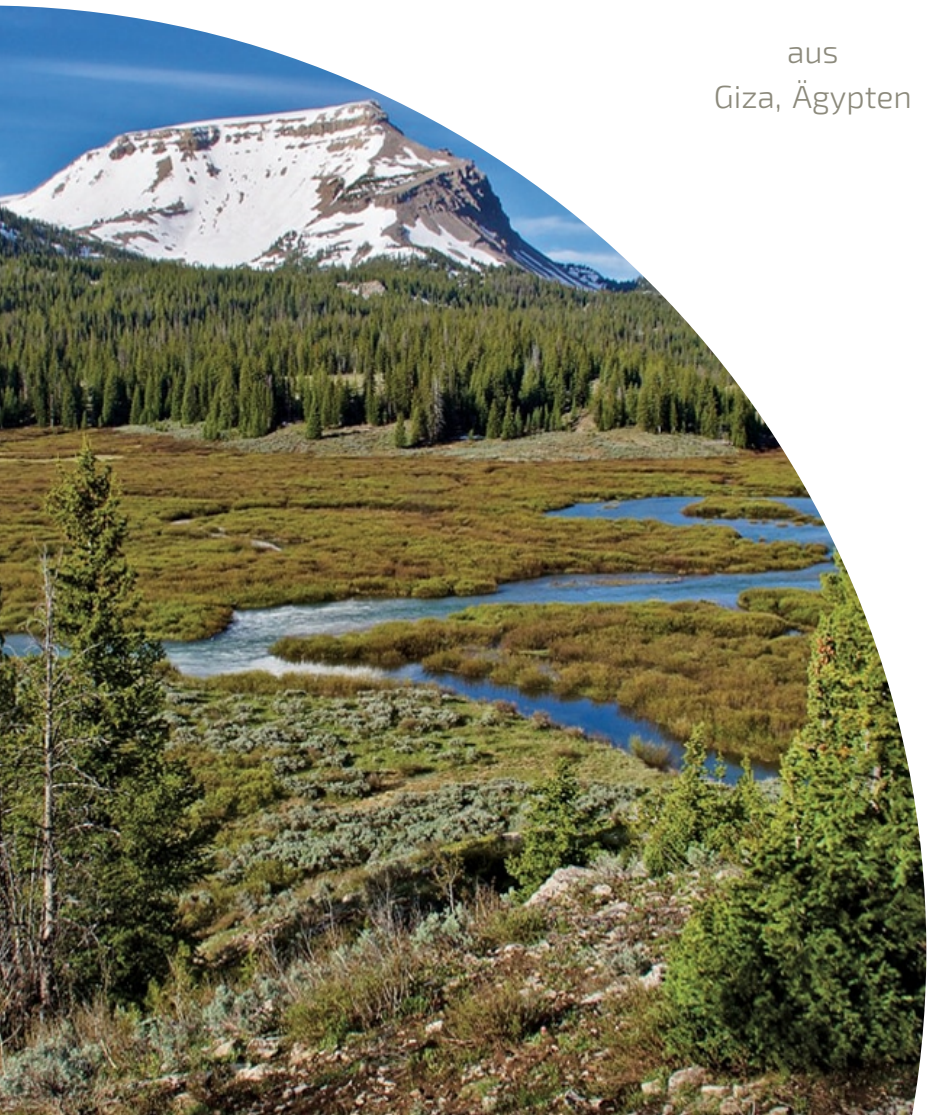
Dissertation
zur Erlangung des Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)
Agrar-, Ernährungs- und Ingenieurwissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn
Institut für Geodäsie und Geoinformation

Explainable Machine Learning-based Frameworks to Investigate The Appearance of Naturalness in Satellite Imagery

von

Ahmed Emam

aus
Giza, Ägypten



Bonn 2026

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

Korreferent:
Univ. Prof. Dr.-Ing. habil. Michael Schmitt, Universität der Bundeswehr München, München, Deutschland

Tag der mündlichen Prüfung: 26.02.2026

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

Zusammenfassung

Natürlichkeit ist ein zentraler Indikator für den Biodiversitätsschutz und das Funktionieren von Ökosystemen und damit von hoher Relevanz für Umweltmonitoring und Entscheidungsunterstützung. Diese Arbeit untersucht das Erscheinungsbild von Natürlichkeit in Satellitenbildern und identifiziert jene Muster und Landbedeckungsklassen, die am stärksten zu ihr beitragen, unter Anwendung erklärbarer Methoden des maschinellen Lernens (XAI) und der Unsicherheitsquantifizierung (UQ). Bestehende Ansätze weisen häufig Einschränkungen auf, wie eine begrenzte Interpretierbarkeit, einen Mangel an datengetriebener Attribution von Landbedeckungsklassen zur Natürlichkeit oder eine unzureichende Berücksichtigung der Unsicherheit in den Modellvorhersagen. Diese Schwächen können einzeln oder in Kombination zu Verzerrungen bei der Erklärung von Natürlichkeit führen. Um diese Lücken zu adressieren, werden drei komplementäre Frameworks eingeführt: [1] ein generatives Framework, das Aktivierungsmaximierung in CycleGAN integriert (AM-GANs for Naturalness) [1], um Bilder zu synthetisieren, die Natürlichkeitshinweise verstärken oder unterdrücken und so visuell aussagekräftige Attributionskarten erzeugen [2, 3]; [2] das Confident Naturalness Explanation (CNE) Framework [4], das semantische Segmentierung, Surrogatmodellierung und Monte-Carlo-Dropout kombiniert, um klassenweise CNE-Werte zu erzeugen, die den Beitrag jeder Landbedeckungsklasse zur Natürlichkeit sowie das Vertrauen in diese Attribution quantifizieren [5, 6]; und [3] Das Transformer-basierte Framework Reliable Explainability for Naturalness (NaT-ReX) [7] ist eine auf Vision Transformer basierende Architektur, die Layer-wise Relevance Propagation mit Attention Rollout und UQ integriert, um feinauflösende, pixelgenaue und unsicherheitsbewusste Relevanzkarten zu erzeugen, die zudem auf Klassenebene aggregiert werden können [8, 9]. Vergleiche mit etablierten Indizes – dem Human Influence Index [10] und dem Naturalness Index [11] – belegen die ökologische Plausibilität und verdeutlichen, wo die vorgeschlagenen Erklärungen mit externen Referenzen übereinstimmen oder von ihnen abweichen. In dieser Dissertation treten Feuchtgebiete, Buschland, Wälder sowie spärlich bewachsene Flächen konsistent als starke Prädiktoren für das visuelle Erscheinungsbild von Natürlichkeit hervor. Alle drei Frameworks wurden als peer-reviewed Forschungsarbeiten veröffentlicht.

Abstract

Naturalness is a key indicator for biodiversity conservation and ecosystem functioning, making its understanding and analysis highly relevant for environmental monitoring and decision support. This thesis explains the appearance of naturalness in satellite imagery and identifies which patterns and land-cover classes contribute most to it, using explainable machine learning (XAI) and uncertainty quantification methods. Existing approaches often suffer from limitations such as limited interpretability, a lack of data-driven attribution of land cover classes to naturalness, or insufficient consideration of the uncertainty associated with model predictions. These shortcomings, individually or in combination, can introduce bias in explaining naturalness. To address these gaps, three complementary frameworks are introduced: [1] the Activation Maximization within CycleGAN (AM-GANs for Naturalness) Framework [1], a generative approach that integrates activation maximization into CycleGAN to synthesize images that emphasize or suppress naturalness cues, producing visually interpretable attribution maps [2, 3]; [2] the Confident Naturalness Explanation (CNE) Framework [4], which combines semantic segmentation, surrogate modeling, and Monte Carlo dropout to produce class-wise CNE scores that quantify the contribution of each land cover class to naturalness together with the confidence in this attribution [5, 6]; and [3] The Transformer-Based Reliable Explainability for Naturalness (NaT-ReX) Framework [7], a vision transformer-based architecture that integrates Layer-wise Relevance Propagation with Attention Rollout and uncertainty quantification to produce fine-grained, pixel-level uncertainty-aware relevance maps that can also be aggregated at the class level [8, 9]. Comparisons with established indices, the Human Influence Index [10], and the Naturalness Index [11], demonstrate ecological plausibility and clarify where the proposed explanations converge or diverge from external references. In this dissertation, wetlands, shrublands, forests, and bare/sparse vegetation consistently emerge as strong contributors to the visual appearance of naturalness. All three frameworks have been published as peer-reviewed research studies.

Acknowledgements

I would like to express my profound gratitude to my first supervisor, Prof. Dr. Ribana Roscher, whose unwavering guidance, encouragement, and thoughtful feedback have been a constant source of strength throughout my doctoral journey. Her expertise not only shaped the direction and quality of this work but also inspired me to grow as an independent researcher. Beyond her academic mentorship, I am deeply thankful for the way she fostered an atmosphere of trust, kindness, and support, always attentive to us as individuals as much as scholars. Her generosity, patience, and genuine care left a lasting impression on me, both professionally and personally, and I feel truly privileged to be guided by her.

I am also deeply grateful to my second examiner, Prof. Dr. Michael Schmitt, for his constructive advice, critical input, and for broadening my perspective through his expertise. I greatly valued our enriching discussions and remain thankful for his readiness to provide support whenever I needed it.

My heartfelt thanks go to my colleagues and team members — Timo, Lukas, Eike, Jana, Johannes, Mohamad, Felix, and Lydia — whose collaboration, stimulating discussions, and constant encouragement created an inspiring and supportive research environment. Their contributions, both academic and personal, made this journey not only productive but also deeply enjoyable. I am especially grateful to Mohamed, my teammate and office companion. The countless discussions we shared — ranging from research ideas to everyday life — together with the laughter in between, enriched my PhD experience immeasurably, making it both intellectually rewarding and personally memorable.

I owe my deepest gratitude to my mother, who has always encouraged me to pursue science and to keep learning. Her countless sacrifices made it possible for me to receive the best possible education and ultimately to complete this dissertation.

Finally, I would like to thank Paula, my life companion, my closest confidante, and my partner in every sense of the word. Her love, patience, and unwavering faith in me gave me strength during moments of doubt and clarity in times of uncertainty. She has been my anchor and my source of joy, sharing both the challenges and the triumphs of this journey. Without her steadfast presence,

support, and belief in me, this work would have been far more difficult to achieve.

I am also grateful to the University of Bonn for providing an excellent academic environment and the resources that enabled me to pursue my doctoral research. More broadly, I would like to extend my sincere thanks to Germany as a country that allowed me to pursue both my master's and doctoral studies and welcomed me as a citizen and member of its community.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the following projects: RO 4839/5-1 and SCHM 3322/4-1 (project number 458156377 — MapInWild), RO 4839/6-1 (project number 459376902 — AID4Crops), and under the DFG's Excellence Strategy via the Cluster of Excellence *PhenoRob — Robotics and Phenotyping for Sustainable Crop Production* (EXC 2070 — 390732324).

Contents

1 Introduction **1**

1.1 Motivation . . . 1
1.2 Contributions . . . 3
1.3 Publications . . . 6
1.4 Collaborations . . . 6

2 Basic Techniques **9**

2.1 Notation . . . 9
2.2 Classical Models . . . 10
2.2.1 Linear Regression . . . 10
2.2.2 Logistic Regression . . . 10
2.3 Neural Networks . . . 14
2.3.1 Multilayer Perceptrons (MLPs) . . . 14
2.3.2 Convolutional Neural Networks (CNNs) . . . 15
2.3.3 Residual Networks (ResNet) . . . 16
2.3.4 Semantic Segmentation with DeepLab . . . 18
2.3.5 Generative Adversarial Networks (GANs) . . . 19
2.3.6 Attention-Based Architectures . . . 21
2.4 Explainable Machine Learning . . . 24
2.4.1 Interpretability vs. Explainability . . . 24
2.4.2 Taxonomy of Explanation Methods in XAI . . . 25
2.4.3 Inherently Interpretable Models . . . 25
2.4.4 Model Visualization Techniques . . . 25
2.4.4.1 Feature Maps . . . 25
2.4.4.2 Activation Maximization . . . 27
2.4.5 Feature Attribution Methods . . . 28
2.4.5.1 Deep Learning Important FeaTures (DeepLIFT) 28
2.4.5.2 Gradient-weighted Class Activation Mapping (Grad-CAM) . . . 29
2.4.6 Perturbation-based Methods . . . 30
2.4.7 Surrogate Models and Approximators . . . 31

2.4.8 Attribution in Attention-Based Models 32
2.5 Evaluation Criteria for Explanations 34
2.6 Uncertainty Quantification (UQ) 35
2.6.1 Confidence Intervals as a Measure of Uncertainty 35
2.6.2 Bayesian Neural Networks (BNNs) 36
2.6.3 Monte Carlo Dropout 36
2.6.4 Calibration of Uncertainty Estimates 37

3 Related Work **39**
3.1 Taxonomy of Naturalness 39
3.2 Heuristic Scoring Approaches to Naturalness Mapping 41
3.2.1 The Human Influence Index (HII) 41
3.2.2 The Naturalness Index (NI) 42
3.2.3 Naturalness Assessment in the Bavarian Forest 42
3.2.4 Multimodal Learning for Naturalness Estimation 43
3.3 XAI-based Frameworks for Naturalness Investigation 44
3.3.1 Latent Space Occlusions to Assess Naturalness 44
3.3.2 Modality Occlusions for Estimating Modality Contribution to Naturalness 45

4 Data **47**
4.1 AnthroProtect Dataset 47
4.1.1 Ecological Coherence 47
4.1.2 Label Design 47
4.1.3 Imagery Source and Preprocessing 48
4.1.4 Dataset Composition 48
4.1.5 Data Splitting Strategy 48
4.2 MapInWild Dataset 50
4.2.1 Dataset Design, Modalities, and Composition 50
4.2.2 Sampling and Curation Strategy 50
4.2.3 Quality Scoring System 51

5 GAN-based Activation Maximization for Naturalness Explanations **55**
5.1 State of the Art 56
5.2 Methodology 57
5.2.1 Regressor Training (Pattern Learning) 57
5.2.2 GANs Training (Pattern Enhancement) 57
5.2.3 Attribution Mapping 59
5.3 Experimental Setup 60
5.3.1 Regressor Training and Performance Evaluation 60

5.3.2 Generative Model Training and Evaluation 61
5.4 Results 62
5.4.1 Qualitative Evaluation of the Attribution Maps 62
5.4.2 Quantitative Results 62
5.5 Discussion 64
5.6 Conclusion 64

6 Surrogate Modeling for Confident Naturalness Explanations 67
6.1 State of the Art 69
6.2 Method 70
6.2.1 Semantic Segmentation and Pattern Vectorization 71
6.2.2 Global Surrogate Model: Logistic Regression 72
6.2.3 Uncertainty Quantification 72
6.2.4 Confident Naturalness Explanation (CNE) Metric 73
6.2.5 Experimental Setup 73
6.3 Results 76
6.3.1 Qualitative Results 76
6.3.2 Quantitative results 76
6.4 Discussion 78
6.5 Conclusion 78

7 Naturalness Assessment with Transformer-Based Reliable Explainability 83
7.1 State of the Art 84
7.1.1 Explainability Approaches for ViT-based architectures . . 84
7.1.2 Quantifying Uncertainty for ViT-based architectures . . . 85
7.1.3 Integration of XAI and UQ techniques 87
7.2 Methodology 88
7.2.1 NaT-ReX Architecture and Training Strategy 88
7.2.2 Reconstruction Head in inference: Uncertainty Quantification via MC-Dropout 89
7.2.3 ReX Score: Combining Naturalness, Relevance, and Uncertainty 89
7.2.4 Experimental Setup 90
7.3 Results 92
7.3.1 Qualitative Evaluation 92
7.3.2 Quantitative Evaluation 92
7.4 Discussion 97
7.5 Conclusion 97

8 Quantitative and Ecological Analysis of Patterns Forming Naturalness in Fennoscandia **99**
8.1 Comparing Frameworks to Naturalness Indices 99
8.1.1 Alignment of AM-GANs Scores with NI and HII 100
8.1.2 Alignment of CNE Scores with NI and HII 101
8.1.3 Alignment of ReX Scores with NI and HII 102
8.1.4 Correlation between the Proposed and the Established Frameworks 103
8.1.5 Sources of Variations in The Proposed Frameworks 104
8.2 Ecological Significance of Key Contributors to Naturalness 105
8.2.1 Shrublands 106
8.2.2 Wetlands 106
8.2.3 Bare and Sparsely Vegetated Areas 107
8.2.4 Forests 109
8.2.5 Natural Grasslands 109
8.2.6 Glaciers, Snow, and Ice 111

9 Conclusion **114**
9.1 Short Summary of Key Contributions 114
9.2 Open Source Contributions 116
9.3 Collaborative Projects and Contributions 116
9.4 Future Work 118
9.4.1 Toward Ecology-Oriented Foundation Models 118
9.4.2 Efficient Uncertainty Quantification 118
9.4.3 Better on-the-ground data and better segmentation maps . 119
9.4.4 Integrate knowledge from text 119

10 Appendix **121**

Chapter 1

Introduction

1.1 Motivation

Naturalness, as used in this thesis, refers to protected natural areas characterized by minimal modern human influence [10, 4]. These areas are ecologically significant, serving as reference points for intact natural systems [10, 12]. Such regions often maintain higher biodiversity, intact ecological processes, and stable habitat structures, making them critical baselines for assessing environmental change and human pressure [13]. Beyond their ecological value, natural areas provide reference conditions for conservation strategies, guide restoration efforts, and serve as benchmarks for evaluating the effectiveness of protected area management [14]. Previous research has attempted to explain naturalness, but these approaches have predominantly relied on predefined assumptions [10, 15, 16, 11]. Conceptualizing naturalness as a fixed construct can limit its applicability in Earth observation research, as it tends to reflect subjective perceptions and introduces systematic biases. For instance, naturalness is frequently equated with the presence of greenery, whereas ecologically relevant environments such as shrublands are often neglected. To overcome these limitations, this thesis explores data-driven frameworks to identify spatial patterns and features in satellite imagery that contribute to the appearance of naturalness and help to mitigate subjective bias.

Machine learning (ML) and deep learning models have been increasingly adopted in Earth observation and remote sensing applications [18, 19]. These technologies are now central to tasks such as land cover classification, ecosystem monitoring, and detecting environmental changes. In domains that directly support conservation strategies and the management of protected natural areas, the reliability and transparency of these models are essential [20, 21, 22].

Many of the most powerful methods—such as deep neural networks are considered black boxes, as their internal decision-making mechanisms are not

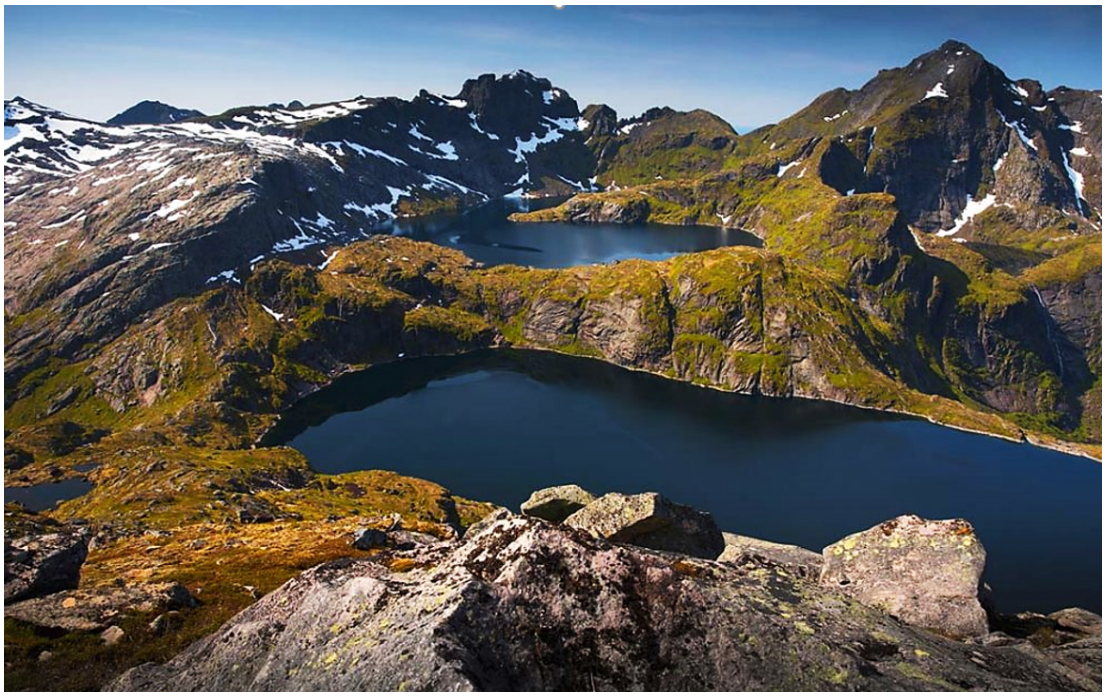


Figure 1.1: High-naturalness landscape in Fennoscandia with minimum human influence: glaciated mountains, alpine lakes, and intact tundra–rock vegetation with no visible infrastructure [17].

directly interpretable to humans [23]. Consequently, these models may rely on spurious or misleading correlations, imbalanced class distributions, or features unrelated to the underlying ecological processes, leading to high performance metrics that do not necessarily reflect ecologically meaningful learning [20, 24]. For example, a classifier could rely on sensor artifacts, atmospheric conditions, or landscape edges that co-occur with certain land cover classes rather than the true underlying ecological structures [25]. Such drawbacks often go unnoticed in standard accuracy metrics but can lead to unreliable behavior when applied to new regions or temporal conditions.

Explainable Machine Learning or Explainable Artificial Intelligence (XAI) addresses this challenge by enhancing the interpretability and transparency of model decisions [24, 20]. In this thesis, the terms Explainable Machine Learning and Explainable Artificial Intelligence are used interchangeably. XAI methods reveal which input features drive the model's predictions. A model is considered explainable if these influences can be traced and related to established ecological knowledge. For example, an explanation method may highlight that vegetation density or the absence of infrastructure strongly contributes to a prediction of high naturalness, which may directly align with expert expectations [24]. In this way, explainability not only supports model validation but also allows experts to verify whether the model is reasoning for ecologically meaningful reasons. Furthermore,

XAI can uncover new patterns in the data, pointing to relationships that may not have been anticipated by experts, thereby contributing to scientific discovery [24, 26].

Uncertainty quantification (UQ) refers to the process of measuring and communicating the level of uncertainty associated with model predictions [6, 27]. Rather than providing a single deterministic prediction, UQ attaches an estimate of confidence, indicating how much trust should be placed in a prediction. In ecological applications, this distinction is critical: two areas may both be predicted as highly natural, but one prediction may be made with high confidence while the other is highly uncertain due to scarce training data, mixed land cover, or atypical spectral signatures [27]. Treating all predictions as equally reliable can lead to flawed decisions, for instance, by prioritizing areas where the model is actually uncertain [4]. Uncertainty estimates help distinguish between predictions that can be trusted and those that require caution or additional field validation [7, 28]. Without UQ, even interpretable models may provide misleading signals if all outputs are assumed to be equally certain. For naturalness explanations, uncertainty-aware decision-making therefore requires systems that explain their predictions while also communicating the confidence associated with them [29, 4].

1.2 Contributions

This thesis presents a set of frameworks designed to enhance our understanding of the appearance of naturalness in satellite imagery. Building on the idea that naturalness is not treated as a fixed or predefined concept, the proposed approaches combine XAI methods with UQ to identify and interpret spatial patterns and land cover classes associated with high naturalness. Specifically, the contributions include:

1. A generative adversarial framework, named in this thesis AM-GANs for Naturalness [1], that produces heatmaps to highlight features in satellite imagery contributing to naturalness. Based on adversarial training[30] and activation maximization [31], this method enables the identification of latent knowledge gained by ML models that distinguish areas of high naturalness. These interpretable representations are then used to quantify the contribution of each land cover class to naturalness [1].

2. A surrogate modeling framework, named the Confident Naturalness Explanation (CNE) framework [4], is proposed to approximate the behavior of a complex black-box model using an inherently interpretable surrogate [32]. This design enables global explanations of naturalness predictions while maintaining high fidelity to the original model's behavior. In parallel,

the UQ technique is employed to assess the confidence associated with each explanatory factor, offering a more interpretable and uncertainty-aware understanding of the model's insights into naturalness. Additionally, the CNE index is introduced to quantitatively assess the contribution of each land cover class to naturalness, along with the associated uncertainty [4].

3. A Transformer-based multitask learning framework, called NaT-ReX [7], is introduced to integrate XAI and UQ for assessing naturalness in protected areas [8]. NaT-ReX employs Layer-wise Relevance Propagation (LRP) Attention Rollouts [9] to compute pixel-level relevance maps and Monte Carlo Dropouts [6] to estimate the associated uncertainty at each pixel. Based on these outputs, a novel metric—the Reliable Explanation (ReX) score—is proposed, capturing both the importance and the certainty of each pixel's contribution to naturalness. To derive more practical insights, these pixel-level ReX scores are averaged within each land cover class, allowing for an interpretable assessment of which land cover classes contribute most confidently to the concept of naturalness across a given region [7].

To clarify the scope of this work, it is important to emphasize what this thesis does not aim to address. It does not involve causal inference, fairness in AI, or the development of general-purpose explainability tools. Furthermore, it does not engage with language-based models or text-driven applications. This thesis is situated in the visual domain, with a specific focus on Earth observation and naturalness understanding. The objective is to understand and quantify the contributions of each pixel and each land cover class to the appearance of naturalness—alongside the uncertainty associated with these insights.

Importantly, this thesis does not attempt to analyze how humans perceive or define naturalness. No user studies or surveys were conducted, and no subjective assumptions about the meaning of naturalness were imposed. Instead, the thesis proposes a set of transparent, data-driven, and uncertainty-aware machine learning frameworks that aim to explain patterns of naturalness as learned by models from the data itself.

The outcomes of this research have been published in peer-reviewed journals and conferences, and the developed code and models are made publicly available to support reproducibility and future research.

Summary of Contributions

- Development of three novel frameworks: AM-GANs for Naturalness, CNE, and NaT-ReX, which combine XAI and UQ for naturalness analysis.

- To the best of our knowledge, the first to integrate activation maximization into a generative adversarial framework for interpretable attribution of land cover features, and to jointly incorporate XAI and UQ to enhance human understanding of the appearance and perception of naturalness.
- Public release of code and models, with results published in peer-reviewed venues to support reproducibility and transparency.

This document is organized as follows:

- **Chapter 2** introduces basic machine learning techniques used in this thesis, including core models for regression and classification, generative adversarial networks, semantic segmentation networks, and transformer-based architectures. It also provides an overview of XAI methods and UQ approaches.
- **Chapter 3** presents related work on the concept of naturalness, particularly in the context of human influence, ecological integrity, and protected areas. It reviews existing mapping approaches and discusses the role of machine learning in environmental monitoring.
- **Chapter 4** describes the datasets used in this work, including the AnthroProtect [12] and MapInWild [33] datasets, which provide diverse environmental samples across protected and non-protected areas.
- **Chapter 5** introduces the first framework, AM-GANs for Naturalness [1], based on generative adversarial training. It uses activation maximization to generate interpretable heatmaps highlighting image features that contribute to naturalness, and quantifies the contribution of land cover classes.
- **Chapter 6** presents the second framework, Confident Naturalness Explanation (CNE) [4], which employs a surrogate modeling approach to explain black-box models through inherently interpretable models. It integrates UQ to assess the reliability of the explanatory factors contributing to naturalness.
- **Chapter 7** introduces the third framework, Transformer-based Reliable Explainability for Naturalness (NaT-ReX) [7], which combines XAI and UQ within a transformer-based architecture. It integrates Layer-wise Relevance Propagation [9] and Monte Carlo Dropout [6] to quantify the class-wise contribution and uncertainty of land cover classes. A novel metric, the ReX score, is proposed.

- **Chapter 8** provides a comprehensive analysis of the naturalness predictions generated by the proposed methods. It discusses regional differences, land cover influences, and patterns across the Fennoscandian landscape.
- **Chapter 9** concludes the thesis by summarizing the main contributions, highlighting open-source components, and outlining possible directions for future work.

1.3 Publications

The following peer-reviewed publications have contributed to the development of this thesis. Each publication has been integrated into the corresponding chapters and revised or extended to ensure coherence and completeness.

Published and Accepted Papers

- A. Emam, T. T. Stomberg, R. Roscher, "Leveraging Activation Maximization and Generative Adversarial Training to Recognize and Explain Patterns in Natural Areas in Satellite Imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024, Art. no. 8500105, DOI: `https://doi.org/10.1109/LGRS.2023.3335473`.
- A. Emam, M. Farag, R. Roscher, "Confident Naturalness Explanation (CNE): A Framework to Explain and Assess Patterns Forming Naturalness," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024, Art. no. 8500505, DOI: `https://doi.org/10.1109/LGRS.2024.3365196`.
- A. Emam, M. Farag, M. Russwurm, and R. Roscher, "NaT-ReX: A Transformer-based Framework to Reliably Explain Naturalness," Accepted at *DAGM German Conference for Pattern Recognition, Freiburg, Germany, 2025.*
- A. Emam, R. Roscher, "Confidence-Filtered Relevance (CFR): An Interpretable and Uncertainty-Aware Machine Learning Framework for Naturalness Assessment in Satellite Imagery," Accepted at *Workshop on Machine Learning for Earth Observation In Conjunction with the ECML/PKDD Portugal, 2025.*

1.4 Collaborations

This thesis also benefited from two interdisciplinary collaborations in the field of digital agriculture, where the methods and perspectives developed in the core research were adapted to agricultural decision-support systems.

- The first collaboration, *"Ahmed Emam, Mohamed Farag, Jana Kierdorf, Lasse Klingbeil, Uwe Rascher and Ribana Roscher. A Framework for Enhanced Decision Support in Digital Agriculture Using Explainable Machine Learning. Computer Vision – ECCV 2024 Workshops, pages 31–45. Springer Nature Switzerland."*, focused on integrating interpretability into model selection and evaluation pipelines for agricultural tasks. In this project, we proposed a three-stage, model-agnostic framework to assess both local and global interpretability of deep learning models, specifically Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) [8]. The framework evaluates models not only based on performance but also on the alignment of their explanations—derived from attribution techniques such as Grad-CAM [34] and attention maps—with domain knowledge. This collaboration highlighted the importance of explainability for building trustworthy AI systems in real-world agricultural environments [35].

- The second collaboration, *"Mohamed Farag, Ahmed Emam, Johannes Leonhardt, and Ribana Roscher. Enhancing decision support in crop production: Analyzing conformal prediction for uncertainty quantification. Computers and Electronics in Agriculture, 237:110559, 2025."*, addressed the challenge of uncertainty in machine learning-based agricultural forecasting. In this work, we investigated conformal prediction as a model-agnostic, distribution-free approach to quantifying both epistemic and aleatoric uncertainty. Through a series of experiments, we benchmarked conformal prediction [36] against more established techniques such as deep ensembles and softmax-based confidence estimation. This work demonstrated the value of reliable UQ in supporting robust, transparent, and risk-aware decision-making in crop production [37].

	AM-GANs for Naturalness [1]	**CNE [4]**	**NaT-ReX [7]**
Data-driven	✓	✓	✓
XAI	✓	✓	✓
UQ	✗	✓	✓
Pixel-level insights	✓	✗	✓
LC class-level insights	✓	✓	✓

Table 1.1: Comparison of the three developed frameworks forming the main components of this thesis. The AM-GANs for Naturalness framework [1] leverages XAI, the CNE framework [4] adds UQ and provides land-cover class-level insights, while the NaT-ReX framework [7] integrates both XAI and UQ for both class- and pixel-level insights.

Chapter 2

Basic Techniques

This chapter introduces the basic machine learning and deep learning techniques that provide the foundation for the methods developed in later chapters or those discussed in related state-of-the-art work. To keep the thesis self-contained, each technique is briefly outlined with its key concepts and mathematical basis.

The chapter begins with classical models such as linear and logistic regression, followed by deep learning architectures including convolutional neural networks (CNNs) [38] and residual networks (ResNets) [39] [40]. Generative adversarial networks (GANs) [41] are introduced as frameworks for data synthesis, while models like DeepLabv3 [5] are discussed for semantic segmentation. Finally, attention-based architectures, particularly vision transformers (ViTs) [8, 42], are presented as alternatives to convolutional models, enabling global context modeling through self-attention.

2.1 Notation

To ensure clarity and consistency, this thesis follows a standardized notation for representing mathematical objects, models, and data:

- Matrices are denoted by upright sans-serif uppercase letters. For example, an image is represented as $\mathsf{X} \in \mathbb{R}^{W \times H \times C}$, where W and H are the spatial dimensions and C is the number of channels. Unless otherwise specified, $C = 3$ is assumed, corresponding to RGB images.

- Vectors are written as bold lowercase letters. For example, $\boldsymbol{x} \in \mathbb{R}^d$ may denote a flattened image or a feature vector.

- Scalars are written as standard lowercase italic letters. For instance, $y \in \{0, 1\}$ represents a binary class label, and $\sigma(z) \in [0, 1]$ denotes the output of a sigmoid activation function.

- Functions are denoted using script font. For example, $\mathcal{L}(\boldsymbol{w}, b)$ represents a loss function such as binary cross-entropy, and $\mathcal{F}(\boldsymbol{x})$ may denote the function learned by a neural network or model block.

- Transpose operations are denoted by a superscript $^{\mathsf{T}}$.

Images in this thesis are represented as $\mathsf{X} \in \mathbb{R}^{W \times H \times C}$, and are typically found in datasets of N samples: $\mathsf{X} = [\mathsf{X}^{(1)}, \mathsf{X}^{(2)}, \ldots, \mathsf{X}^{(N)}]$. Each image $\mathsf{X}^{(i)}$ is associated with a binary label $y^{(i)} \in \{0, 1\}$, resulting in a labeled dataset $\{(\mathsf{X}^{(i)}, y^{(i)})\}_{i=1}^{N}$. Throughout this thesis, the indices i, j will be explicitly defined when they refer to a spatial location, a specific pixel, or an image patch.

2.2 Classical Models

Linear models assume a linear relationship between the input features and the target variable. They are among the most fundamental tools in statistical learning and provide a useful reference point for more complex models. This subsection introduces linear regression and logistic regression. [40, 43].

2.2.1 Linear Regression

Linear regression models the relationship between a real-valued target $y \in \mathbb{R}$ and an input vector $\boldsymbol{x} \in \mathbb{R}^d$. The target is assumed to be a linear combination of the input features [40, 43]:

$$\hat{y} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b \tag{2.1}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ is the weight vector, $b \in \mathbb{R}$ is the bias, $y^{(i)}$ is the true target, and $\hat{y}^{(i)} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}^{(i)} + b$ is the prediction. The parameters are learned by minimizing the mean squared error (MSE):

$$\mathcal{L}(\boldsymbol{w}, b) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 \tag{2.2}$$

The parameters can be estimated using gradient descent or closed-form solutions (e.g., the normal equation). Linear regression assumes normally distributed residuals and low multicollinearity among input features [40, 43]. Figure 2.1 presents a simple illustrative example of linear regression.

2.2.2 Logistic Regression

Logistic regression is an essential model for binary classification. It estimates the probability that a given input feature vector $\boldsymbol{x} \in \mathbb{R}^d$ belongs to class 1 (i.e.,

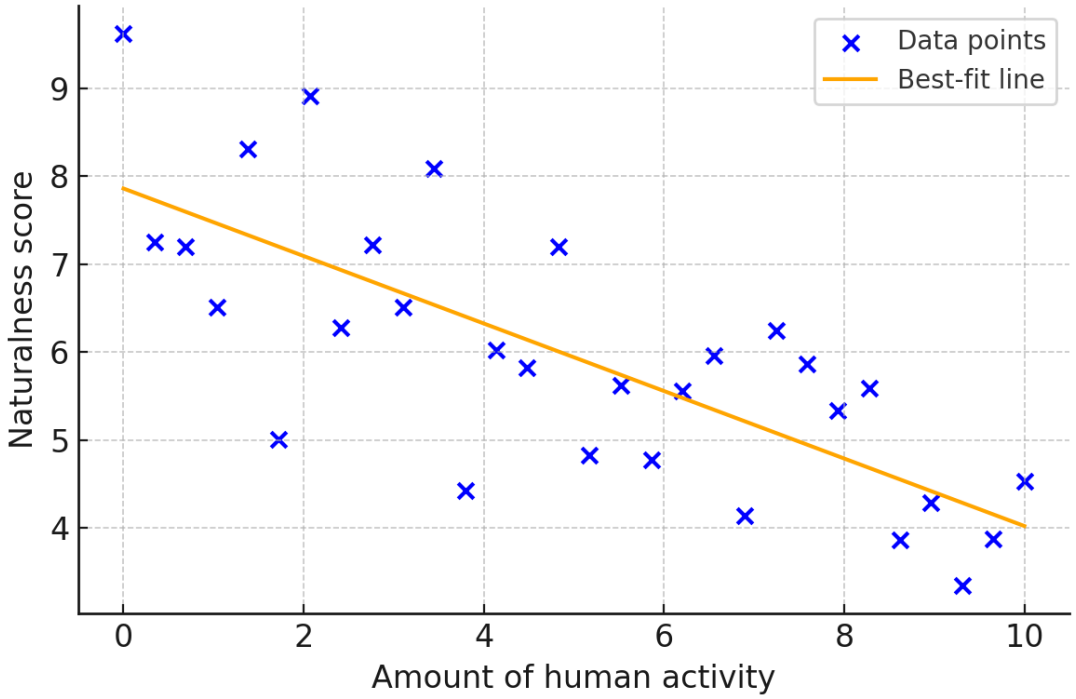


Figure 2.1: The plot illustrates a negative linear relationship between the amount of human activity and the naturalness score, showing that higher levels of human activity are associated with lower naturalness. The underlying data were generated by the author solely for illustrative purposes to demonstrate how linear regression works.

$y = 1$), where $y \in \{0, 1\}$. The model computes a linear score [40, 43]:

$$z = \boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b \tag{2.3}$$

Which is transformed into a probability using the sigmoid function:

$$p(y = 1 \mid \boldsymbol{x}) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.4}$$

The predicted output is thus:

$$\hat{y} = \sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b) \tag{2.5}$$

A binary prediction is obtained by thresholding the probability:

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{y} > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

Training involves minimizing the binary cross-entropy loss:

$$\mathcal{L}(\boldsymbol{w}, b) = -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\log(\hat{y}^{(i)}) + (1 - y^{(i)})\log(1 - \hat{y}^{(i)})\right] \tag{2.7}$$

Gradient descent is used to update the parameters:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})\boldsymbol{x}^{(i)} \tag{2.8}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)}) \tag{2.9}$$

Logistic regression defines a linear decision boundary via:

$$\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b = 0 \tag{2.10}$$

and the model's output $\hat{y}$ represents the confidence in class membership. The learned weights $\boldsymbol{w}$ can be interpreted: a positive weight increases the probability of predicting class 1 as the corresponding feature increases, while a negative weight decreases it.

The model relies on several assumptions: (1) the log-odds are linearly related to the input features, (2) data points are independent, (3) input features are not strongly collinear, and (4) the sample size is sufficient to estimate parameters reliably. Figure 2.2 presents a simple illustrative example of logistic regression. [44, 40, 43].

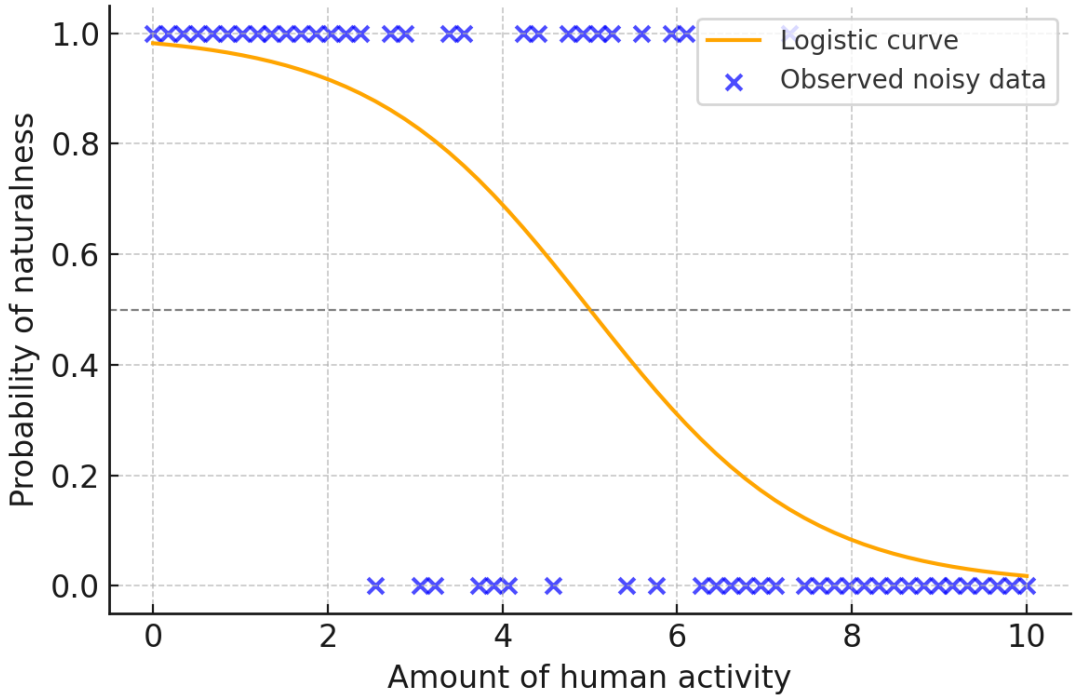


Figure 2.2: The plot illustrates logistic regression modeling the inverse relationship between human activity and the probability of naturalness. The orange curve shows the predicted probability, while the blue points represent class labels, assigned as 1 (naturalness) when the probability exceeds 0.5 and 0 (non-naturalness) otherwise. The underlying data were generated by the author solely for illustrative purposes to demonstrate how logistic regression works.

2.3 Neural Networks

2.3.1 Multilayer Perceptrons (MLPs)

MLPs are the simplest feedforward neural networks, consisting of layers of linear transformations followed by non-linear activations. Given an input $\boldsymbol{x} \in \mathbb{R}^d$, a hidden layer with weights $\mathsf{W} \in \mathbb{R}^{m \times d}$, bias $\boldsymbol{b} \in \mathbb{R}^m$, and activation $\sigma(\cdot)$ computes

$$\boldsymbol{h} = \sigma(\mathsf{W}\boldsymbol{x} + \boldsymbol{b}), \tag{2.11}$$

where d is the input dimension and m the number of hidden units. Stacking multiple layers forms a deep MLP.

The output layer maps the final hidden representation to predictions. For binary classification with output $\hat{y} \in [0, 1]$, this can be expressed as:

$$\hat{y} = \sigma(\boldsymbol{w}^\mathsf{T}\boldsymbol{h} + b) \tag{2.12}$$

where $\boldsymbol{w} \in \mathbb{R}^m$ and $b \in \mathbb{R}$ are the weight vector and bias of the output layer.

Although MLPs can approximate complex functions, they treat the input as a flat vector and ignore spatial or structural information, making them less efficient for image data. This limitation motivates the use of specialized architectures such as Convolutional Neural Networks (CNNs) [40].

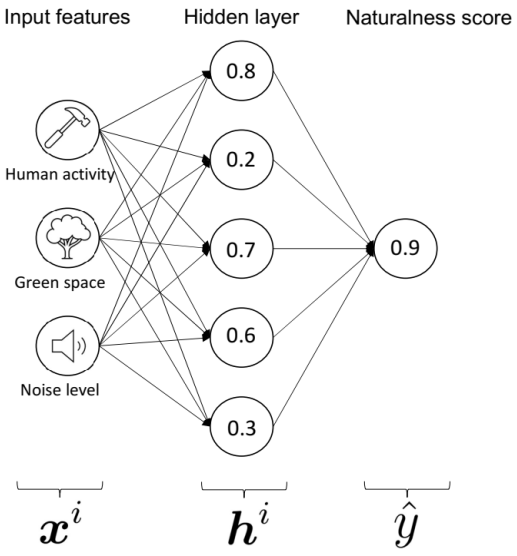


Figure 2.3: Schematic of a single-hidden-layer multilayer perceptron (MLP). The input features $\boldsymbol{x}^{(i)}$ (e.g., human activity, green space, noise level) are transformed through a hidden layer into intermediate representations $\boldsymbol{h}^{(i)}$, which are subsequently mapped to the output $\hat{y}$, representing the predicted naturalness score.

2.3.2 Convolutional Neural Networks (CNNs)

CNNs are a class of deep neural networks designed to process data with a grid-like topology, such as images. The core idea of CNNs is to leverage local spatial correlations through convolution operations, enabling efficient learning of hierarchical representations [40, 43, 45].

Each convolutional layer applies a kernel $\mathsf{K} \in \mathbb{R}^{k \times k}$ across spatial locations of the input $\mathsf{X} \in \mathbb{R}^{H \times W}$ to produce an output feature map $\mathsf{Y} \in \mathbb{R}^{H' \times W'}$. The output at spatial location (u, v) is computed as:

$$\mathsf{Y}_{u,v} = (\mathsf{X} * \mathsf{K})_{u,v} = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} \mathsf{X}_{u+m,v+n} \mathsf{K}_{m,n} \tag{2.13}$$

where u, v index the spatial location in the output feature map, m, n index positions within the kernel window, and k denotes the kernel size.

CNNs typically consist of the following components [40, 43]:

- Convolutional layers: extract spatial features by sliding learnable filters over the input.
- Activation functions: introduce non-linearities into the network, enabling it to model complex, non-linear relationships in the data. Common examples include ReLU and the Sigmoid function, which allow the network to capture detailed feature interactions beyond linear transformations [45].
- Pooling layers: perform downsampling (e.g., max pooling) to reduce spatial resolution and increase translational invariance.
- Fully connected layers: flatten and map the high-level features into output predictions.

For binary classification, the final layer applies a sigmoid activation:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad z = \boldsymbol{w}^{\mathsf{T}} \boldsymbol{h} + b \tag{2.14}$$

where $\boldsymbol{h}$ is the final hidden feature vector, $\boldsymbol{w} \in \mathbb{R}^d$ is a weight vector, and $b \in \mathbb{R}$ is the bias.

The model is trained using the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^{n} \left[y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right] \tag{2.15}$$

CNNs are also applicable to regression tasks by replacing the final sigmoid with a linear activation and using mean squared error (MSE) as the loss:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)} \right)^2 \tag{2.16}$$

In semantic segmentation, CNNs are adapted to output dense predictions. Fully Convolutional Networks (FCNs) [46] replace fully connected layers with convolutional layers to produce pixel-level class scores. The output is usually upsampled to match the input resolution using deconvolution layers or bilinear interpolation. Models such as U-Net [47] and DeepLab [5] extend this design by integrating skip connections and atrous convolutions to preserve fine spatial details and capture multi-scale context.

Historically, CNNs gained popularity with LeNet-5 [48] for digit classification, followed by AlexNet [45]. Subsequent architectures such as VGG [49], GoogLeNet [38], and ResNet [39] progressively increased network depth and introduced architectural innovations like residual connections to improve training stability and accuracy on large-scale vision tasks.

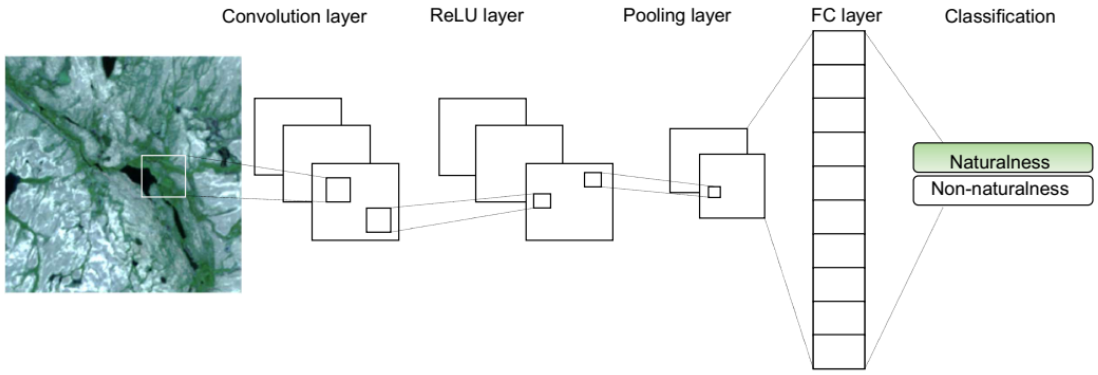


Figure 2.4: Convolutional neural network to classify areas of naturalness. Convolution extracts local features, ReLU activation function adds nonlinearity, pooling downsamples, and the flattened features feed a fully connected layer that outputs probabilities for the classes Naturalness and Non-naturalness.

2.3.3 Residual Networks (ResNet)

ResNet, introduced by He et al. [39], addressed the degradation problem observed in very deep neural networks. As network depth increases, traditional architectures often suffer from higher training error, not due to overfitting, but because deeper models become harder to optimize. This difficulty arises from vanishing gradients and disrupted signal propagation during backpropagation.

ResNet mitigates these limitations by introducing residual connections, which allow layers to learn modifications (residuals) to the input rather than the full transformation. Instead of modeling a target function $\mathcal{H}(\boldsymbol{x})$, the network learns a residual function $\mathcal{F}(\boldsymbol{x}) = \mathcal{H}(\boldsymbol{x}) - \boldsymbol{x}$, resulting in the reformulated expression:

$$\mathcal{H}(\boldsymbol{x}) = \mathcal{F}(\boldsymbol{x}; \mathsf{W}) + \boldsymbol{x} \tag{2.17}$$

A typical residual block is defined as:

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \mathsf{W}) + \boldsymbol{x} \tag{2.18}$$

If the dimensions of $\boldsymbol{x}$ and $\mathcal{F}(\boldsymbol{x}; \mathsf{W})$ do not match, a projection is introduced:

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \mathsf{W}) + \mathsf{W}_s \boldsymbol{x} \tag{2.19}$$

where W_s is a learnable projection matrix.

Residual connections preserve both signal and gradients, enabling the training of very deep networks such as ResNet-50, ResNet-101, and ResNet-152 [39]. A typical architecture begins with a convolution and pooling layer, followed by stacked residual blocks, and ends with global average pooling and a fully connected output. By improving optimization and gradient flow, ResNet made training networks with hundreds of layers feasible. Its principles have since become standard in image classification, detection, segmentation, and as backbones for vision transformers [39, 43].

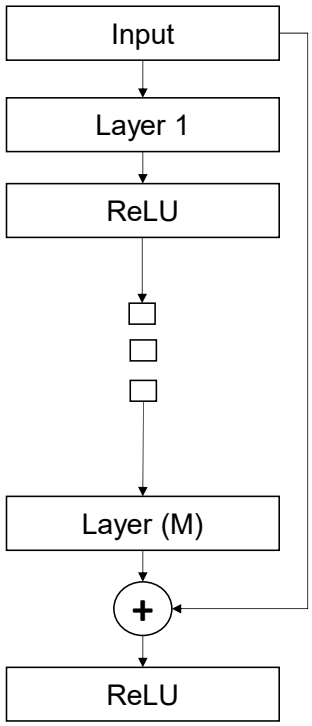


Figure 2.5: Residual block in ResNet. Deep networks learn a residual function $\mathcal{F}(\boldsymbol{x}; \mathsf{W})$ that is added to an identity shortcut to preserve signal and gradients, addressing degradation in very deep models. The block output is $\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \mathsf{W}) + \boldsymbol{x}$; if dimensions differ, the shortcut uses a projection W_s so that $\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \mathsf{W}) + \mathsf{W}_s \boldsymbol{x}$ [39].

2.3.4 Semantic Segmentation with DeepLab

DeepLabv3 [5] is a CNN-based architecture for semantic segmentation, where the goal is to assign a class label to each pixel of an input image. To balance spatial detail and contextual understanding, DeepLabv3 introduces atrous (dilated) convolutions and a multi-scale context aggregation module called Atrous Spatial Pyramid Pooling (ASPP).

Atrous convolution enlarges the receptive field by inserting zeros between kernel elements without increasing the number of parameters or downsampling the feature map. The one-dimensional case is defined as

$$y[u] = \sum_{k} x[u + r \cdot k] \cdot w[k], \tag{2.20}$$

where $y[u]$ is the output at position u, r is the dilation rate, and k indexes the kernel elements. Its two-dimensional counterpart is given by

$$\mathsf{Y}_{u,v} = \sum_{m} \sum_{n} \mathsf{X}_{u+r\cdot m,\, v+r\cdot n} \cdot \mathsf{W}_{m,n}, \tag{2.21}$$

where $\mathsf{Y}_{u,v}$ denotes the output value at pixel location (u, v), X is the input feature map, W is the convolution kernel, and m, n index positions within the kernel window.

To capture multi-scale context, ASPP applies parallel atrous convolutions with different dilation rates to the same input, including a 1×1 convolution, several 3×3 atrous convolutions, and a global average pooling branch. The resulting feature maps $\mathsf{f}_1, \ldots, \mathsf{f}_5$ are concatenated and fused:

$$\mathsf{F}^{\text{asp}} = (\mathsf{f}_1 \parallel \mathsf{f}_2 \parallel \mathsf{f}_3 \parallel \mathsf{f}_4 \parallel \mathsf{f}_5) *_{1\times1} \mathsf{W} \tag{2.22}$$

where $\parallel$ denotes concatenation, $*_{1\times1}$ denotes a 1×1 convolution with kernel W, and F^{asp} is the fused ASPP feature map.

The fused representation is upsampled to the input resolution using bilinear interpolation. Class probabilities are then obtained per pixel by applying a softmax:

$$P_{u,v}(c) = \frac{e^{z^c_{u,v}}}{\sum_{c'=1}^{C} e^{z^{c'}_{u,v}}}, \tag{2.23}$$

where $z^c_{u,v}$ denotes the logit for class c at pixel (u, v), C is the total number of classes, and c' is a dummy class index used in the normalization term to sum over all classes.

The model is trained by minimizing the categorical cross-entropy loss over all pixels:

$$\mathcal{L} = -\frac{1}{N} \sum_{u=1}^{H} \sum_{v=1}^{W} \sum_{c=1}^{C} y^c_{u,v} \log P_{u,v}(c) \tag{2.24}$$

where $y^c_{u,v} \in \{0, 1\}$ is the ground truth indicator for class c at pixel (u, v), H and W are the image dimensions, and $N = H \cdot W$ is the total number of pixels.

DeepLabv3 typically uses a CNN backbone such as ResNet [39] for feature extraction. The extended version, DeepLabv3+ [50], fuses high-level semantic features from deeper layers with low-level features from earlier layers that preserve fine spatial details such as object edges, improving segmentation accuracy especially at boundaries.

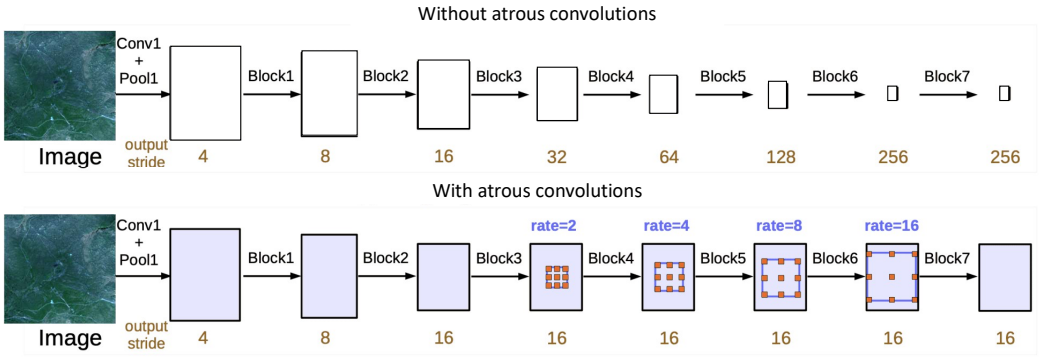


Figure 2.6: Concept of atrous (dilated) convolution in DeepLab for preserving spatial resolution. top: In a standard CNN, the output stride increases with network depth (e.g., 4, 8, 16, 32), progressively reducing spatial resolution. bottom: DeepLab replaces the later downsampling operations with atrous convolutions using increasing dilation rates (e.g., 2, 4, 8, 16), maintaining a constant output stride (e.g., 16) while enlarging the receptive field for dense prediction [5].

2.3.5 Generative Adversarial Networks (GANs)

GANs, introduced by Goodfellow et al. [41], consist of two neural networks: the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$. The generator $\mathcal{G} : \mathbb{R}^k \to \mathbb{R}^n$ maps a latent vector $\boldsymbol{z} \sim p_z(\boldsymbol{z})$ from a prior distribution p_z to a synthetic sample $\boldsymbol{x} \sim p_g(\boldsymbol{x})$, while the discriminator $\mathcal{D} : \mathbb{R}^n \to [0, 1]$ aims to distinguish between real samples $\boldsymbol{x} \sim p_r(\boldsymbol{x})$ and generated ones.

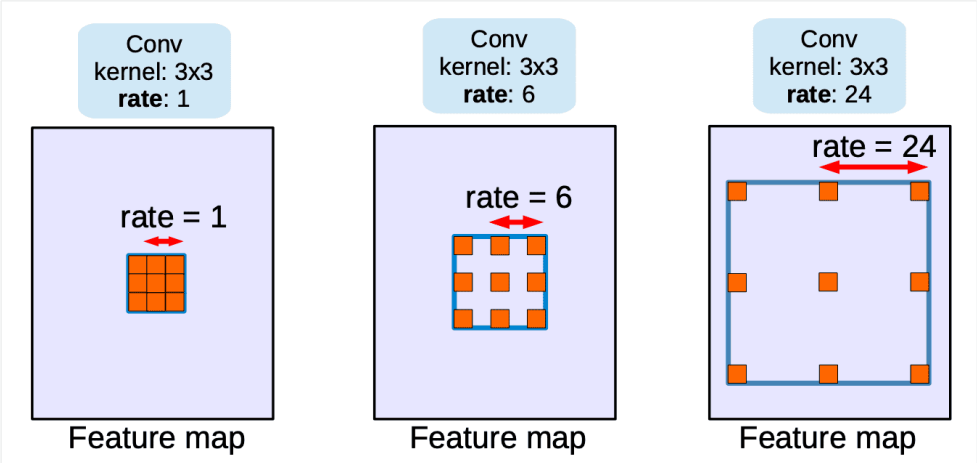


Figure 2.7: Effect of the dilation rate in atrous convolution. With a fixed 3×3 kernel, increasing the dilation rate (e.g., 1, 6, 24) enlarges the spacing between sampled pixels, thereby expanding the receptive field while preserving the feature-map resolution without additional downsampling [5]

.

The standard GAN training objective is formulated as a minimax game:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{L}_{\text{GAN}}(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{\boldsymbol{x} \sim p_r(\boldsymbol{x})}[\log \mathcal{D}(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z})}[\log(1 - \mathcal{D}(\mathcal{G}(\boldsymbol{z})))], \quad (2.25)$$

where $p_r(\boldsymbol{x})$ denotes the real data distribution and $p_g(\boldsymbol{x})$ the generator distribution.

The generator typically uses upsampling layers (e.g., transposed convolutions) to transform $\boldsymbol{z} \in \mathbb{R}^k$ into an image-like output $\mathcal{G}(\boldsymbol{z}) \in \mathbb{R}^{H \times W \times C}$. The discriminator uses convolutional and dense layers to estimate the probability that its input is real.

Conditional GANs (cGANs) [51] extend this framework by incorporating auxiliary information y, such as class labels or images. This means the generator can be "told" what type of image to produce, and the discriminator checks both the image and the condition. The loss becomes:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{L}_{\text{cGAN}}(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{\boldsymbol{x} \sim p_r(\boldsymbol{x}|y)}[\log \mathcal{D}(\boldsymbol{x}, y)] + \mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z})}[\log(1 - \mathcal{D}(\mathcal{G}(\boldsymbol{z}, y), y))]. \quad (2.26)$$

For example, a cGAN trained on satellite imagery could generate a forest scene when conditioned on the label "forest" or an urban scene when conditioned on "city."

CycleGAN [3] tackles unpaired image-to-image translation by enforcing cycle-consistency: if an image is translated from domain A to B and back to A, it

should return to its original form. This is achieved using two generators and two discriminators. For instance, CycleGAN can translate between optical and radar images without requiring perfectly aligned pairs of images.

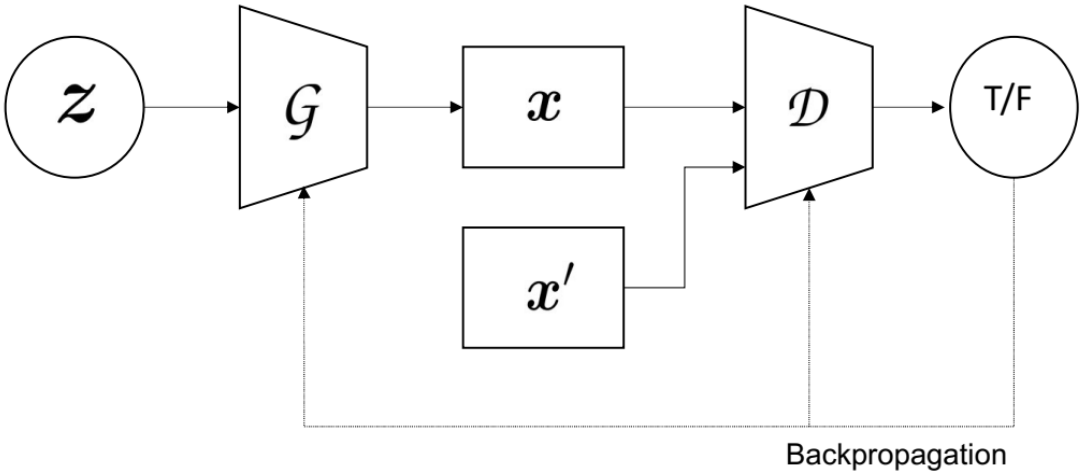


Figure 2.8: Basic GAN schematic. The generator $\mathcal{G}$ maps a latent vector $\mathbf{z} \sim p_z(\mathbf{z})$ to a fake sample $\mathbf{x}' = \mathcal{G}(\mathbf{z})$. The discriminator $\mathcal{D}$ receives real data $\mathbf{x} \sim p_r(\mathbf{x})$ and fake data $\mathbf{x}'$ and outputs $\mathcal{D}(\cdot) \in [0, 1]$. Gradients update $\mathcal{D}$ directly and update $\mathcal{G}$ by backpropagating through $\mathcal{D}$ to increase $\mathcal{D}(\mathcal{G}(\mathbf{z}))$ [41].

2.3.6 Attention-Based Architectures

Transformers, introduced by Vaswani et al. [52], established attention-based architectures as a core paradigm in machine learning. The central operation of the Transformer is the attention mechanism, which computes a weighted representation of values based on the similarity between queries and keys:

$$\text{Attention}(\mathsf{Q}, \mathsf{K}, \mathsf{V}) = \text{softmax}\left(\frac{\mathsf{Q}\mathsf{K}^\mathsf{T}}{\sqrt{d_k}}\right)\mathsf{V}, \tag{2.27}$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors. Intuitively: - A query asks the question "What am I looking for?" - A key describes the content of each element ("what do I contain?"). - A value is the actual information carried by each element. The attention mechanism compares queries with keys to decide how much weight each element should give to all the others, and then combines the corresponding values.

This mechanism enables each input element to attend to all others, allowing the model to capture long-range dependencies and contextual relationships in data. Initially developed for natural language processing, Transformers have since been successfully adapted to computer vision tasks [52, 8].

A Transformer encoder consists of stacked layers, each combining multi-head self-attention with a feed-forward network, residual connections, and layer normalization. Multi-head attention enables the model to focus on different types of relationships at the same time:

$$\text{MultiHead}(\mathsf{Q}, \mathsf{K}, \mathsf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathsf{W}^O \tag{2.28}$$

with

$$\text{head}_i = \text{Attention}(\mathsf{Q}\mathsf{W}_i^Q, \mathsf{K}\mathsf{W}_i^K, \mathsf{V}\mathsf{W}_i^V),$$

where $\mathsf{W}_i^Q, \mathsf{W}_i^K, \mathsf{W}_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are learned projection matrices and $\mathsf{W}^O \in \mathbb{R}^{(h \cdot d_k) \times d_{\text{model}}}$ is the output projection. Each head looks at the input in a slightly different way, like multiple experts focusing on different aspects of the same problem.

Vision Transformers (ViTs) [8] adapt this architecture to image data by dividing an image of size $H \times W \times C$ into $N = \frac{HW}{P^2}$ non-overlapping patches of size $P \times P$. Each patch is flattened and linearly projected:

$$\boldsymbol{z}_p^i = \text{Flatten}(\mathsf{x}[i])\mathsf{E} \in \mathbb{R}^D \tag{2.29}$$

where $\mathsf{x}[i] \in \mathbb{R}^{P^2 \cdot C}$ is the vectorized i-th patch and $\mathsf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is a learnable embedding matrix.

A learnable classification token $\boldsymbol{x}_{\text{class}} \in \mathbb{R}^D$ is prepended to the sequence. To encode spatial positions, a learnable positional embedding $\mathsf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ is added:

$$\mathsf{z}_0 = [\boldsymbol{x}_{\text{class}}; \boldsymbol{z}_p^1; \ldots; \boldsymbol{z}_p^N] + \mathsf{E}_{\text{pos}} \tag{2.30}$$

This ensures that the model knows the relative location of each patch, since self-attention alone does not encode order. The sequence z_0 is passed through the Transformer encoder layers, producing contextualized representations for all tokens. The output embedding $\boldsymbol{z}_{\text{class}}$ corresponds to the final hidden state of the classification token $\boldsymbol{x}_{\text{class}}$, which aggregates information from all image patches through self-attention. For binary classification, this embedding is used to predict the output as

$$\hat{y} = \sigma(\boldsymbol{w}^\mathsf{T} \boldsymbol{z}_{\text{class}} + b), \tag{2.31}$$

where $\boldsymbol{w} \in \mathbb{R}^D$ and $b \in \mathbb{R}$ are learnable parameters, and σ denotes the sigmoid activation function.

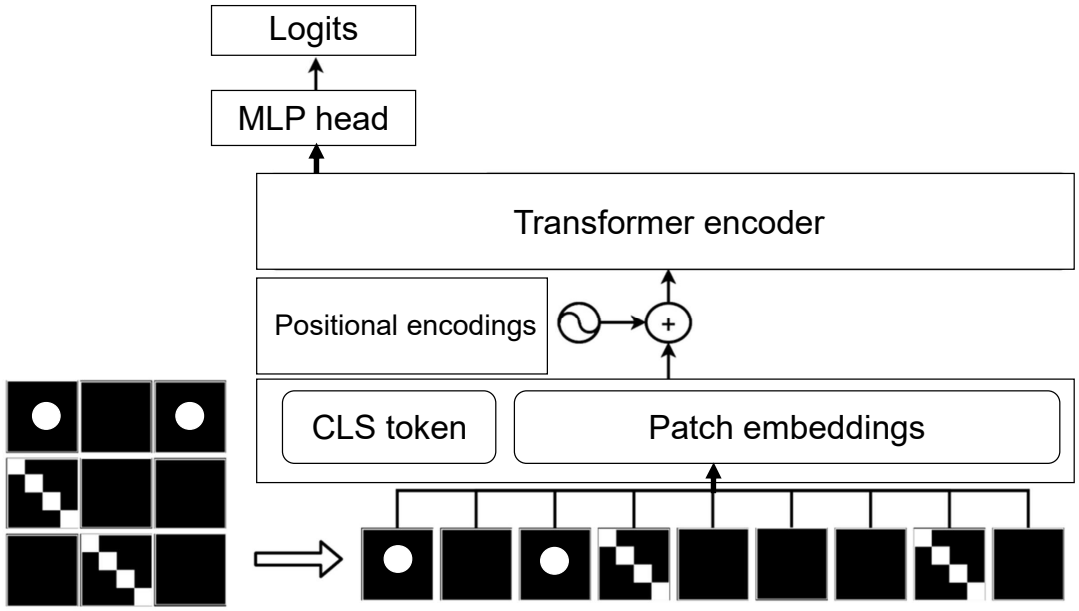


Figure 2.9: Vision Transformer (ViT) with self-attention notation. An image $\mathsf{X} \in \mathbb{R}^{H \times W \times C}$ is split into $N = \frac{HW}{P^2}$ patches; each vectorized patch $\mathsf{x}[i] \in \mathbb{R}^{P^2 C}$ is linearly embedded $\boldsymbol{z}_p^i = \mathsf{x}[i]\mathsf{E} \in \mathbb{R}^D$. A learnable token $\boldsymbol{x}_{\text{class}}$ is prepended and positional embeddings E_{pos} are added to form $\mathsf{z}_0 = [\boldsymbol{x}_{\text{class}}; \boldsymbol{z}_p^1; \ldots; \boldsymbol{z}_p^N] + \mathsf{E}_{\text{pos}}$. The sequence passes through encoder layers using self-attention $\text{Attention}(\mathsf{Q}, \mathsf{K}, \mathsf{V}) = \text{softmax}\big(\mathsf{Q}\mathsf{K}^\mathsf{T}/\sqrt{d_k}\big)\,\mathsf{V}$ (applied in multi-head form). The final class token representation $\boldsymbol{z}_{\text{class}}$ is fed to an MLP head to produce $\hat{y}$ [8].

2.4 Explainable Machine Learning

Explainable Machine Learning (XAI) aims to make the behavior of complex models transparent, interpretable, and trustworthy for human users. As machine learning models are increasingly deployed in high-stakes applications, such as medicine, finance, and environmental monitoring, understanding how a model arrives at a particular decision becomes essential [24]. This chapter introduces foundational concepts in explainability, clarifies the distinction between interpretability and explainability, and provides a taxonomy of explanation techniques. It further discusses intrinsic and post hoc methods, including visualization tools, attribution strategies, and perturbation-based approaches, along with criteria for evaluating their explanations.

2.4.1 Interpretability vs. Explainability

In the field of XAI, the distinction between interpretability and explainability is widely recognized [53, 54, 55]. Interpretability refers to the degree to which a human can consistently understand the model's decision-making mechanism given a certain input [54]. This property is typically associated with inherently transparent models, such as linear regression, decision trees, or rule-based systems, where the influence of each input feature on the prediction can be directly traced and understood.

On the other hand, explainability refers to the ability to describe the internal mechanisms or output of a model in a way that is comprehensible to humans, especially when the model is not inherently interpretable, such as deep neural networks [55, 56]. It is often achieved post hoc by integrating interpretive techniques, such as feature attributions or relevance maps, with domain-specific knowledge [57].

This distinction implies that interpretability concerns the model's structure and its transparency in representing input-output relationships, while explainability focuses on how well the model's decisions can be communicated and understood. For example, a model may be simple in structure but still fail to provide meaningful explanations if its components do not correspond to concepts familiar to users. Conversely, a complex model can be made explainable through techniques that identify which input regions influenced a prediction or how changes in the input affect the output, especially when these insights are linked to domain-relevant concepts that make the model's behavior understandable to users [24, 57].

2.4.2 Taxonomy of Explanation Methods in XAI

Explanatory methods in XAI can be categorized along several key dimensions, reflecting when the explanation is generated, what aspect of the model it aims to describe, and how the explanation is produced. Table 2.1 summarizes these dimensions, stage, scope, and mechanism, and highlights representative techniques within each category.

2.4.3 Inherently Interpretable Models

Inherently interpretable models are characterized by transparent structures that make it straightforward to understand how input features influence predictions. The mathematical details of linear and logistic regression were explained earlier in 2.2. The emphasis here is on their interpretability. In these models, the learned parameters can be directly inspected, with their magnitude and sign indicating the importance and direction of each feature in the decision process [40]. This accessibility contrasts with complex models such as convolutional neural networks, where feature importance cannot be read off directly from the model weights and requires post-hoc explanation methods. While less flexible than deep learning approaches, inherently interpretable models remain valuable in applications where transparency and ease of interpretation are primary requirements [20].

2.4.4 Model Visualization Techniques

Model visualization techniques aim to provide insight into the internal representations and decision-making processes of neural networks. By examining intermediate activations or optimizing inputs to elicit specific responses, these approaches help reveal what information a model has learned and how it processes data across layers. Such techniques contribute to the broader goals of explainability and interpretability in deep learning.

2.4.4.1 Feature Maps

Convolutional Neural Networks (CNNs) learn hierarchical representations of input data through a series of convolutional layers. Each layer produces a set of feature maps, which are 2D arrays capturing the activation of learned filters across spatial regions of the input. Early layers extract low-level features such as edges and textures, while deeper layers encode more abstract semantic patterns like object parts or high-level concepts [64, 66].

Mathematically, the feature map $\mathsf{F}_k^{(l)}$ at layer l and channel k is computed as:

$$\mathsf{F}_k^{(l)} = \sigma\left(\sum_{c=1}^{C} \mathsf{W}_{k,c}^{(l)} * \mathsf{F}_c^{(l-1)} + b_k^{(l)}\right) \tag{2.32}$$

Dimension	Category	Description
Stage	Post-hoc methods	Analyze a trained model to provide insights without modifying its internal structure. Examples include LIME and SHAP [58, 59].
	Intrinsic methods	Encourage interpretability during training using inherently transparent models or architectural constraints (e.g., linear regression, tree-based methods) [60].
Scope	Global explanations	Describe the model's behavior across the entire input space, summarizing general decision logic (e.g., prototype methods) [61].
	Local explanations	Focus on individual predictions by attributing relevance to specific input features (e.g., GradCAM [59, 58].
Mechanism	Feature attribution methods	Assign importance scores to input features based on backpropagated gradients or relevance scores. Examples include Grad-CAM, Integrated Gradients, and DeepLIFT [34, 62, 63].
	Perturbation-based methods	Measure output changes in response to occluding or modifying parts of the input.(e.g., Occlusion and LIME [64, 58].
	Model-based visualization	Reveal internal representations by generating inputs that maximally activate neurons (e.g., Activation Maximization) [65].
	Surrogate models	Use interpretable models (e.g., decision trees) to approximate the behavior of a complex model locally or globally [58, 4].

Table 2.1: Classification of Explanation Methods in XAI

where $\mathsf{F}_c^{(l-1)}$ is the c-th input feature map, $\mathsf{W}_{k,c}^{(l)}$ is the convolution kernel, $b_k^{(l)}$ is the bias term, and $\sigma(\cdot)$ is a non-linear activation function such as ReLU.

These feature maps reflect spatially distributed patterns captured by the network and are central to understanding how CNNs make predictions.

2.4.4.2 Activation Maximization

Activation maximization visualizes what a neural network has learned by synthesizing an input matrix $\mathsf{X} \in \mathbb{R}^{H \times W}$ that maximally activates a particular output neuron [67]. Let $f_c(\mathsf{X})$ denote the logit or activation for class c. The objective is to find an input $\hat{\mathsf{X}}$ that maximizes this activation while incorporating a regularization term $\mathcal{R}(\mathsf{X})$ to improve interpretability:

$$\hat{\mathsf{X}} = \arg\max_{\mathsf{X}} f_c(\mathsf{X}) - \lambda \mathcal{R}(\mathsf{X}) \tag{2.33}$$

where $\lambda \in \mathbb{R}^+$ balances the influence of regularization. The input is iteratively updated using gradient ascent:

$$\mathsf{X}^{(n+1)} = \mathsf{X}^{(n)} + \alpha \nabla_{\mathsf{X}} \left(f_c(\mathsf{X}^{(n)}) - \lambda \mathcal{R}(\mathsf{X}^{(n)}) \right) \tag{2.34}$$

where α denotes the learning rate.

To avoid unrealistic patterns, regularization plays a crucial role. A common choice is the L2 norm $\|\mathsf{X}\|_2^2$, which penalizes large pixel values and prevents excessive activations [68]. Another strategy is total variation loss, which enforces spatial smoothness:

$$\mathcal{R}(\mathsf{X}) = \sum_{u,v} (\mathsf{X}_{u,v} - \mathsf{X}_{u+1,v})^2 + (\mathsf{X}_{u,v} - \mathsf{X}_{u,v+1})^2 \tag{2.35}$$

where u, v denote the spatial coordinates of pixels in the image. Other regularization terms include frequency-domain penalties to suppress high-frequency artifacts, as well as sparsity constraints or pixel value clipping.

To improve the robustness and structure of the synthesized inputs, techniques such as momentum, octave-based optimization, and jittering can be applied. The input initialization $\mathsf{X}^{(0)}$ is typically chosen as random noise or a sample from the training distribution [68].

A more recent approach replaces direct optimization in image space with latent-space optimization using a pretrained generator $\mathcal{G}$. In this case, the latent code $\boldsymbol{z}$ is optimized to produce an image $\mathcal{G}(\boldsymbol{z})$ that maximally activates the target output:

$$\hat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} f_c(\mathcal{G}(\boldsymbol{z})) - \lambda \mathcal{R}(\mathcal{G}(\boldsymbol{z})) \tag{2.36}$$

This formulation, known as Deep Generator Networks for Activation Maximization (DGN-AM) [31, 2], constrains the optimization to the learned data manifold, resulting in more coherent and interpretable visualizations.

The synthesized inputs highlight the features that the model considers most discriminative for a given class. In addition to revealing learned patterns, they can also expose potential biases in the training data, such as the presence of spurious background correlations.

2.4.5 Feature Attribution Methods

Feature attribution methods aim to explain model predictions by assigning importance scores to input features. These methods provide insight into which parts of the input contributed most to the model's decision, making them particularly useful for evaluating complex neural networks. Common techniques include backpropagation-based approaches and gradient-weighted visualizations, each offering different perspectives on model behavior [69].

2.4.5.1 Deep Learning Important FeaTures (DeepLIFT)

DeepLIFT [63] assigns relevance scores to each input feature by comparing the activation of the model output on the input $\mathsf{X} \in \mathbb{R}^{H \times W}$ to that on a reference input X^{ref}. Let $\mathcal{T}(\mathsf{X})$ denote the target output (e.g., a class logit). DeepLIFT computes contributions by analyzing finite differences between X and X^{ref}.

$$\Delta\mathsf{X}_{u,v} = \mathsf{X}_{u,v} - \mathsf{X}^{\text{ref}}_{u,v}, \quad \Delta\mathcal{T} = \mathcal{T}(\mathsf{X}) - \mathcal{T}(\mathsf{X}^{\text{ref}}) \tag{2.37}$$

The attribution $A_{u,v}$ for each input element is computed such that the sum of all attributions equals the total output change:

$$\sum_{u,v} A_{u,v} = \Delta\mathcal{T} \tag{2.38}$$

Each attribution is obtained by scaling the input difference by a multiplier $m_{u,v}$:

$$A_{u,v} = m_{u,v} \cdot \Delta\mathsf{X}_{u,v} \tag{2.39}$$

For linear operations of the form

$$y = \sum_{u,v} w_{u,v}\mathsf{X}_{u,v} + b, \tag{2.40}$$

the change in output is

$$\Delta y = \sum_{u,v} w_{u,v} \cdot \Delta\mathsf{X}_{u,v}, \tag{2.41}$$

and the attributions become

$$A_{u,v} = w_{u,v} \cdot \Delta \mathsf{X}_{u,v} \tag{2.42}$$

For nonlinear functions such as ReLU, DeepLIFT applies specialized propagation rules, such as the rescale rule, which distributes the output difference proportionally to the input deviations:

$$\Delta y = \sum_{u,v} m_{\mathsf{X}_{u,v} \to y} \cdot \Delta \mathsf{X}_{u,v} \tag{2.43}$$

This approach allows DeepLIFT to remain effective even in regions where gradient-based methods fail due to saturation or flat activations. The attributions are propagated backward from the output through the network using recorded activations and reference values.

The choice of the reference input X^{ref} is critical. It should represent the absence of a relevant signal while remaining within the data distribution. In image tasks, common references include zero-valued (black) images or blurred versions of the input. In other domains, feature-wise means are often used [63].

2.4.5.2 Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM [34] generates class-specific heatmaps that highlight spatial regions relevant for a given prediction. Let $\mathit{y}^c \in \mathbb{R}$ denote the scalar logit corresponding to class c, and let $\mathsf{A}^k \in \mathbb{R}^{H \times W}$ denote the activation map of feature channel k from a chosen convolutional layer. Grad-CAM computes the gradients of y^c with respect to A^k and averages them spatially:

$$\alpha_k^c = \frac{1}{H \cdot W} \sum_{u=1}^{H} \sum_{v=1}^{W} \frac{\partial \mathit{y}^c}{\partial \mathsf{A}_{u,v}^k} \tag{2.44}$$

Here, u and v index the vertical and horizontal positions in the activation map.

The final heatmap for class c is:

$$\mathsf{L}_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c \mathsf{A}^k \right) \tag{2.45}$$

where $\text{ReLU}(x) = \max(0, x)$ ensures that only positively contributing regions are highlighted. The heatmap is then upsampled to match the input resolution and overlaid on the image for interpretation.

Grad-CAM shows which regions of the image the network focused on when making a decision. For example, in classifying an image of a bird, the heatmap might highlight the wings and head rather than the background, making the model's reasoning more understandable.

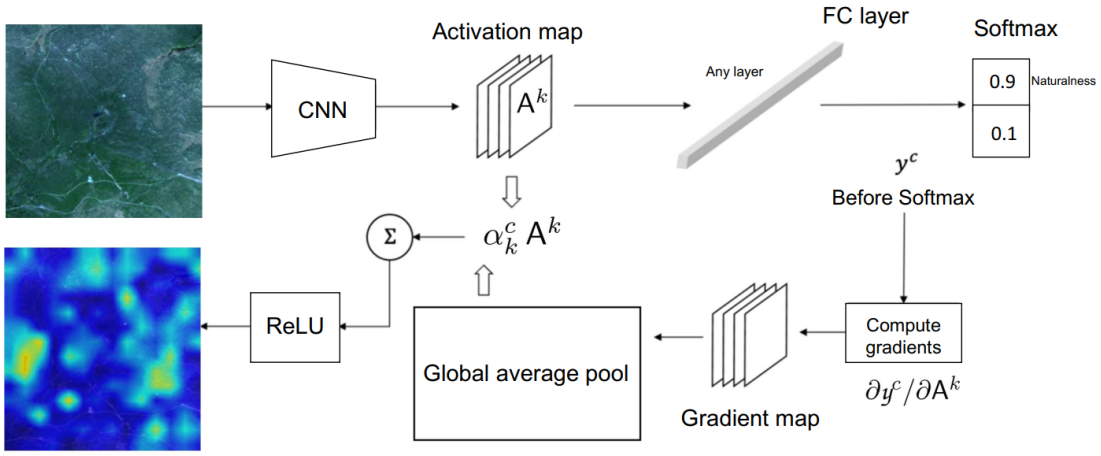


Figure 2.10: Grad-CAM pipeline in our notation. A forward pass provides activation maps A^k at a chosen layer and the class logit y^c. Backpropagation yields gradients $\partial y^c / \partial \mathsf{A}^k$, from which weights are computed as $\alpha_k^c = \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} \partial y^c / \partial \mathsf{A}_{u,v}^k$. The class-specific heatmap is $\mathsf{L}^c_{\text{Grad-CAM}} = \text{ReLU}\big(\sum_k \alpha_k^c \mathsf{A}^k\big)$, then upsampled and overlaid on the input image.

2.4.6 Perturbation-based Methods

Perturbation-based methods interpret model predictions by directly altering the input and observing the resulting changes in output. These methods are model-agnostic, making them applicable to both white-box and black-box models. Examples include occlusion sensitivity [64], Local Interpretable Model-agnostic Explanations (LIME) [58], SHapley Additive exPlanations (SHAP) [59], and deletion-insertion metrics [70].

Occlusion sensitivity, also referred to as sliding window occlusion [64], evaluates the importance of spatial regions by systematically masking parts of a 2D input matrix $\mathsf{X} \in \mathbb{R}^{H \times W}$ and observing the change in the model's output. In image classification, this typically involves sliding a patch (e.g., black, blurred, or filled with noise) across the input and recording the effect on the class score.

Let $f(\mathsf{X})$ denote the model's output for a target class c, and let $\mathsf{X}_{\text{occl}(u,v)}$ denote the input with a patch centered at location (u, v) replaced by a baseline value. The attribution at location (u, v) is then computed as:

$$A(u, v) = f(\mathsf{X}) - f\big(\mathsf{X}_{\text{occl}(u,v)}\big) \tag{2.46}$$

A large value of $A(u, v)$ implies that the region around (u, v) is important for the model's decision. Negative values indicate that masking the region may improve the prediction, suggesting a potentially suppressive influence.

The resolution of the resulting attribution map depends on the patch size and stride. Larger patches capture broader context but reduce spatial precision, while smaller patches improve localization at the cost of increased computation. The number of forward passes for an input of size $H \times W$, patch size $p \times p$, and stride

s is approximately:

$$\left\lceil \frac{H}{s} \right\rceil \cdot \left\lceil \frac{W}{s} \right\rceil \tag{2.47}$$

For example, applying a 15×15 patch with a stride 15 on a 224×224 image results in around 225 forward passes. Reducing the stride increases the fidelity of the heatmap but also the computational burden. Compared to gradient-based methods such as Grad-CAM, occlusion sensitivity is significantly more expensive since it requires many forward evaluations.

The choice of baseline affects the interpretability of the results. Zero-valued patches are simple but may introduce unnatural artifacts, whereas blurred or mean-value patches tend to preserve global image structure.

The final scores $A(u, v)$ are often visualized as heatmaps. Negative values may be clipped to focus on regions that support the prediction. For instance, if occluding the head of a cat leads to a large drop in the class score, that region is considered crucial for the classification.

Despite its computational cost, occlusion sensitivity remains a robust and model-agnostic approach, especially useful when internal model details or gradients are inaccessible.

2.4.7 Surrogate Models and Approximators

Surrogate models provide post hoc explanations by approximating the behavior of complex models with simpler, interpretable ones. They enable users to understand a black-box model's decision process without requiring access to internal parameters or gradients [71].

Global surrogate models aim to approximate the overall decision function of a black-box model across the full input space. This involves generating synthetic input-output pairs $\{\mathsf{X}^{(i)}, f(\mathsf{X}^{(i)})\}$, where $\mathsf{X}^{(i)} \in \mathbb{R}^{H \times W}$ is a 2D input matrix and f is the original model. An interpretable model $g(\mathsf{X})$, such as a linear model or shallow decision tree, is trained to mimic f. The surrogate's performance is evaluated using metrics such as the coefficient of determination (R^2) or agreement rate. While global surrogates offer high-level insights, they often fail to capture complex nonlinear dependencies unless their own complexity is increased—potentially compromising interpretability.

Local surrogate models explain individual predictions by approximating the model's behavior in the neighborhood of a specific input $\mathsf{X}^{(0)}$. A widely used method is LIME (Local Interpretable Model-agnostic Explanations) [58], which generates perturbed versions $\mathsf{Z}^{(i)} \in \mathcal{Z}$ around $\mathsf{X}^{(0)}$, queries the black-box model f, and fits a simple model g to approximate the predictions locally.

The perturbed inputs are weighted based on their similarity to $\mathsf{X}^{(0)}$ using a kernel function:

$$\pi_{\mathsf{X}^{(0)}}(\mathsf{Z}^{(i)}) = \exp\left(-\frac{D(\mathsf{X}^{(0)}, \mathsf{Z}^{(i)})^2}{\sigma^2}\right) \tag{2.48}$$

where $D(\mathsf{X}^{(0)}, \mathsf{Z}^{(i)})$ is a distance function (e.g., Euclidean), and σ controls the scale of locality.

The surrogate g is then obtained by minimizing a weighted and regularized loss:

$$\xi(\mathsf{X}^{(0)}) = \arg\min_{g \in \mathcal{G}} \sum_{\mathsf{Z}^{(i)} \in \mathcal{Z}} \pi_{\mathsf{X}^{(0)}}(\mathsf{Z}^{(i)})\, \mathcal{L}(f(\mathsf{Z}^{(i)}), g(\mathsf{Z}^{(i)})) + \Omega(g), \tag{2.49}$$

where $\mathcal{G}$ denotes the space of interpretable models, $\mathcal{L}$ is a loss function (e.g., squared error), and $\Omega(g)$ is a complexity penalty to promote simplicity.

Local surrogates yield explanations specific to a given input. Their fidelity depends on the expressiveness of g and the kernel bandwidth σ, which determines the trade-off between locality and robustness.

2.4.8 Attribution in Attention-Based Models

In Transformer-based architectures, attention weights can be utilized to interpret how information is propagated across tokens. Each attention head produces an attention matrix that quantifies the relative contribution of one token to another, which can be visualized as a heatmap to analyze information flow [8].

$$\mathsf{A} = \text{softmax}\left(\frac{\mathsf{Q}\mathsf{K}^\mathsf{T}}{\sqrt{d_h}}\right), \tag{2.50}$$

where Q and K denote the query and key matrices, and d_h is the dimensionality of the attention head. The entry A represents the attention weight from the whole sequence. The complete attention matrix $\mathsf{A} \in \mathbb{R}^{T \times T}$, with T denoting the number of tokens, can be visualized to reveal how attention is distributed across the sequence.

However, standard attention visualizations have notable limitations. They are not class-specific and do not indicate which attention pathways are causally relevant to the model's output. Furthermore, attention rollout methods—based on recursively multiplying averaged attention maps across layers—ignore gradient information and may yield misleading attributions.

To address these shortcomings, LRP Attention Rollout [9] integrates gradient-based relevance propagation into the attention computation. It computes the element-wise product between attention scores and their gradients, yielding relevance-aware maps specific to the target prediction.

The method adapts Layer-wise Relevance Propagation (LRP) [72] for Transformer architectures, addressing challenges such as skip connections and

non-ReLU activations that can disrupt relevance conservation. This ensures numerically stable and faithful propagation.

Relevance propagation for a target class t is defined as:

$$R_j^{(n)} = \sum_i \frac{\mathsf{X}_j \, \partial L_i^{(n)}(\mathsf{X}, \mathsf{w}) / \partial \mathsf{X}_j}{L_i^{(n)}(\mathsf{X}, \mathsf{w})} \, R_i^{(n-1)}, \tag{2.51}$$

where i and j index neurons in consecutive layers, with i referring to neurons in layer $n-1$ and j to neurons in layer n. The relevance $R_j^{(n)}$ is thus redistributed from neurons i in the previous layer to neurons j in the current one.

To preserve positivity and stability, only positive contributions are propagated:

$$R_j^{(n)} = \sum_{i:(i,j)\in q} \frac{\mathsf{X}_j w_{ji}}{\sum_{j':(j',i)\in q} \mathsf{X}_{j'} w_{j'i}} R_i^{(n-1)}, \tag{2.52}$$

where $q = \{(i,j) \mid \mathsf{X}_j w_{ji} \geq 0\}$. The index j' refers to neurons connected to neuron i within the positive relevance set q, and the prime notation is used to distinguish inner summation indices from the outer ones.

For operations involving skip connections or matrix compositions, relevance is normalized to preserve total relevance:

$$\bar{R}_u = \frac{R_u}{|R_u| + |R_v|} \cdot \sum_i R_i^{(n-1)}, \quad \bar{R}_v = \frac{R_v}{|R_u| + |R_v|} \cdot \sum_i R_i^{(n-1)}, \tag{2.53}$$

ensuring the conservation of relevance:

$$\sum_j \bar{R}_u + \sum_k \bar{R}_v = \sum_i R_i^{(n-1)}. \tag{2.54}$$

In each attention block b, the raw attention map is computed as:

$$\mathsf{A}^{(b)} = \text{softmax}\left(\frac{\mathsf{Q}^{(b)}\mathsf{K}^{(b)\mathsf{T}}}{\sqrt{d_h}}\right), \tag{2.55}$$

and the relevance-aware attention map is:

$$\bar{\mathsf{A}}^{(b)} = \mathsf{I} + \mathbb{E}_h\left(\nabla \mathsf{A}^{(b)} \odot \mathsf{R}^{(n_b)}\right), \tag{2.56}$$

where $\mathbb{E}_h$ denotes the average over attention heads, and $\odot$ is the element-wise product.

The final relevance matrix is obtained by chaining relevance-aware attention maps across layers:

$$\mathsf{C} = \bar{\mathsf{A}}^{(1)} \cdot \bar{\mathsf{A}}^{(2)} \cdot \ldots \cdot \bar{\mathsf{A}}^{(B)}. \tag{2.57}$$

The explanation for the input X is derived from the [CLS] token row of C, reshaped and upsampled into a spatial heatmap.

In contrast, the classical rollout method multiplies head-averaged attention matrices without using gradients:

$$\hat{\mathsf{A}}^{(b)} = \mathsf{I} + \mathbb{E}_h \mathsf{A}^{(b)}, \quad \text{rollout} = \hat{\mathsf{A}}^{(1)} \cdot \ldots \cdot \hat{\mathsf{A}}^{(B)}. \tag{2.58}$$

This approach is class-agnostic and does not reflect causal influence on the prediction.

2.5 Evaluation Criteria for Explanations

Explanations generated by XAI methods are commonly evaluated based on three essential properties: validity, completeness, and objectivity [73, 20, 74, 24, 4]. These criteria assess the quality of the explanation itself rather than the underlying explanation method.

Validity refers to whether an explanation faithfully reflects the expected internal reasoning process of the model. A valid explanation should contain both the entity being explained and the content that contributed to the model's output. For instance, in image classification, a valid explanation for a "cat" prediction should highlight the cat itself—such as its fur texture or facial features—rather than unrelated background areas like grass or furniture. Explanations that attribute relevance to irrelevant regions or features are considered invalid, as they misrepresent the model's true decision basis. If the explanations consistently fail to highlight the relevant parts of the input, this indicates that the model's prediction is not trustworthy and that the explanation produced by the XAI method lacks validity. Validity is essential for legal accountability and trustworthy artificial intelligence [25, 4].

Completeness evaluates whether the explanation includes all important features or regions responsible for the model's decision. For example, in a cat classification task, a complete explanation should highlight all regions that contribute to recognizing the cat—such as the head, body, and fur patterns—rather than focusing on a single part like the ears and ignoring the other features [73, 20, 4].

Objectivity concerns the possibility of evaluating explanations quantitatively without imposing human bias. This enables reproducible and fair comparisons of different explanation methods across models and datasets [4, 28].

Together, these properties form a foundational criterion for assessing the explanations produced by XAI methods.

2.6 Uncertainty Quantification (UQ)

Understanding a model's confidence in its predictions is essential, particularly in high-stakes applications such as healthcare or remote sensing. UQ provides mechanisms to assess this confidence, offering insight into when a model's outputs can be trusted. This section outlines common techniques for modeling and evaluating uncertainty [37, 6, 28].

2.6.1 Confidence Intervals as a Measure of Uncertainty

A classical approach to uncertainty quantification involves constructing confidence intervals, especially in linear regression models. These intervals provide a range around an estimate within which the true value is expected to lie with a specified level of confidence [75]. Consider a simple linear regression model that predicts a continuous target y from an input vector $\boldsymbol{x} \in \mathbb{R}^d$:

$$\hat{y} = \boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b \tag{2.59}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ are the regression weights and $b \in \mathbb{R}$ is the bias term. The model assumes a linear relationship between inputs and outputs with additive Gaussian noise.

After fitting the model to data, we obtain estimates $\hat{\boldsymbol{w}}$ and $\hat{b}$. The standard error of an estimated coefficient $\hat{w}_j$ reflects its uncertainty. A $(1-\alpha)\%$ confidence interval for $\hat{w}_j$ is given by:

$$\hat{w}_j \pm z_{\alpha/2} \cdot \mathrm{SE}(\hat{w}_j) \tag{2.60}$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution and $\mathrm{SE}(\hat{w}_j)$ is the standard error of the coefficient, and α denotes the significance level, where for example $\alpha = 0.05$ corresponds to a 95% confidence interval.

Similarly, a confidence interval for a model prediction $\hat{y}$ at a new input $\boldsymbol{x}^*$ accounts for both the variance of the model and the residual variance:

$$\hat{y}(\boldsymbol{x}^*) \pm z_{\alpha/2} \cdot \mathrm{SE}(\hat{y}(\boldsymbol{x}^*)) \tag{2.61}$$

Here, $\mathrm{SE}(\hat{y}(\boldsymbol{x}^*))$ reflects the uncertainty of the predicted output due to both parameter estimation error and inherent data noise.

This method provides efficient and interpretable uncertainty estimates but relies on assumptions such as linearity, homoscedasticity, and normally distributed errors. They also do not capture uncertainty in the model structure itself, which limits their applicability in complex or nonlinear models such as neural networks.

2.6.2 Bayesian Neural Networks (BNNs)

BNNs [76] quantify uncertainty by modeling the weights w of a neural network as probability distributions rather than fixed values. The weights $\mathsf{w} = \{w_1, w_2, \ldots, w_P\}$ represent all trainable parameters of the model, where P is the total number of parameters. Given a dataset $\mathsf{D} = \{(\mathsf{X}^{(i)}, y^{(i)})\}_{i=1}^{N}$, the posterior distribution over weights is derived via Bayes' theorem:

$$p(\mathsf{w} \mid \mathsf{D}) = \frac{p(\mathsf{D} \mid \mathsf{w})\, p(\mathsf{w})}{p(\mathsf{D})} \tag{2.62}$$

The predictive distribution for a new input X^* is obtained by marginalizing over the weight posterior:

$$p(y^* \mid \mathsf{X}^*, \mathsf{D}) = \int p(y^* \mid \mathsf{X}^*, \mathsf{w})\, p(\mathsf{w} \mid \mathsf{D})\, d\mathsf{w} \tag{2.63}$$

Since the exact posterior is intractable in deep networks, variational inference is used to approximate it with a simpler distribution $q(\mathsf{w})$. Sampling from this approximation yields:

$$p(y^* \mid \mathsf{X}^*, \mathsf{D}) \approx \frac{1}{T} \sum_{t=1}^{T} p(y^* \mid \mathsf{X}^*, \mathsf{w}^{(t)}), \quad \mathsf{w}^{(t)} \sim q(\mathsf{w}) \tag{2.64}$$

By integrating over weight uncertainty, BNNs provide a principled way to model epistemic uncertainty. However, they are computationally intensive and more challenging to train than standard neural networks [76, 6, 37].

2.6.3 Monte Carlo Dropout

Due to the computational limitations of Bayesian Neural Networks, Monte Carlo (MC) Dropout [6] offers a practical approximation for uncertainty estimation with minimal architectural changes. Dropout is a regularization technique that randomly deactivates a subset of neurons during training to prevent overfitting. By keeping dropout active during inference, the model introduces stochasticity in predictions, enabling approximate Bayesian inference [6, 4, 7]. MC Dropout provides estimates of epistemic uncertainty while remaining computationally more efficient than full Bayesian inference, although it increases inference time due to multiple stochastic forward passes [76].

Given an input X, multiple stochastic forward passes are performed:

$$\{\hat{y}^{(t)}\}_{t=1}^{T}, \quad \hat{y}^{(t)} = f_{\text{drop}}^{(t)}(\mathsf{X}) \tag{2.65}$$

The predictive mean is:

$$\mathbb{E}[\hat{y}] \approx \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)} \tag{2.66}$$

The predictive variance, capturing model uncertainty, is:

$$\mathrm{Var}[\hat{y}] \approx \frac{1}{T} \sum_{t=1}^{T} \left(\hat{y}^{(t)} - \mathbb{E}[\hat{y}]\right)^2 \tag{2.67}$$

Here, T denotes the number of stochastic forward passes, $\hat{y}^{(t)}$ represents the prediction from the t-th pass of the network with dropout active, $f_{\text{drop}}^{(t)}$ is the model instance under a specific dropout mask.

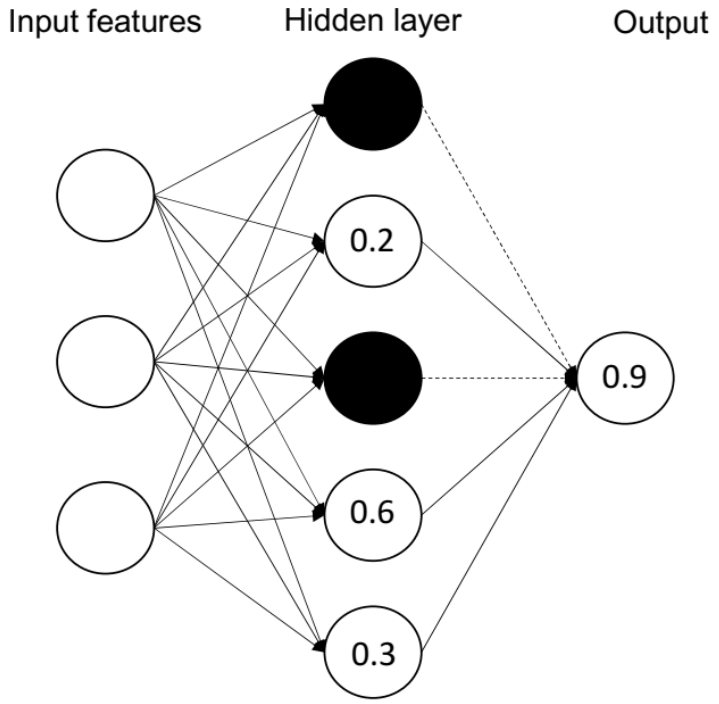


Figure 2.11: Monte Carlo Dropout (MC-Dropout). Dropout remains active at inference, randomly closing a subset of hidden units (filled black; dashed connections ignored). For an input X, run T stochastic forward passes $\{\hat{y}^{(t)}\}_{t=1}^{T}$ with $\hat{y}^{(t)} = f_{\text{drop}}^{(t)}(\mathsf{X})$. The predictive mean and variance are estimated by $\mathbb{E}[\hat{y}] \approx \frac{1}{T}\sum_{t=1}^{T} \hat{y}^{(t)}$ and $\mathrm{Var}[\hat{y}] \approx \frac{1}{T}\sum_{t=1}^{T} \left(\hat{y}^{(t)} - \mathbb{E}[\hat{y}]\right)^2$, providing an approximation to epistemic uncertainty.

2.6.4 Calibration of Uncertainty Estimates

Uncertainty estimates are meaningful only if they are well-calibrated, which means that the predicted confidence aligns with actual correctness. A model is calibrated if, for instance, predictions made with 80% confidence are correct 80% of the time.

A common metric for calibration is the Expected Calibration Error (ECE) [77], defined as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{n_m}{n} \left|a_m - c_m\right| \tag{2.68}$$

where M is the number of confidence bins, n is the total number of predictions, n_m is the number of predictions in bin m, a_m is the empirical accuracy in that bin, and c_m is the average predicted confidence. The ECE quantifies the average discrepancy between confidence and accuracy, weighted by bin size. A lower ECE indicates better calibration. Poorly calibrated models may be overconfident, leading to unreliable decisions.

Post-hoc calibration methods adjust model outputs to better align predicted confidence with actual correctness. A widely used technique is temperature scaling [77], where a scalar T is applied to the logits z before the softmax function:

$$\text{softmax}\left(\frac{\mathsf{z}}{T}\right) \tag{2.69}$$

The temperature T is optimized on a validation set to minimize calibration error. Other post-hoc methods include Platt scaling, histogram binning, and isotonic regression [78, 79], each offering different strategies to map predicted probabilities to calibrated outputs.

Beyond such post-processing, intrinsic calibration approaches aim to produce well-calibrated uncertainty estimates directly during training. Bayesian neural networks and deep ensembles [76, 80] capture uncertainty over model parameters, often resulting in more reliable confidence estimates, particularly under distribution shifts or limited data. An alternative strategy involves modifying the training objective by incorporating calibration-sensitive loss functions, such as the Brier score or focal loss [81], which penalize overconfident predictions. These approaches enhance the quality of uncertainty estimates without requiring post-hoc adjustments [27, 82].

Chapter 3

Related Work

Explaining and mapping naturalness is essential for monitoring human impact and guiding conservation [12, 1]. Despite methodological differences, almost all approaches share a common principle: they estimate naturalness indirectly through proxies such as population density, land use, infrastructure, or remote sensing signals, or rely on datasets that are themselves constructed based on these assumptions [10, 11, 15]. This chapter compares six approaches investigating naturalness, highlighting their assumptions, inputs, and outputs. Table 3.1 concludes the chapter by comparing their key characteristics.

3.1 Taxonomy of Naturalness

Naturalness is central to conservation science, yet it remains variably defined across ecological and philosophical contexts. Grabherr et al. (1998) introduced a human-impact framework, which classifies ecosystems according to their degree of human modification, equating naturalness inversely with levels of human disturbance [83]. Landres et al. [13] approached the concept from a different perspective, defining naturalness as the degree to which ecological processes and components are self-organizing, minimally influenced by human activities, and retain their native characteristics. Winter et al. (2010) [16] expanded on this by introducing a gradient-based understanding using the Relative Quantitative Reference Approach for Naturalness Assessments (RANA), which estimates naturalness as a continuum from 0% (maximum human impact) to 100% (least human impact), highlighting the impossibility of identifying truly untouched references in modern landscapes due to pervasive human influence.

Although often used interchangeably, wilderness and naturalness represent related but distinct perspectives on the human–nature relationship. The U.S. Wilderness Act of 1964 defines wilderness as an area "untrammeled by man," emphasizing non-intervention and freedom from human control [84]. Cookson

offers a more nuanced definition of wildness as a "quality of interactive processing between an organism and its surroundings," positioning wildness as an emergent, relational property rather than a spatial designation [85]. This diverges from older conceptions that equate wilderness simply with remoteness or a lack of human presence.

Human influence or human impact is quantitatively operationalized through indices like the Human Influence Index (HII) [10], as proposed by Sanderson et al. (2002) and further modernized by Ekim et al These frameworks use proxies such as population density, land transformation, and infrastructure presence to map anthropogenic impact at various spatial resolutions [10, 11].

Ecological integrity, closely related to naturalness, refers to the wholeness and functioning of ecosystems in terms of composition, structure, and processes. Winter et al. (2010) emphasize that integrity is most evident in systems where historical conditions are largely retained or recoverable, though even core protected areas in Europe may lack fully reliable reference ecosystems due to past interventions [16]. Across these terms and studies, the variety of definitions reflects disciplinary perspectives; however, I argue that they ultimately seek to conceptualize the same underlying phenomenon—the differentiation of areas along a spectrum of human impact, approached from different viewpoints.

In this thesis, we adopt the term *naturalness* to describe areas that are strictly protected and show little to no modern human influence. As an objective proxy for naturalness, we focus on preserved natural areas within the Fennoscandian region, where strict protection regimes and long-standing conservation efforts have ensured that these areas remain minimally influenced by human activity [86, 87, 88].

3.2 Heuristic Scoring Approaches to Naturalness Mapping

Before the advent of data-driven models, naturalness was commonly estimated using heuristic scoring systems based on expert knowledge and rule-based fusion of geospatial indicators. These approaches assign impact scores to proxies such as population density, land cover, infrastructure, and accessibility, which are then aggregated into an overall naturalness or human influence index. While conceptually transparent and ecologically motivated, they often rely on static assumptions, outdated data, and uniform weights that may not capture regional variability or ecological nuance [13, 10].

This section reviews three representative heuristic frameworks: the global Human Footprint Index [10], the forest-specific RANA model for Central Europe [16], and the high-resolution Naturalness Index by Ekim et al. [11], which aims to modernize earlier approaches by incorporating recent datasets and finer spatial granularity.

3.2.1 The Human Influence Index (HII)

Sanderson et al. [10] introduced the HII, a global method to quantify human influence on terrestrial ecosystems. The goal was to create a unified index that identifies areas with minimal human disturbance—termed the "Last of the Wild."

HII is based on four main proxies of human impact: (1) population density [89], (2) land transformation through agriculture and built environments, (3) accessibility via roads, rivers, and coastlines [90], and (4) nighttime illumination as a proxy for infrastructure [91, 92]. Each layer was ranked on a scale from 0 to 10, where 10 indicates the highest impact. These scores were summed to form the Human Influence Index (HII), mapped at $1\,\mathrm{km}^2$ resolution.

A strength of the method lies in its integration of ecological context. The scoring was applied within specific biomes and ecoregions, acknowledging that the same ecological pattern may have different HII depending on location. This allowed for identifying naturalness within each biome.

Despite its impact, the method has limitations. It provides a static snapshot and lacks temporal dynamics. Some input datasets were outdated or coarse. Additionally, assigning equal weights to all layers may overlook regional differences in sensitivity to human impact. The scoring mechanism is based on logical assumptions about human influence, but it remains heuristic and may introduce bias, as it reflects expert-defined notions of naturalness rather than deriving from data-driven optimization.

3.2.2 The Naturalness Index (NI)

Ekim et al introduced the NI [11], a high-resolution metric designed to modernize the HII [10] by offering finer spatial granularity and updated input data. While the HII operates at a $1\,\mathrm{km}^2$ scale, the NI improves resolution to 10 meters and adapts the framework to better reflect current anthropogenic pressures, particularly in fragmented or densely populated regions.

The NI is constructed through a rule-based aggregation of four geospatial indicators of human presence: (1) population density [89], (2) land cover transformation [93], (3) nighttime illumination [94], and (4) accessibility. Accessibility is modeled using OpenStreetMap infrastructure and cost-distance surfaces based on terrain elevation data (ALOS and SRTM), with Tobler's hiking function [95, 96] simulating realistic travel effort.

Each layer is normalized to a 0–10 scale, and land cover classes are weighted according to their degree of human modification. Urban and industrial surfaces receive maximum impact scores (10), while forests, wetlands, and other semi-natural areas are assigned intermediate scores depending on their presumed naturalness. Bare ground and snow/ice are excluded from scoring due to classification uncertainty. These weighted layers are summed into a modified Human Influence Index, which is then inverted and rescaled into the final Naturalness Index, ranging from 0 (least natural) to 100 (most natural).

Despite its strengths, the method has several limitations. Some input datasets, such as population and nighttime lights, have coarser native resolution than the output map, which may affect accuracy at fine spatial scales. Importantly, the rule-based scoring relies on expert judgment rather than a data-driven method, which may lead to biased assumptions about land cover naturalness. Nevertheless, the NI offers a scalable and reproducible framework for mapping naturalness at high resolution. In chapter 8, I compare my proposed data-driven approaches with the HII and NI to examine whether their outcomes align conceptually, without treating these indices as ground truth or expecting a perfect correspondence in an optimal scenario

3.2.3 Naturalness Assessment in the Bavarian Forest

Winter et al. introduced the Relative Quantitative Reference Approach for Naturalness Assessments (RANA) [16], a method developed to evaluate the naturalness of forests in Central Europe, where truly pristine ecosystems—defined as areas entirely untouched by human activity—are no longer available [97, 98]. The study was conducted at a local scale in the Bavarian Forest National Park, focusing on forested land only.

The methodology is based on systematic field data from two sources: a dense

200 m by 200 m forest inventory grid covering 4599 plots and an additional 100 stratified plots with expanded ecological variables [16]. Each plot was evaluated based on eight ecological indicators: tree species composition, diameter distribution of living trees, quality and quantity of deadwood [99, 100, 15], presence of epiphytic lichens [101, 102], degree of canopy layering, forest continuity, deviation from potential natural vegetation [103], and fragmentation assessed via effective mesh size [104].

Naturalness scores were assigned on a scale from 0 to 100, with the least-managed zones in the park used as internal references for naturalness instead of idealized, fully untouched forests. These internal benchmarks helped calibrate the scoring, enabling relative assessments across the landscape.

The method offers a way to evaluate naturalness in managed landscapes. However, it provides only a static snapshot, relies on internal reference zones that may still reflect past human influence [105], and assumes equal weight for all indicators without sensitivity analysis. Additionally, the scoring strategy may suffer from bias, as it is based on the authors' interpretation of naturalness, which is modeled using Potential Natural Vegetation (PNV) [16] as a simulated ecological baseline. While regionally effective, the approach would be difficult to apply at a global scale and would require adaptation for use in other forest types or climatic zones.

3.2.4 Multimodal Learning for Naturalness Estimation

To improve the predictive quality of naturalness estimation, Ekim and Schmitt introduced a multimodal approach that integrates multiple sources of geospatial information [106]. Specifically, the model combines spectral features from Sentinel-2 imagery, spatial context derived from a larger surrounding area encoded via a pretrained autoencoder, and geographic coordinates represented through cyclic encoding. These modalities are fused within a U-Net architecture [47] to generate high-resolution, pixel-level naturalness predictions.

The motivation behind this design was to leverage richer contextual and spatial information to produce more accurate and fine-grained naturalness maps than those based on handcrafted proxies alone. By learning directly from image data and spatial cues, the method aims to replace manual indicator fusion with a scalable, data-driven solution.

However, a key limitation lies in the fact that the model is trained on labels derived from the NI [11], which is based on heuristic rules and expert-defined assumptions [106]. As a result, the model inherits the biases present in the NI, which stem from its rule-based formulation, rather than learning ecological patterns independently. This dependency constrains its ability to move beyond the conceptual limitations of its training data.

3.3 XAI-based Frameworks for Naturalness Investigation

As alternatives to heuristic, rule-based scoring methods, recent research explored XAI-based frameworks that estimate naturalness directly in satellite imagery. These approaches explain model behavior and help to better understand how different input features or data sources contribute to naturalness estimation. This shift reflects a growing emphasis on transparency, ecological reasoning, and the need to validate machine learning outputs in high-stakes environmental applications.

This section reviews key contributions in this area, including: (1) the interpretable AnthroProtect model by Stomberg et al. [107] that learns directly from land protection status, and (2) a modality occlusion framework designed to quantify the relative importance of different input modalities in naturalness prediction [108].

3.3.1 Latent Space Occlusions to Assess Naturalness

Stomberg et. al proposed Activation Space Occlusion Sensitivity (ASOS) [12], an interpretable deep learning framework designed to distinguish naturalness from anthropogenically modified landscapes in Sentinel-2 imagery. The primary goal of the work is to leverage the internal knowledge learned by the model to improve our current understanding of what constitutes naturalness, rather than using machine learning purely for prediction. By examining how the model learns to separate natural from human-impacted areas, the authors aim to support ecological reasoning and inform conservation planning.

The model is trained in a weakly supervised manner on 10 m resolution Sentinel-2 multispectral imagery, using global labels constructed from the World Database on Protected Areas (WDPA) [109] and OpenStreetMap (OSM) [110] annotations. WDPA polygons are treated as proxies for naturalness, while OSM infrastructure layers (e.g., roads, buildings) represent anthropogenic modification. The dataset covers diverse ecological regions in Fennoscandia [12].

ASOS quantifies changes in latent activation space when specific input patches are occluded, to map the contribution of input patches to the naturalness of the scene. This approach is driven entirely by data and seeks to reveal new insights from the learned representations. This shift from rule-based scoring to data-driven interpretation reflects a methodological advancement toward explainable, empirically grounded models.

Nonetheless, the approach is not without limitations. Occlusion-based methods such as ASOS may introduce out-of-distribution artifacts when

activations are occluded, potentially distorting interpretability and introducing uncertainties in the model [1, 111].

3.3.2 Modality Occlusions for Estimating Modality Contribution to Naturalness

Ekim and Schmitt [108] recently proposed an interpretable deep learning framework to analyze the contribution of different data modalities in predicting land naturalness. A U-Net [47] is trained to regress a naturalness index using multiple input sources: Sentinel-2 and Sentinel-1 imagery [112, 113], land cover maps [93], and nighttime lights [114]. To estimate the contribution of each modality to naturalness, the framework applies modality-wise occlusion—removing one input modality at a time by replacing it with zero values—and measures the resulting change in the network's latent representation. Specifically, the Euclidean distance between the latent embedding of the full input and that of the occluded input serves as a proxy for the modality's contribution. This allows the model to provide interpretability at both sample and dataset levels by quantifying which modalities most influence the final prediction. These insights are intended to guide future models, such as those by Ekim et al. [106], where only the most informative modalities might be used to reduce computational costs while preserving accuracy.

However, the proposed occlusion-based method comes with several limitations. Using zero values for occlusion may introduce unrealistic, out-of-distribution inputs that the model was not trained to handle, potentially skewing the relevance estimates. Additionally, interpreting distances in latent space as indicators of prediction relevance is an indirect approximation and must be accompanied by the assumption that the latent space is continuous, which may not always hold in practice. Furthermore, occlusion sensitivity can be highly dependent on the choice of occlusion strategy and the characteristics of the input data, which may affect reproducibility and robustness.

Method	Rule-based	XAI-based	High-res Output
HII [10]	✓	✗	✗
NI [11]	✓	✗	✓
RANA [16]	✓	✗	✗
ASOS [12]	✗	✓	✓
Modality Occlusion [108]	✗	✓	✓
Multimodal Learning [106]	✓	✗	✓

Table 3.1: Tick matrix comparing six methods for naturalness mapping or assessment across key dimensions. Columns indicate whether the method is rule-based (uses manually defined scoring logic), XAI-based (employs explainable AI techniques), and provides high-resolution output (e.g., 10 m maps). Methods include Human Influence Index (HII) [10], Naturalness Index (NI) [11], Relative Qualitative Naturalness Assessment (RANA) [16], Activation Space Occlusion Sensitivity (ASOS) [12], Modality Occlusion [108], and Multimodal Learning [106].

Chapter 4

Data

4.1 AnthroProtect Dataset

AnthroProtect is a high-resolution remote sensing dataset focused on the Fennoscandian region (Norway, Sweden, Finland), developed to distinguish naturalness from human-influenced landscapes [107]. It combines Sentinel-2 imagery with land cover and naturalness status annotations, providing class balance and ecological consistency to support applications such as scene classification and semantic segmentation.

4.1.1 Ecological Coherence

The study area forms an ecologically coherent biome. It features extensive coniferous forests, peatlands, wetlands, and low levels of natural fragmentation, extending continuously across national borders with consistent climatic, vegetative, and soil conditions. According to the Human Influence Index [10], protected areas in Fennoscandia are among the least impacted by human influence regions in Europe, highlighting its suitability for naturalness mapping.

4.1.2 Label Design

The dataset includes two land cover classes: *naturalness* and *non-naturalness*, derived from authoritative land use and land cover (LULC) sources to capture the structural differences between natural and human-altered environments.

The naturalness class comprises strictly protected areas listed in the World Database on Protected Areas (WDPA) [109], that include categories such as Ia (strict nature reserve), Ib (naturalness area), and II (national park). Only terrestrial polygons larger than $50\,\mathrm{km}^2$ were retained to ensure spatial relevance and minimize anthropogenic influence.

The non-naturalness class was constructed from the CORINE Land Cover (CLC) 2018 dataset [115], including CLC classes 1 (artificial surfaces) and 2 (agricultural areas). To produce spatially coherent masks, morphological operations (binary closing, opening, dilation) were applied, followed by filtering to exclude regions smaller than 50 km^2.

4.1.3 Imagery Source and Preprocessing

Imagery was sourced from Sentinel-2 Level-2A products, which provide atmospherically corrected surface reflectance. To ensure consistent seasonal and illumination conditions, only images from July 1 to August 30, 2020, were used. Cloud contamination was minimized using the QA60 and Scene Classification Layer (SCL) masks [116]. A temporal composite was computed using the 25th percentile of clear-sky reflectance values per band, followed by manual quality inspection. The resulting composites were tiled into non-overlapping 256×256 pixel patches, each covering 2.56 km $\times$ 2.56 km on the ground.

4.1.4 Dataset Composition

The final dataset contains 23,919 multispectral tiles, each including 10 Sentinel-2 bands: B2 (blue), B3 (green), B4 (red), B5–B7 (red edge), B8 (NIR), B8A (narrow NIR), and B11–B12 (SWIR). Class distribution is as follows:

- **Naturalness**: 7,003 tiles (369 from Ia, 4,512 from Ib, and 2,122 from II)
- **Non-naturalness**: 16,916 tiles from various rural and urban areas

4.1.5 Data Splitting Strategy

To mitigate spatial autocorrelation bias, a two-stage clustering strategy was employed. DBSCAN [117] was first used to identify spatially distinct regions, followed by k-means clustering [118] for refinement. All tiles within a cluster were assigned to the same data split to ensure statistical independence between training, validation, and test sets.

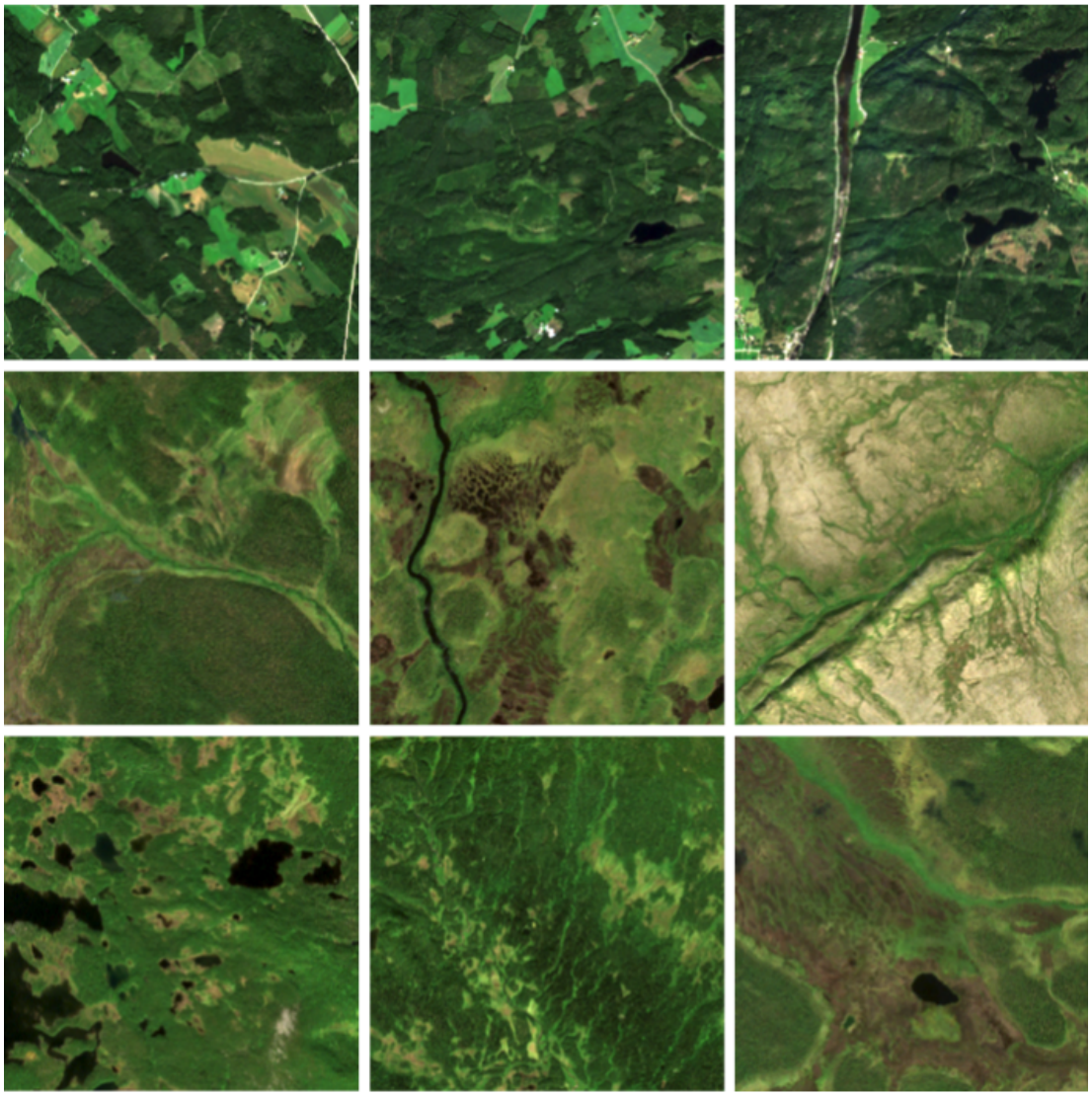


Figure 4.1: Example tiles from the AnthroProtect dataset showing protected areas in Fennoscandia. The satellite image patches depict landscapes with minimal direct human influence [107].

4.2 MapInWild Dataset

MapInWild [33] is a global, multi-modal remote sensing dataset designed for machine learning research on naturalness mapping and human impact assessment. In contrast to the regionally focused AnthroProtect, MapInWild covers diverse ecosystems and climate zones, offering broader geographic and spectral diversity.

4.2.1 Dataset Design, Modalities, and Composition

MapInWild integrates five complementary remote sensing sources:

- **Sentinel-1 (SAR)** [112]: GRD imagery with VV and VH polarizations at 10 m resolution
- **Sentinel-2** [113]: Level-2A surface reflectance with 10 spectral bands (B2–B12), composited separately for each season
- **ESA WorldCover** [93]: Global land cover map at 10 m resolution
- **VIIRS DNB** [114]: Nighttime lights as a proxy for human activity
- **WDPA polygons** [109]: Used to assign wilderness labels based on IUCN categories Ia, Ib, and II

Sentinel-2 composites were generated using the 25th percentile of reflectance values from 2020, with hemisphere-aware seasonal alignment. All sources were co-registered to a common spatial grid using a rubber-sheet algorithm for pixel-level alignment. The final dataset comprises 8,144 multi-modal image patches, each containing all five layers resampled to 1920×1920 pixels [33].

4.2.2 Sampling and Curation Strategy

A stratified and semi-automated pipeline ensured ecological representativeness and geographic diversity. WDPA polygons were filtered by area ($5\,\text{km}^2$ for Ia/Ib; $100\,\text{km}^2$ for II) and terrestrial extent. Minimum inter-sample distance thresholds (30 km for Ia/Ib; 50 km for II) were applied to avoid spatial clustering. Sampling was further guided by a weighting scheme based on the joint distribution of Köppen-Geiger climate zones [119] and ESA WorldCover classes [93], promoting biome diversity. For each selected polygon, a 20×20 km Area of Interest (AOI) centered on the polygon centroid was extracted.

In total, 910 protected AOIs were selected. To increase variability, 108 additional AOIs representing high human impact areas (e.g., cities, transport networks) were manually added. Each AOI is processed into eight co-registered image layers with a spatial extent of 1920×1920 pixels.

4.2.3 Quality Scoring System

Each seasonal Sentinel-2 composite was manually assessed by remote sensing experts. A quality score from 0 to 10 was assigned, with images having over 50% negative ratings discarded. For single-season experiments, the best-rated composite per AOI was selected, with preference given to summer in the case of ties.

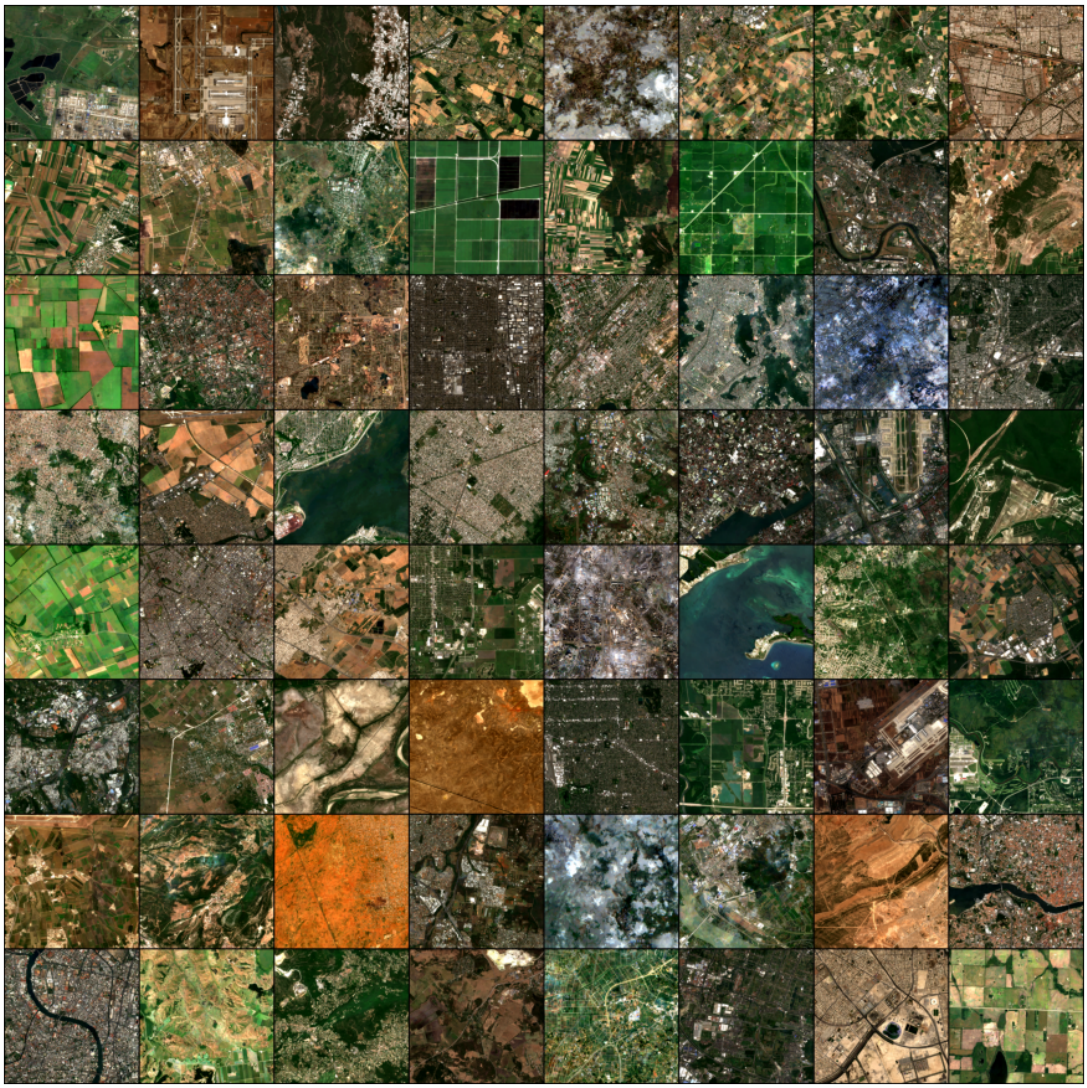


Figure 4.2: Collage of Sentinel-2 tiles from the MapInWild dataset, illustrating diverse land-use patterns and degrees of human activity across regions [33].

The main differences between the AnthroProtect and MapInWild datasets are summarized in Table 4.1.

Table 4.1: Key differences between the AnthroProtect [12] and MapInWild [33] datasets.

Feature	**AnthroProtect** [107]	**MapInWild** [33]
Geographic Scope	Regional (Fennoscandia)	Global
Imagery Modalities	Sentinel-2 only	Sentinel-1, Sentinel-2, VIIRS
Seasonal Coverage	Summer only	All seasons
Land Cover Source	CORINE Land Cover [115]	ESA WorldCover [93]

Chapter 5

GAN-based Activation Maximization for Naturalness Explanations

This work aims to explain the patterns and land cover classes contributing to naturalness in satellite imagery by integrating Activation Maximization (AM) [31, 2] and Generative Adversarial Networks (GANs) [41]. Specifically, it introduces a novel framework, AM-GANs for Naturalness, which modifies the Cycle Consistent Generative Adversarial Networks (CycleGANs) [3] architecture to generate images highlighting class-specific patterns, using the prediction of a classification network as feedback to guide the image generation process in the GANs.

The main contributions of this chapter are:

1. Develop a framework that integrates Activation Maximization [31, 2] into CycleGAN-styled [3] objective function to generate image pairs with maximized and minimized class-specific patterns.

2. Compare the images in each pair to generate heatmaps (attribution maps) highlighting patterns contributing to naturalness.

3. Provide quantitative and qualitative explanations of land cover classes contributing to naturalness using two datasets, and benchmark the framework against alternative methods (see Chapter 8).

5.1 State of the Art

AM interprets neural networks by altering or generating inputs that maximize the activation of specific neurons or layers, revealing the patterns a model associates with particular outputs [65, 2]. To improve the interpretability of the altered input, AM methods often employ regularizations such as L2 penalties, total variation, and image jittering or blurring [120, 2]. Despite these measures, AM can still produce unrealistic explainability outputs, be sensitive to initialization, and lack diversity or consistency in the discovered features [2, 121, 122].

To address these limitations, generative models—particularly GANs [41]—have been integrated into AM frameworks. For example, Nguyen et al. [2] employed Deep Generator Networks trained via GANs as priors, constraining optimization to the manifold of real images and producing more realistic, interpretable outputs. Later work, such as Activation Maximization GANs by Zhou et al. [31], incorporated class information into the GAN training process, enabling the generation of high-quality, class-specific visualizations. In the context of 3D point clouds, domain-specific generative models, such as those designed for 3D point clouds, have been shown to produce more perceptible and representative explanations than conventional AM methods [122]. However, in the field of remote sensing, such generative AM approaches remain underinvestigated.

Generative approaches have also been explored outside the AM setting for direct visual explanations. In medical imaging, the GANterfactual method [123] adapts a CycleGAN architecture [3] with an additional counterfactual loss based on the classifier's predictions to understand the patterns contributing to pneumonia in x-ray images. This ensures that generated images are realistic, belong to the opposite class, and preserve all features unrelated to the classification. GANterfactual does not output attribution maps; instead, it produces a pair of images—the original and a counterfactual—where only class-relevant features are modified while the overall structure remains intact. In a pneumonia detection case study, this meant maintaining lung anatomy while altering texture and opacity patterns indicative of disease [123].

To the best of my knowledge, at the time of writing this thesis, I am the first to develop a framework that combines a CyclGAN-styled model and AM for remote sensing applications, specifically aimed at improving explainability and enhancing the knowledge about the appearance of naturalness satellite imagery.

5.2 Methodology

The framework integrates AM [31, 2] with CycleGAN [3] to identify and highlight the patterns that characterize the appearance of naturalness in satellite imagery. Figure 5.1 summarizes three phases that happen within the framework's training and inference: (i) Regressor training, (ii) GANs training, and (iii) Attribution mapping. By combining domain knowledge with generative modeling [41], the framework generates class-specific image pairs and attribution maps that provide explanations of the appearance of naturalness.

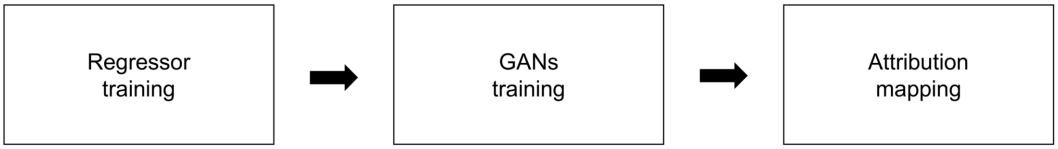


Figure 5.1: Pipeline overview consisting of three sequential phases: Regressor training, which learns class-specific representations; GANs training, which enhances these representations via CycleGAN-guided activation maximization to generate maximized and minimized image pairs; and Attribution mapping, which explains patterns contributing to naturalness by comparing the generated image pairs.

5.2.1 Regressor Training (Pattern Learning)

A ResNet-50 [45] regressor r is trained to predict a continuous naturalness score between 0 (fully non-naturalness) and 1 (strictly naturalness) from Sentinel-2 RGB images. CutMix augmentation [124] blends patches from regions representing non-naturalness and naturalness, exposing the model to mixed compositions and promoting robust learning of intermediate naturalness levels. The training data consist of both the original images and their CutMix-augmented counterparts. For example, if a CutMix image contains 70% non-natural and 30% natural regions, its target naturalness score is set to 0.3. Given training pairs (X^i, y^i) with $y^i \in [0, 1]$, the regression objective is

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \left(y^i - r(\mathsf{X}^i) \right)^2. \tag{5.1}$$

After training, the regressor is frozen and serves as a semantic guide for the naturalness-maximizing and -minimizing generators.

5.2.2 GANs Training (Pattern Enhancement)

GAN training proceeds in two stages. In the first stage (pretraining), the generators w^+ and w^- learn to synthesize realistic Sentinel-2 imagery without

AM integration. The focus is on adversarial, cyclic, LPIPS, and feature-matching losses to ensure visual fidelity. In the second stage (semantic training), the frozen regressor r is integrated to guide both generators via AM losses, enhancing or suppressing semantic features related to naturalness.

Each sub-network employs a StyleGAN2-inspired generator [125] with adaptive instance normalization (AdaIN) layers [126] and a latent mapping network, paired with PatchGAN discriminators d^+ and d^- [127]. After training the regressor r, it is used in inference mode to guide the generation process. Specifically, its prediction serves as feedback for the generators within the AM loss, steering them toward image transformations that respectively increase or decrease the regressor's output.

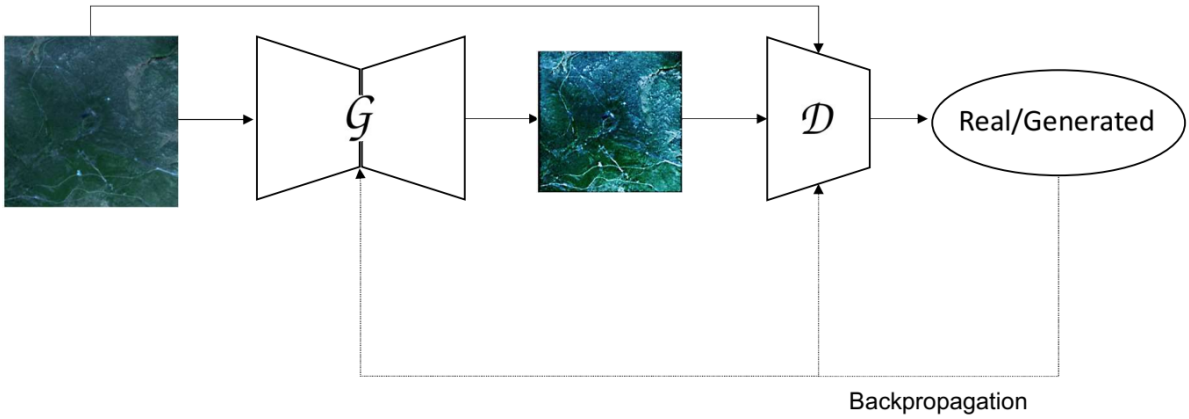


Figure 5.2: GAN pretraining phase (Stage 1): generators learn to produce realistic satellite imagery while discriminators enforce realism and perceptual similarity. Activation maximization is excluded at this stage.

To reveal discriminative patterns, AM is embedded in a CycleGAN-style framework [3] with generators w^+ (naturalness maximizer) and w^- (naturalness minimizer), each paired with a PatchGAN discriminator d^+ and d^-. Given an input image X, the generators produce $\hat{\mathsf{X}}^+ = w^+(\mathsf{X})$ and $\hat{\mathsf{X}}^- = w^-(\mathsf{X})$. A composite objective integrates multiple loss terms to steer the generators while preserving structure and realism. where X is the input RGB image; $\hat{\mathsf{X}}^\pm$ are the maximized/minimized outputs; $w^\pm$ and $d^\pm$ are generators and discriminators; r is the trained regressor ($[0, 1]$); $\|\cdot\|_1$ is the L_1 norm; BCE is binary cross-entropy; and $\lambda_{\text{cyc}}, \lambda_{\text{AM}}, \lambda_{\text{LPIPS}}, \lambda_{\text{FM}}$ are the corresponding loss weights.

Cycle consistency maintains reversibility to preserve content [3]:

$$\mathcal{L}_{\text{cyc}} = \left\|\mathsf{X} - w^-\big(w^+(\mathsf{X})\big)\right\|_1 + \left\|\mathsf{X} - w^+\big(w^-(\mathsf{X})\big)\right\|_1 . \tag{5.2}$$

Adversarial losses per GAN encourage the generators to produce realistic

outputs:

$$\begin{aligned} \mathcal{L}_{\text{adv}}^{+} &= \Big(d^{+}(\hat{\mathsf{X}}^{+}) - 1 \Big)^2 , \\ \mathcal{L}_{\text{adv}}^{-} &= \Big(d^{-}(\hat{\mathsf{X}}^{-}) - 1 \Big)^2 . \end{aligned} \tag{5.3}$$

AM utilizes the regressor's feedback to steer the generators to generate images with maximized or minimized naturalness scores:

$$\begin{aligned} \mathcal{L}_{\text{AM}}^{+} &= \text{BCE}\Big(1,\, r(\hat{\mathsf{X}}^{+}) \Big) , \\ \mathcal{L}_{\text{AM}}^{-} &= \text{BCE}\Big(0,\, r(\hat{\mathsf{X}}^{-}) \Big) . \end{aligned} \tag{5.4}$$

A perceptual penalty (LPIPS) [128] constrains unintended changes by keeping generated images close to the input in deep feature space:

$$\begin{aligned} \mathcal{L}_{\text{LPIPS}}^{+} &= \sum_{l} \frac{1}{H_l W_l} \sum_{u,v} \sum_{c} w_c^{(l)} \left(\hat{F}^{(l)}_{w^{+},c}(\mathsf{X})_{u,v} - \hat{F}^{(l)}_{w^{+},c}(\hat{\mathsf{X}}^{+})_{u,v} \right)^2 , \\ \mathcal{L}_{\text{LPIPS}}^{-} &= \sum_{l} \frac{1}{H_l W_l} \sum_{u,v} \sum_{c} w_c^{(l)} \left(\hat{F}^{(l)}_{w^{-},c}(\mathsf{X})_{u,v} - \hat{F}^{(l)}_{w^{-},c}(\hat{\mathsf{X}}^{-})_{u,v} \right)^2 . \end{aligned} \tag{5.5}$$

where $\hat{F}^{(l)}_{w^{\pm},c}(\cdot)$ denotes the feature map from layer l of the corresponding generator w^{+} or w^{-}, channel c at spatial location (u, v). The weights $w_c^{(l)}$ balance the contribution of each channel. H_l and W_l are the spatial height and width of the feature maps at layer l.

Feature matching stabilizes training by aligning intermediate discriminator statistics rather than final real/fake scores [129]:

$$\begin{aligned} \mathcal{L}_{\text{FM}}^{+} &= \sum_{l} \left\| \mathsf{F}^{(l)}_{d^{+}}(\mathsf{X}) - \mathsf{F}^{(l)}_{d^{+}}(\hat{\mathsf{X}}^{+}) \right\|_1 , \\ \mathcal{L}_{\text{FM}}^{-} &= \sum_{l} \left\| \mathsf{F}^{(l)}_{d^{-}}(\mathsf{X}) - \mathsf{F}^{(l)}_{d^{-}}(\hat{\mathsf{X}}^{-}) \right\|_1 . \end{aligned} \tag{5.6}$$

The generator objectives combine these terms with scalar weights:

$$\mathcal{L}_{\text{gen}}^{+} = \mathcal{L}_{\text{adv}}^{+} + \lambda_{\text{cyc}} \mathcal{L}_{\text{cyc}} + \lambda_{\text{AM}} \mathcal{L}_{\text{AM}}^{+} + \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}}^{+} + \lambda_{\text{FM}} \mathcal{L}_{\text{FM}}^{+}. \tag{5.7}$$

$$\mathcal{L}_{\text{gen}}^{-} = \mathcal{L}_{\text{adv}}^{-} + \lambda_{\text{cyc}} \mathcal{L}_{\text{cyc}} + \lambda_{\text{AM}} \mathcal{L}_{\text{AM}}^{-} + \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}}^{-} + \lambda_{\text{FM}} \mathcal{L}_{\text{FM}}^{-}. \tag{5.8}$$

Although this two-stage design increases computational cost, it yields more coherent and semantically meaningful attribution maps than traditional single-stage approaches [30, 2].

5.2.3 Attribution Mapping

To highlight spatially discriminative patterns, attribution maps are computed as the absolute difference between naturalness-maximized and minimized outputs:

$$\text{Attribution}(u, v) = \frac{1}{C} \sum_{c=1}^{C} \left| \hat{\mathsf{X}}^{+}_{c,u,v} - \hat{\mathsf{X}}^{-}_{c,u,v} \right| \tag{5.9}$$

where C is the number of channels, and X^+, X^- are the maximized and minimized images. These maps are overlaid on the original inputs to visualize spectral–spatial regions influencing the regressor's prediction. Figure 5.3 illustrates the interaction between the framework components during attribution generation.

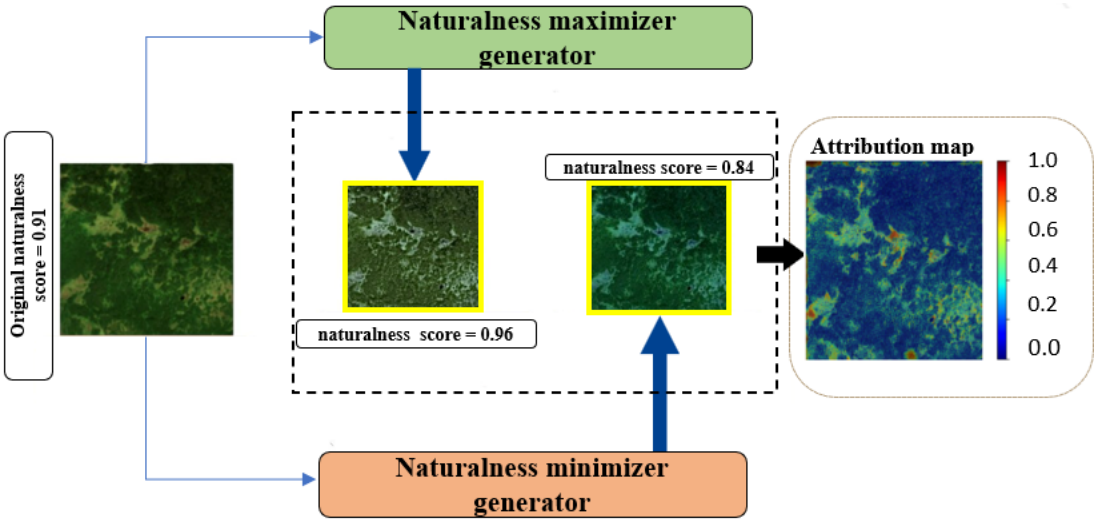


Figure 5.3: The AM-GANs for Naturalness in inference: the trained maximizer and minimizer generators map the input image into generated images with higher or lower naturalness scores, enabling attribution map generation based on comparing the generated image pairs.

5.3 Experimental Setup

We evaluate the framework on AnthroProtect [12] and the MapInWild [33] over Fennoscandia, using only RGB channels and the associated land cover segmentation maps. Channels are normalized per training set statistics, and predefined training, validation, and test splits of 80%, 10%, and 10% are applied.

The experiments use images with a spatial size of 256×256 pixels and three spectral bands (red, green, blue). The discriminators operated on overlapping 70×70 pixel patches to evaluate local realism and guide generator updates.

5.3.1 Regressor Training and Performance Evaluation

The regressor r demonstrates high accuracy and stability across both training and validation sets, as summarized in Table 5.1. The low mean absolute error (MAE) and root mean squared error (RMSE) indicate that r reliably estimates naturalness scores with minimal deviation from the reference values. Similarly, the small bias and error standard deviation confirm the absence of systematic over- or underestimation, suggesting strong generalization to unseen data.

Table 5.1: Performance metrics of the regressor r.

Set	**MAE** $\downarrow$	**RMSE** $\downarrow$	**Bias** $\downarrow$	**Error Std** $\downarrow$
Train	0.0061	0.0078	0.0031	0.0072
Validation	0.0065	0.0082	0.0034	0.0075

5.3.2 Generative Model Training and Evaluation

Training followed the two-stage process as described earlier. In total, the framework was trained for 20 pretraining epochs followed by 50 semantic training epochs. The complete training required approximately 8 hours on an NVIDIA A100 GPU. The main hyperparameters and optimization settings used during both stages are summarized in Table 5.2.

Table 5.2: Hyperparameters used for training the framework.

Component	**Setting**
Optimizer (G)	Adam, lr $= 1 \times 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$
Optimizer (D)	Adam, lr $= 4 \times 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$
Batch Size	16
Epochs	20 (pretraining) + 50 (main training)
λ_{adv}	0.5
λ_{LPIPS}	11.0
λ_{FM}	5.0
λ_{ACT}	7.0

To assess the quality and realism of the generated imagery, we use four complementary evaluation metrics:

- **L1:** Mean absolute error between generated and reference pixels (lower is better).
- **SSIM:** Structural similarity index [130], measuring perceived structural consistency (higher is better).
- **LPIPS:** Learned perceptual image patch similarity [128], comparing deep features (lower is better).
- **PSNR:** Peak signal-to-noise ratio [131], evaluating pixel-level fidelity in decibels (higher is better).

Table 5.3: Reconstruction performance of the trained GAN on training and validation sets. Arrows indicate whether higher (↑) or lower (↓) values are preferable.

Set	L1 ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Train	0.0325	0.9486	0.0252	40.2280
Validation	0.0401	0.9464	0.0272	38.8361

5.4 Results

We present both qualitative and quantitative evaluations of the appearance of naturalness using the proposed AM-GANs for Naturalness framework. The training performance metrics and results for the regressor and generative model were presented in the previous section 5.3. Chapter 8 further discusses the alignment of the framework with other established and proposed naturalness assessment approaches.

5.4.1 Qualitative Evaluation of the Attribution Maps

Figure 5.4 presents examples of attribution maps generated by the framework. The highlighted patterns appear continuous and spatially coherent, capturing semantically meaningful features that contribute to the model's naturalness estimation [1, 2].

Additional qualitative examples are provided in chapter 10 to support extended visual interpretation across diverse landscapes and land cover compositions.

5.4.2 Quantitative Results

We computed the average of each land cover class to naturalness by spatially overlaying the generated attribution maps with the corresponding land cover segmentation maps. Since each original image is associated with an attribution map and a corresponding ground truth segmentation map, we calculated the sum of attribution values for all pixels belonging to each land cover class and divided it by the number of pixels that class occupies within the image. This procedure was repeated for all images in the dataset, resulting in an average attribution value for each land cover class, computed as:

$$\text{Attribution}_c = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathsf{M}_{i,c}|} \sum_{(u,v) \in \mathsf{M}_{i,c}} \text{Attribution}_i(u, v) \tag{5.10}$$

where $\text{Attribution}_i(u, v)$ is the attribution score at pixel (u, v) in image i, $\mathsf{M}_{i,c}$ is the set of pixels belonging to land cover class c in image i, $|\mathsf{M}_{i,c}|$ is

the number of pixels of class c in image i, and N is the total number of images in the dataset [115, 93, 33, 12]. As shown in Table 5.4, wetlands and sparse vegetation classes consistently exhibit the highest contributions to naturalness. Other anthropogenic land cover types, such as Agricultural areas, Cropland built-up areas, roads, and dump sites, exhibit zero contribution values. Water-related classes display moderate contributions, reflecting their varied ecological importance in Fennoscandian environments.

Table 5.4: Mean and standard deviation of each land cover class contribution to naturalness for the AnthroProtect and MapInWild datasets.

AnthroProtect with CORINE Land Cover Classes [12, 115]		
Class Name	Mean	Std
Transitional woodland shrub	0.146	0.012
Coniferous forest	0.239	0.013
Sparsely vegetated areas	0.252	0.004
Natural grassland	0.215	0.012
Water bodies	0.103	0.003
Bare rock	0.252	0.004
Inland marshes	0.900	0.015
Water courses	0.103	0.003
Mixed forest	0.239	0.013
Moors and heathland	0.146	0.012
Peat bogs	0.900	0.015
Glaciers and perpetual snow	0.129	0.024
Broad-leaved forest	0.239	0.013
Sea and ocean	0.103	0.003
MapInWild with ESA WorldCover Classes [33, 93]		
Shrubland	0.146	0.012
Trees	0.239	0.013
Grassland	0.215	0.012
Bare/Sparse vegetation	0.252	0.004
Herbaceous wetland	0.900	0.015
Open water	0.103	0.003
Snow and ice	0.129	0.024

5.5 Discussion

The results confirm that the proposed GAN-based activation maximization framework effectively identifies patterns in satellite imagery that contribute to naturalness. The attribution maps, derived from the difference between naturalness-maximized and minimized outputs, consistently highlight wetland ecosystems—including Inland marshes, Peat bogs, and Herbaceous wetlands—as the strongest contributors to naturalness (Table 5.4). These ecosystems are known for their importance in carbon storage, habitat diversity, and water regulation, and their prominence in the attribution results aligns with ecological literature on Nordic biodiversity hotspots [132, 10, 13, 133].

5.6 Conclusion

This chapter presented a framework that combines activation maximization [2, 31] with CycleGAN-based generative modeling [3] to interpret the naturalness in satellite imagery. A regressor predicts naturalness scores, and two generators generate maximized and minimized variants to reveal contributing patterns. The resulting attribution maps consistently highlight ecologically relevant land cover types, such as wetlands, trees, and sparse vegetation classes, emerging as dominant contributors.

The method was evaluated on the AnthroProtect dataset [12] and the corresponding overlapping areas within [33] datasets, with attribution scores aligning with known ecological indicators [132] and external naturalness indices. By leveraging generative modeling [41, 3], the approach produces valid, semantically consistent explanations that extend beyond traditional saliency maps [34, 64].

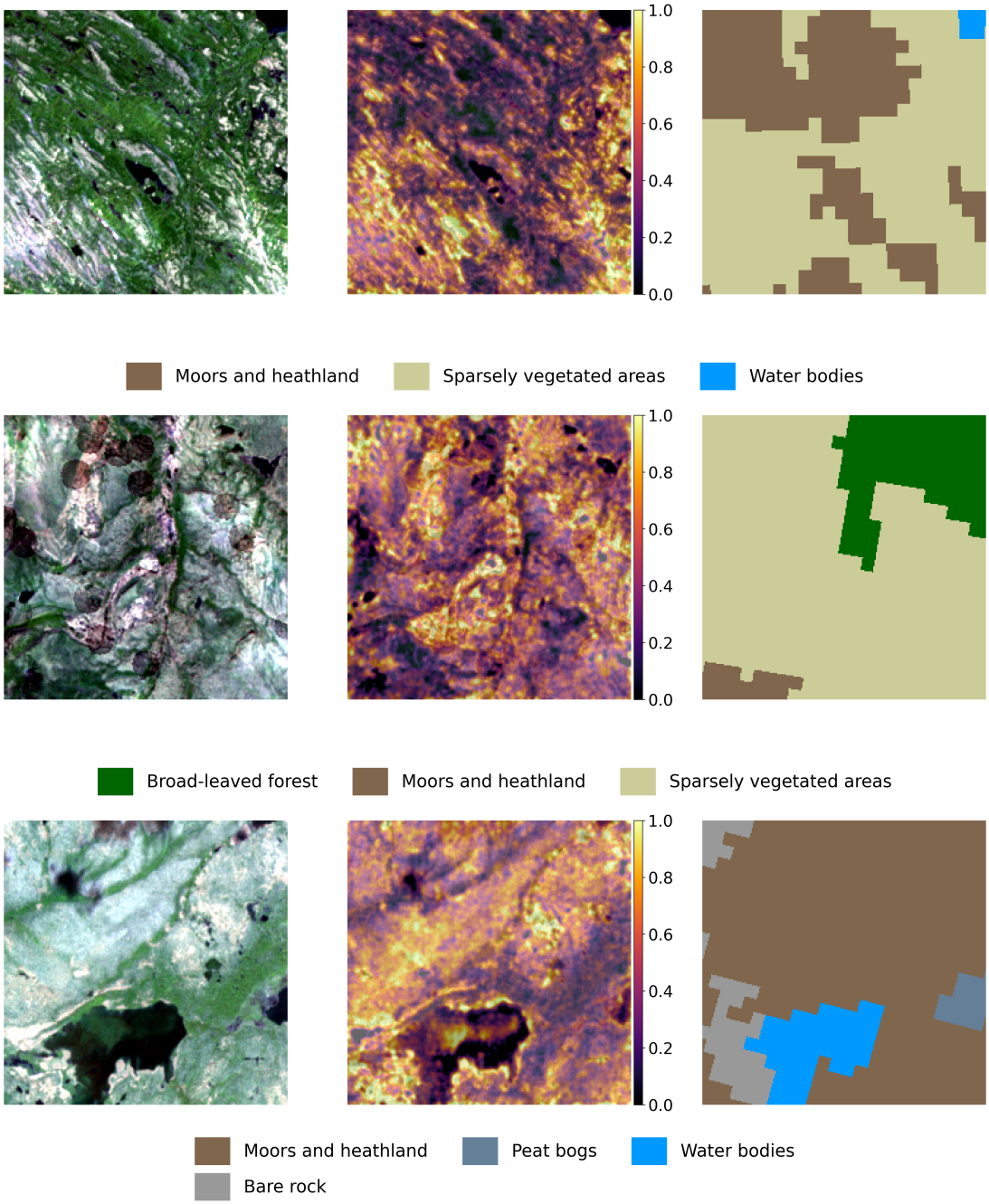


Figure 5.4: The qualitative results of the AM-GANs for the Naturalness framework (Part 1 of 2). Each triplet of images includes: (1) the original RGB input, (2) the corresponding attribution map highlighting spatial patterns contributing to naturalness estimation, and (3) the associated semantic segmentation mask for land cover association. The examples illustrate how the framework selectively identifies the ecological patterns that contribute to the appearance of naturalness.

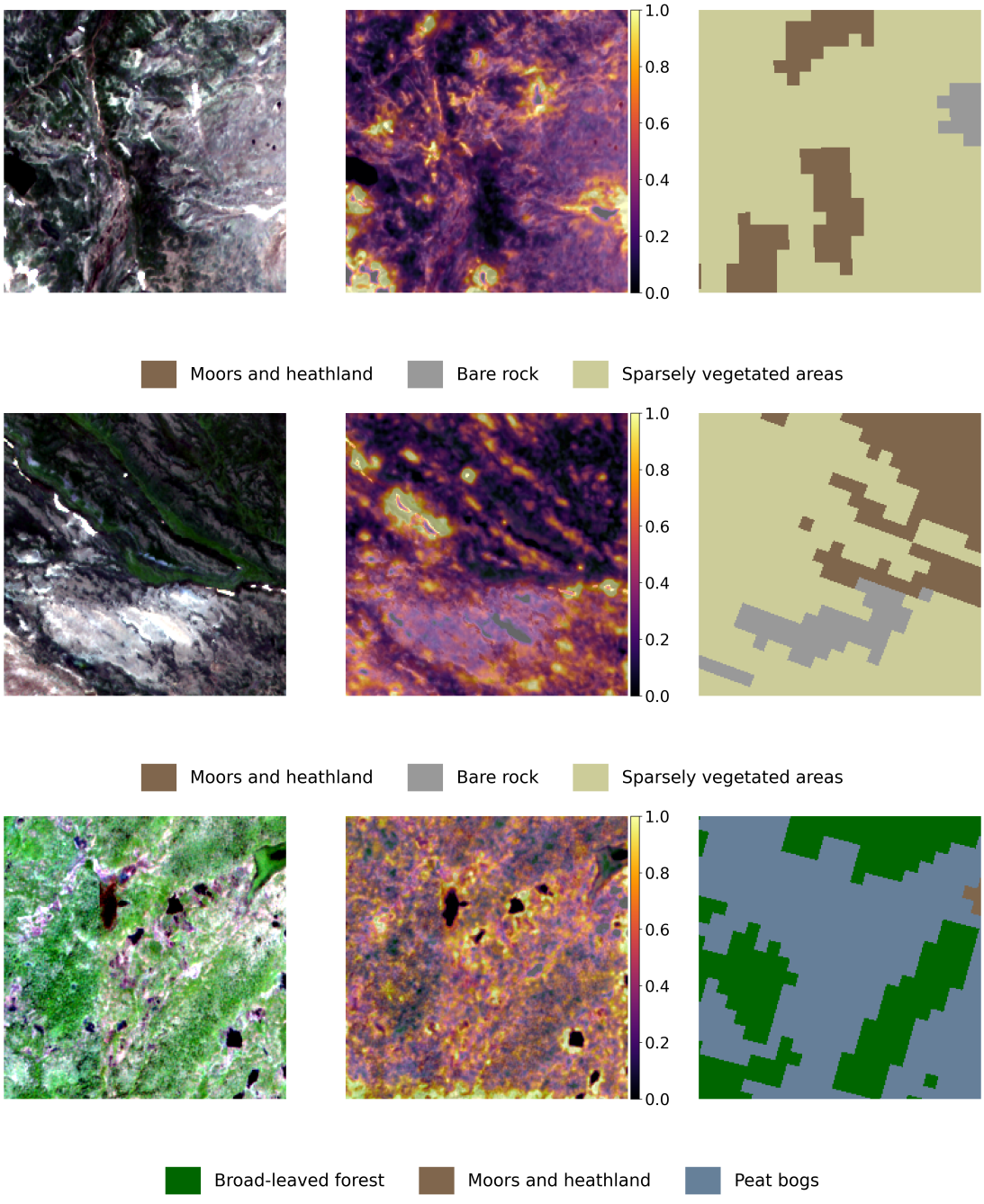


Figure 5.5: The qualitative results of the AM-GANs for the Naturalness framework (Part 2 of 2). Each triplet of images includes: (1) the original RGB input, (2) the corresponding attribution map highlighting spatial patterns contributing to naturalness estimation, and (3) the associated semantic segmentation mask for land cover association. The examples illustrate how the framework selectively identifies the ecological patterns that contribute to the appearance of naturalness.

Chapter 6

Surrogate Modeling for Confident Naturalness Explanations

Semantic segmentation models can learn complex patterns contributing to classes such as naturalness in satellite imagery [50]. Compared to standard classification models, they typically have larger capacities and operate at the pixel level, assigning a semantic label to each pixel in the image. This enables them to capture fine-grained spatial and spectral information about land cover types [5, 50]. The knowledge learned by a trained semantic segmentation model can therefore be highly valuable for interpreting naturalness. By applying explainability techniques, we can access and utilize this latent knowledge to gain a deeper understanding of the spatial patterns and land cover features driving the model's predictions [134, 135, 71, 24].

Unlike existing naturalness assessment methods such as those of Sanderson [10], Winter [16], and Ekim [33], which are rule-based and may suffer from bias, and unlike the XAI-based only approach presented in the previous chapter [1], this chapter introduces a new framework that confidently explains naturalness from a semantic segmentation model using surrogate modeling. The framework also incorporates UQ associated with the model's predictions, enabling an assessment of the uncertainty associated with the knowledge extracted [28].

The proposed framework, termed Confident Naturalness Explanations (CNE) [4], uses a logistic regression model as a surrogate to approximate the behavior of a semantic segmentation model. Since logistic regression is inherently explainable, its coefficient weights can be directly interpreted to assess the contribution of different land cover patterns to naturalness. To account for predictive uncertainty, CNE applies Monte Carlo Dropout (MC-Dropout) [6] during inference. Finally, the CNE framework introduces the CNE index, which integrates both the contribution of land cover classes and their associated uncertainty, providing a naturalness assessment for a specific region that reflects not only the strength

of contribution to naturalness but also the confidence of the prediction. To the best of our knowledge, this is the first approach to explain the appearance of naturalness in satellite imagery using this combination of surrogate modeling as an XAI approach and UQ.

The key contributions of this chapter are:

1. A novel framework that incorporates XAI and UQ for interpreting semantic segmentation models.

2. Introduction of the CNE index to assess naturalness by integrating pattern contribution and prediction uncertainty.

3. Quantitative and qualitative analysis of naturalness in Fennoscandia based on the results of the CNE index.

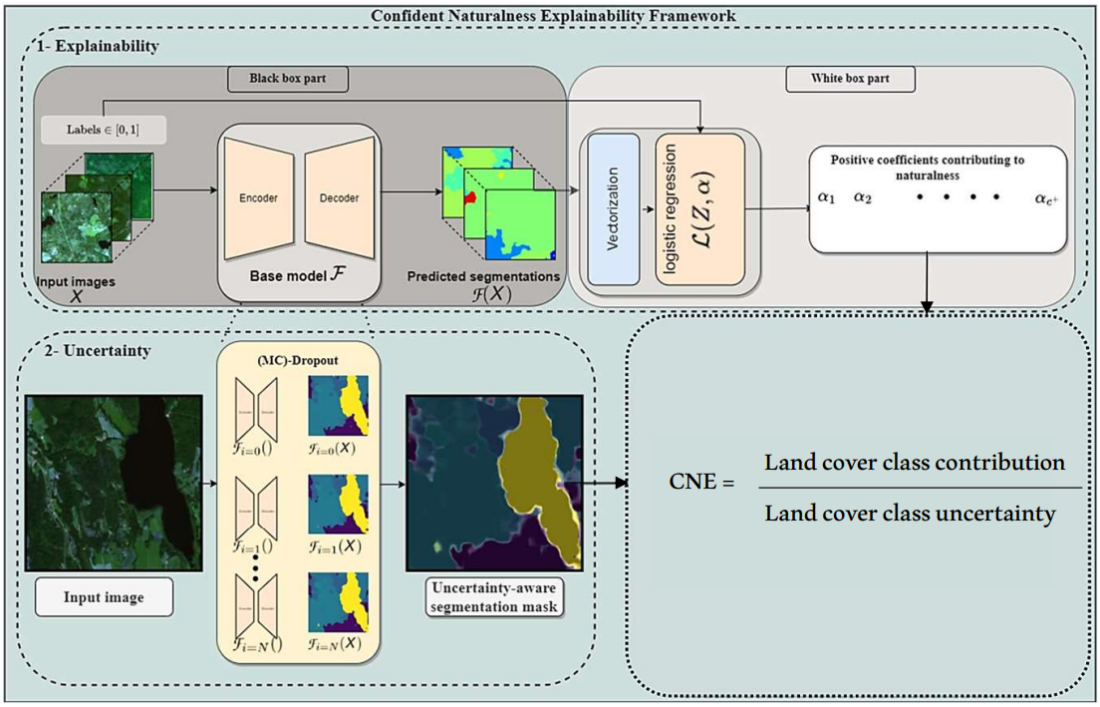


Figure 6.1: Overview of the CNE framework. In the explainability path (top), input images are processed by the segmentation model to produce predicted segmentation masks Y, which are vectorized and paired with naturalness labels to train a logistic regression surrogate $\mathcal{L}$. Only positive coefficients are retained to compute attribution. In the uncertainty path (bottom-left), MC-Dropout generates multiple segmentation outputs, allowing estimation of class-wise uncertainty maps S. Both paths are then combined (bottom-right) to compute the CNE index, assigning confidence-weighted relevance scores to land cover classes. The uncertainty estimation process is detailed in Figure 6.3.

6.1 State of the Art

Semantic segmentation refers to the task of assigning a class label to each pixel in an image, enabling fine-grained representation of the components of satellite imagery. Architectures such as U-Net [47] and DeepLab [5] are widely used for this purpose. U-Net, originally developed for biomedical image segmentation, employs an encoder–decoder structure with skip connections to preserve spatial details, while DeepLab [5] incorporates atrous convolution to efficiently capture multi-scale contextual information. These architectural advances have led to strong performance across diverse domains, including medical imaging, autonomous driving, and environmental monitoring.

Despite these advances, models like U-Net and DeepLab remain black-box systems with limited interpretability. This lack of transparency poses challenges in applications where understanding the model's reasoning is essential for trust, accountability, and informed decision-making [134, 135]. A more detailed overview of semantic segmentation architectures, such as DeepLab and U-Net, is provided in chapter 2.

Several methods have been proposed to improve the interpretability of semantic segmentation models. Vinogradova et al. [136] proposed Seg-Grad-CAM, the first adaptation of Gradient-weighted Class Activation Mapping (Grad-CAM) [34] to semantic segmentation networks. Seg-Grad-CAM computes class-specific relevance maps at the pixel level by backpropagating gradients of class scores through convolutional layers, providing coarse yet interpretable visual explanations of segmentation decisions. However, its spatial resolution remains limited, especially for fine-grained structures or overlapping classes. Building upon this idea, Schorr et al. [135] introduced an extended framework with the Neuroscope toolbox, which provides a graphical interface that enables systematic inspection of encoder and decoder layers. Their analysis revealed that late encoder layers yield smoother and more class-consistent explanations, whereas decoder layers offer finer spatial detail but more fragmented attributions.

Surrogate modeling represents another strategy for interpretability, where a simpler, interpretable model approximates the predictions of a complex network. A well-known example is Local Interpretable Model-agnostic Explanations (LIME) [58], described in detail in Section 2.4.7. LIME perturbs the input by modifying superpixels, observes the effect on segmentation outputs, and fits a simple linear model to approximate the local decision boundary. While effective for class-specific, localized explanations, its inherent locality limits the ability to capture dataset-wide behavior. In the context of semantic segmentation, DSEG-LIME [137] extends this approach by replacing arbitrary superpixels with semantically meaningful segments derived from foundation models such as SAM [138], while preserving LIME's surrogate modeling principle. This

adaptation enhances region-level interpretability by producing smoother and more semantically coherent explanations that align with object-level structures in segmentation outputs, although it introduces a higher computational cost due to the complexity of foundation-model-based segmentation.

Research about explainability of semantic segmentation models, especially within remote sensing applications, remains an emerging field that requires further investigation, with most existing work focusing on scene classification tasks [134]. This highlights the need for global surrogate models capable of capturing dataset-wide relationships and approximating the decision process over the full data distribution, thereby providing comprehensive and interpretable insights into complex phenomena such as naturalness. Logistic regression as a global surrogate model is particularly suited for this purpose due to its inherent interpretability and ability to quantify the contributions of each input feature—such as land cover classes—to the model's outputs.

In the context of uncertainty quantification for semantic segmentation of remote sensing data, Rottmann et al. [139] introduced a meta-classification framework that predicts the reliability of segmentation outputs using aggregated dispersion measures of softmax probabilities. Their post-hoc, model-agnostic approach identifies erroneous or uncertain segments and estimates segment-wise Intersection over Union (IoU) values without modifying the segmentation network.

Bayesian approximation methods such as deep ensembles [80, 6] and MC-dropout [140, 141] remain the most widely adopted alternatives. Both approximate Bayesian inference through stochasticity during inference: MC-Dropout [6] applies dropout masks at test time, while deep ensembles train multiple independently initialized networks. MC-Dropout is particularly suited for satellite imagery semantic segmentation tasks, due to its simplicity and ability to produce spatially resolved uncertainty maps [4]. However, it may be sensitive to the dropout rate selection. Since the main focus of this thesis lies on explainability techniques, the UQ approaches are discussed only briefly.

6.2 Method

We propose the Confident Naturalness Explanation (CNE) framework [4], which combines semantic segmentation, UQ, and a globally interpretable surrogate model. The CNE framework quantifies the confidence-weighted contribution of land cover classes to naturalness using satellite imagery. The framework is independent of specific choices of surrogate or uncertainty quantification models: logistic regression can be replaced by other interpretable models such as decision trees, and MC-Dropout can be substituted with deep ensembles or even measures

derived directly from softmax dispersion [6, 140, 141, 139].

The overall design is a grey-box approach [142], in which a black-box segmentation network is coupled with a white-box logistic regression model. The segmentation model provides spatially detailed land cover predictions, while the regression model assigns interpretable importance weights to predefined patterns using the ground truth naturalness labels (naturalness or non-naturalness). This combination bridges data-driven feature extraction with model transparency, enabling interpretable analysis of the patterns contributing to naturalness.

In this work, we adopt logistic regression together with MC-Dropout, since this combination yields globally interpretable explanations, intuitive confidence estimates. Through this design, we make use of the latent knowledge of a trained semantic segmentation model about naturalness with the uncertainty associated with each land cover class, thereby enhancing our understanding of the appearance of naturalness in an uncertainty-aware and data-driven way.

6.2.1 Semantic Segmentation and Pattern Vectorization

Semantic segmentation is performed using the DeepLabV3 architecture [143], built on a ResNet-50 backbone [45]. This model effectively balances fine spatial detail with large-scale contextual understanding through atrous convolutions and an Atrous Spatial Pyramid Pooling (ASPP) module [5], enabling precise object boundary delineation.

The segmentation model $\mathcal{F}$ maps the input image X to its segmentation output:

$$\mathcal{F}(\mathsf{X}) = \mathsf{Y}. \tag{6.1}$$

The output tensor $\mathsf{Y}^{C\times H\times W}$ contains one channel per land cover class, where C is the number of classes and $H \times W$ denotes the spatial resolution. Each predicted segmentation mask Y comprises C binary masks, one for each class, indicating which pixels are assigned to class c.

To obtain a compact yet meaningful representation, the segmentation mask is transformed into a $C \times 1$-dimensional pattern vector $\boldsymbol{z}$, where each element represents the abundance of a specific land cover class in the predicted segmentation map:

$$\boldsymbol{z}_c = \sum_{h,w} \mathsf{Y}_{c,h,w}. \tag{6.2}$$

This transformation converts the segmentation map into a $1 \times C$ vector representing land cover composition, formatted for direct use as input to logistic regression, while omitting fine spatial context. Although this vectorization discards spatial dependencies, it preserves the compositional structure of the scene, enabling interpretable global analysis of land cover contributions to naturalness.

6.2.2 Global Surrogate Model: Logistic Regression

The second component of the framework introduces a global surrogate model to achieve interpretability. A logistic regression model $\mathcal{L}(\boldsymbol{z}; \boldsymbol{\alpha})$ is trained on the pattern vectors derived from the segmentation outputs:

$$\mathcal{L}(\boldsymbol{z}; \boldsymbol{\alpha}) = \frac{1}{1 + e^{-\boldsymbol{\alpha}^\top \boldsymbol{z}}}, \tag{6.3}$$

where $\boldsymbol{\alpha}$ is a C-dimensional coefficient vector, with each coefficient $\boldsymbol{\alpha}_c$ corresponding to a land cover class c. The magnitude and sign of $\boldsymbol{\alpha}_c$ indicate how strongly and in which direction class c contributes to the probability of an image being classified as natural. Positive coefficients are contributing to naturalness, while negative ones are contributing to non-naturalness. Logistic regression was selected for its transparency and alignment with algorithmic interpretability principles, providing an accessible means of quantifying class-level contributions to naturalness. Although the transformation sacrifices some spatial information, this trade-off ensures the generation of globally interpretable and explainable results.

6.2.3 Uncertainty Quantification

MC-Dropout [6] offers a simple and model-agnostic way to estimate epistemic uncertainty by enabling stochastic predictions during inference. A detailed explanation of its theoretical background is provided in Chapter 2.

To assess spatial uncertainty in the segmentation predictions, MC-Dropout is applied during inference. For each input image, J stochastic forward passes are performed, resulting in a set of J segmentation masks Y_j. This stochastic sampling provides an efficient approximation to Bayesian inference in deep learning, enabling uncertainty estimation without modifying the model architecture.

For each land cover class c, the mean prediction map A_c and the corresponding uncertainty map S_c are computed as:

$$\mathsf{A}_c = \frac{1}{J} \sum_{j=1}^{J} \mathsf{Y}_{c,j}, \qquad \mathsf{S}_c = \sqrt{\frac{1}{J} \sum_{j=1}^{J} (\mathsf{Y}_{c,j} - \mathsf{A}_c)^2}. \tag{6.4}$$

The resulting uncertainty map S_c represents the spatial standard deviation across predictions and provides a quantitative estimate of model confidence for each land cover class. High values in S_c indicate regions of high uncertainty, whereas low values correspond to confident predictions.

6.2.4 Confident Naturalness Explanation (CNE) Metric

Finally, the Confident Naturalness Explanation metric integrates attribution with uncertainty to quantify the confidence-weighted importance of each land cover class c:

$$\mathrm{CNE}_c = \frac{\max(\boldsymbol{\alpha}_c, 0)}{u_c}, \tag{6.5}$$

where $\boldsymbol{\alpha}_c$ is the logistic regression coefficient for class c, and u_c is the aggregated uncertainty across all pixels of that class:

$$u_c = \sum_{h,w} \mathsf{S}_{c,h,w}. \tag{6.6}$$

A high CNE_c score indicates that a class is both strongly associated with naturalness and predicted with high confidence. This integration moves beyond attribution alone by providing a systematic, uncertainty-aware interpretation of land cover contributions to naturalness.

6.2.5 Experimental Setup

The semantic segmentation model is trained on the training subsets of the AnthroProtect dataset and the overlapping region in the MapInWild dataset [107, 33] using the RGB channels. Training was conducted for 100 epochs with stochastic gradient descent (SGD), an initial learning rate of 0.001, momentum of 0.9, and a polynomial learning rate decay schedule. A mini-batch size of 16 was used. To support uncertainty quantification, dropout with probability $p_{\mathrm{drop}} = 0.1$ was enabled during both training and inference, following the MC-Dropout paradigm [6]. During inference, dropout was applied only in the last fully connected layer, and 25 stochastic forward passes were performed to generate the uncertainty-aware segmentation maps. The training of the logistic regression component was optimized using SGD to estimate the coefficients associated with each pattern contributing to naturalness.

After training, the segmentation model's predictive reliability and calibration were evaluated to ensure suitability for subsequent attribution and uncertainty analyses. The mean IoU (mIoU) across all classes is calculated as an overall measure of segmentation accuracy.

In addition to segmentation accuracy, calibration quality was evaluated using the Expected Calibration Error (ECE) [144], which quantifies the discrepancy between predicted confidence and empirical correctness. The trained DeepLabV3 model achieved a mIoU of 80.3% on the test set, with an ECE of 0.132, indicating strong segmentation performance and acceptable calibration.

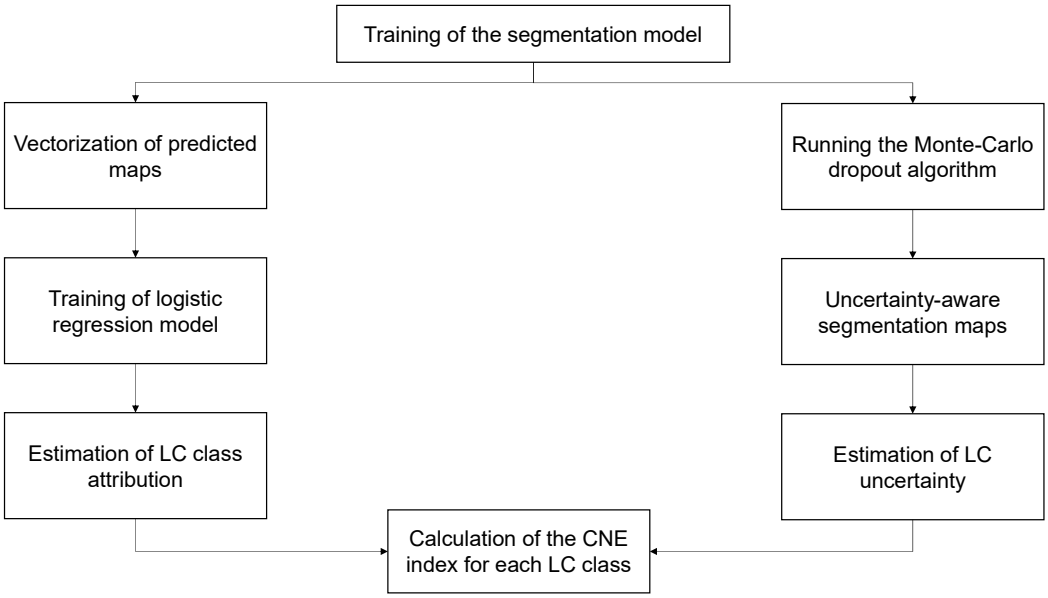


Figure 6.2: Flowchart of the CNE framework. It consists of two parallel paths operating on segmentation model outputs. The left path performs the explainability task: predicted segmentation masks are vectorized and used to train a logistic regression model with naturalness labels, resulting in attribution scores for each land cover class. The right path estimates uncertainty via MC-Dropout: multiple stochastic forward passes yield class-wise uncertainty maps. Both paths are combined to compute the final CNE index, which assigns confidence-weighted relevance scores to each land cover pattern. Details of the uncertainty estimation process are shown in Figure 6.3.

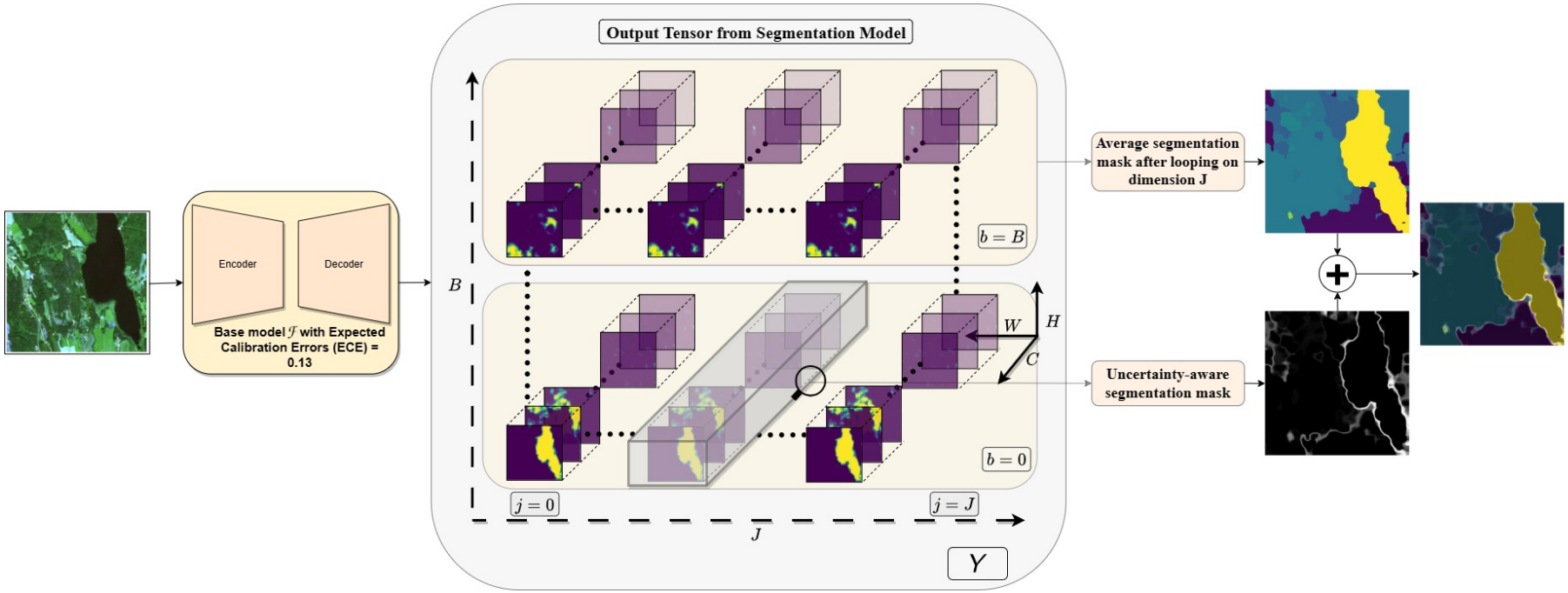


Figure 6.3: Illustration of the uncertainty estimation process. Input images are passed through the segmentation model with MC-Dropout. The output tensor Y is indexed by batch b and sample j. On the right, two outputs are shown: the upper image is the mean prediction over J runs (argmax per pixel), while the lower image shows the standard deviation map, where bright pixels indicate high uncertainty.

6.3 Results

In the following, we first present qualitative visualizations of uncertainty-aware segmentation maps, followed by quantitative results that illustrate the contribution of each land cover class to naturalness through the CNE metric.

6.3.1 Qualitative Results

Figure 6.4 shows two representative cases where predicted segmentation masks are complemented with uncertainty maps, highlighting regions of higher ambiguity in bright tones. These maps enable a spatial understanding of model reliability beyond class-level metrics. Further examples are provided in Figure 6.7, which present diverse land cover patterns. The uncertainty overlays reveal that boundaries between classes and spectrally heterogeneous areas tend to yield higher uncertainty, while regions such as shrublands or forests are predicted with higher confidence. These visualizations demonstrate the role of uncertainty in refining the interpretation of the appearance of naturalness and in identifying areas where additional validation or targeted data collection may be beneficial [37]. Additional qualitative examples are presented in the Appendix chapter.

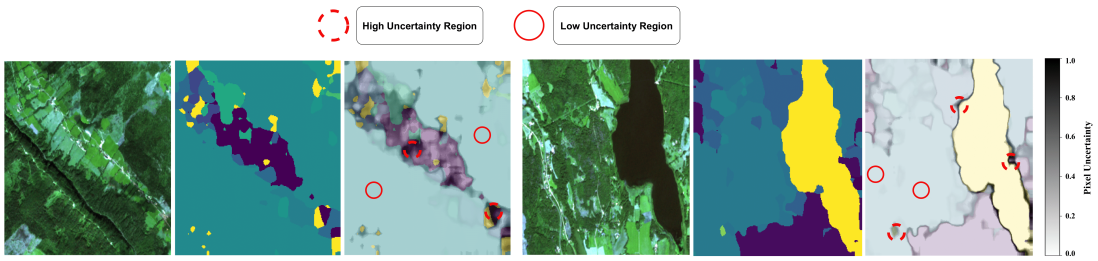


Figure 6.4: Uncertainty-aware segmentation maps. Two examples showing Sentinel-2 RGB images, predicted segmentation masks, and grayscale overlaid uncertainty maps. Brighter pixels indicate higher uncertainty. Segmentation colors correspond to different land cover patterns. Dashed and solid red circles show the areas with the lowest and highest confidence, respectively.

6.3.2 Quantitative results

The Confident Naturalness Explanation (CNE) metric was computed for each land cover pattern by combining attribution strength with class-specific uncertainty [4]. Table 6.1 reports the resulting CNE values together with the relative distribution of each pattern in the data.

High CNE values for patterns such as Shrublands, Moors and Heathland, and Peat Bogs in the AnthroProtect dataset indicate that these classes are both strongly associated with naturalness and predicted with high confidence. Analogous classes in MapInWild, such as Shrublands and Wetlands, also show

Table 6.1: CNE metric scores and datasets distribution in the regions representing naturalness. The table presents a comparative analysis across two datasets: the top section details insights derived from the AnthroProtect dataset [12], while the bottom section displays results from the MapInWild dataset [33].

Pattern	**CNE Metric**	**Distribution (%)**
Transitional Woodland Shrub	0.91	10.12
Moors and Heathland	0.92	13.20
Peat Bogs	0.81	6.12
Bare Rock	0.65	6.81
Broad-leaved Forest	0.61	13.40
Sparsely Vegetated Areas	0.49	24.11
Coniferous Forest	0.44	18.23
Watercourses	0.23	0.22
Glaciers and Perpetual Snow	0.21	1.11
Natural Grassland	0.19	0.05
Water Bodies	0.18	5.23
Shrubland	0.90	23.32
Wetlands	0.79	6.12
Tree Cover	0.55	31.63
Bare / Sparse Vegetation	0.60	30.92
Grassland	0.30	0.05
Open Water	0.21	5.45
Snow and Ice	0.21	1.11

consistently high scores. In contrast, patterns such as Water Bodies and Grassland exhibit low CNE values across both datasets.

6.4 Discussion

Shrublands achieved the highest CNE value, indicating a particularly strong and confident association with naturalness. Alongside Shrublands, wetland-related patterns such as Moors and Heathland and Peat Bogs also contribute strongly, with high CNE scores and low associated uncertainty. These results align with the ecological importance of these land cover types, which often function as biodiversity hotspots, carbon sinks, and relatively undisturbed habitats [10, 13, 16]. In contrast, patterns such as Water Bodies and Natural Grassland display lower confidence-weighted contributions, which may be due to greater ambiguity in their relationship to naturalness or challenges in segmentation caused by spectral variability, seasonal effects, or underrepresentation in the training data.

These findings illustrate the ability of the CNE framework to distinguish between confidently relevant and uncertain land cover patterns. This differentiation enables the identification of classes or regions that may require further validation or targeted data augmentation and provides a more reliable basis for conservation strategies by clarifying which land cover types consistently indicate naturalness [4].

The framework has several limitations. The use of logistic regression as a global surrogate, while interpretable, may oversimplify nonlinear interactions between land cover patterns. Additionally, the framework uses MC-Dropout, which requires multiple forward runs to produce the uncertainty-aware segmentation maps, which can introduce a bottleneck for scalability.

6.5 Conclusion

This chapter presented the Confident Naturalness Explanation (CNE) framework, which integrates XAI with UQ to assess the appearance of naturalness in satellite imagery. The framework builds on a DeepLabV3 segmentation model [143] to derive land cover patterns that are vectorized into a simpler form as an input for a logistic regression model that works as a surrogate model for the interpretation of the semantic segmentation model [24, 20]. The surrogate provides class-level contributions to naturalness, while predictive uncertainty is estimated using MC-Dropout [6, 145].

The combination of the land cover's contributions and UQ yields the CNE score, a confidence-weighted relevance measure for each land cover class.

High CNE values indicate strong and reliable contributions to naturalness, whereas low values reflect weaker or uncertain contributions. Results from the AnthroProtect [107] and MapInWild [33] datasets showed that shrublands, moors, heathland, and peat bogs consistently exhibit high CNE scores, showing their ecological relevance to naturalness. In contrast, water bodies and natural grassland displayed lower scores, indicating greater uncertainty or smaller contributions.

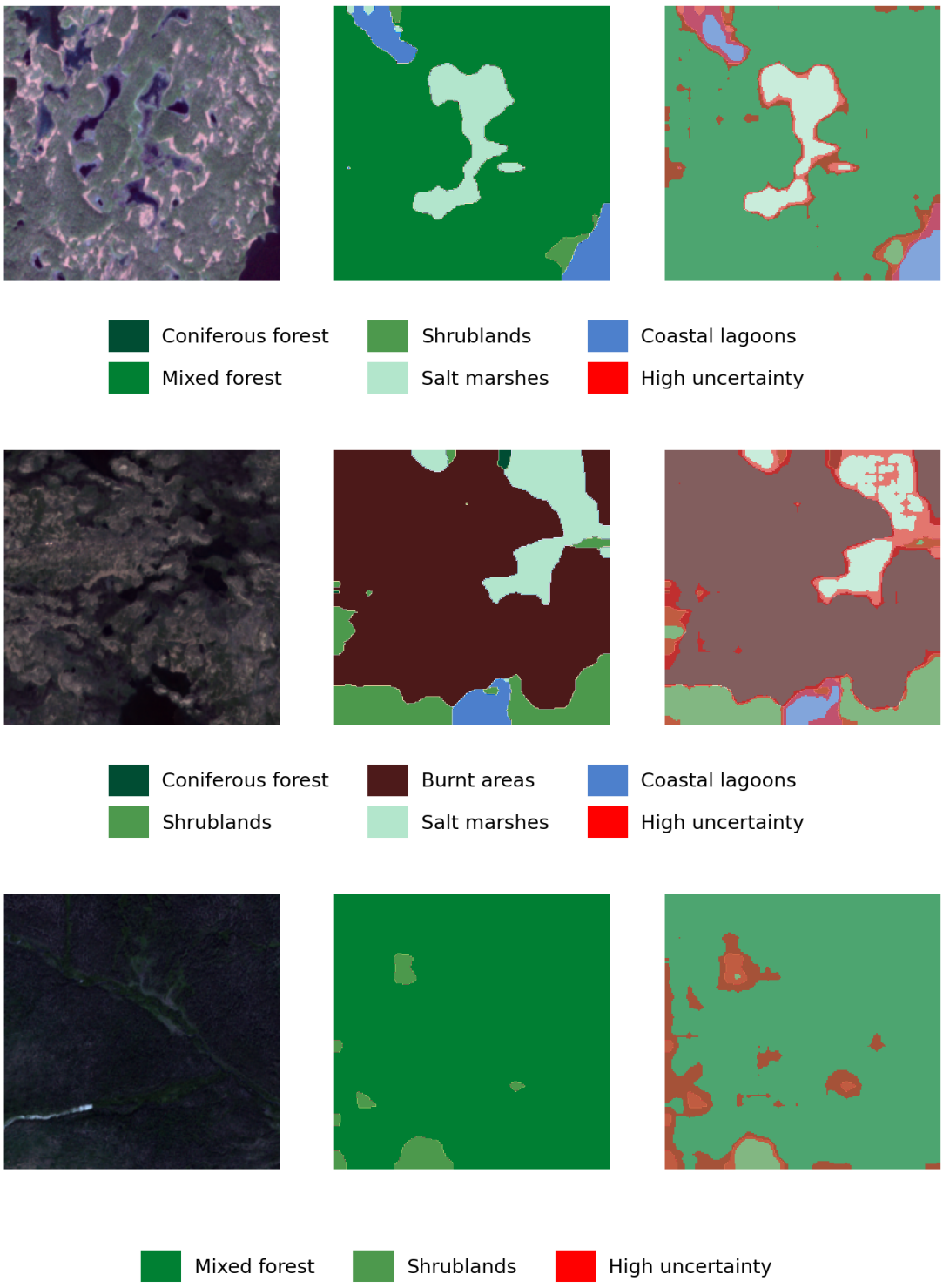


Figure 6.5: Example visualizations of qualitative results from the CNE framework (Part 1 of 3). Each row presents three images: on the left, the original Sentinel-2 image from areas of varying naturalness; in the middle, the average segmentation map obtained from multiple MC-Dropout runs; and on the right, the corresponding uncertainty-aware segmentation map, where red intensity represents the magnitude of predictive uncertainty for each pixel.

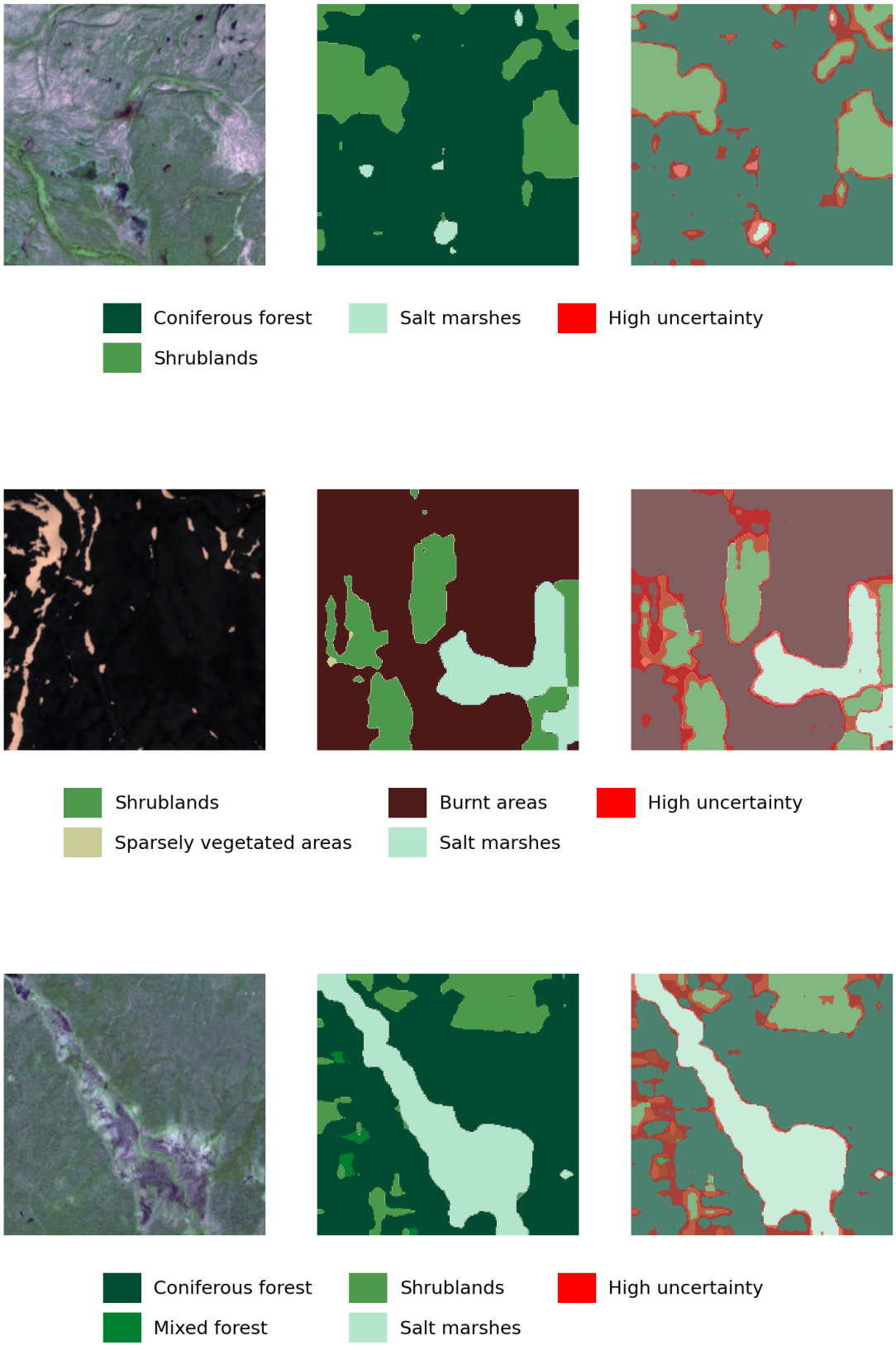


Figure 6.6: Example visualizations of qualitative results from the CNE framework (Part 2 of 3). Each row presents three images: on the left, the original Sentinel-2 image from areas of varying naturalness; in the middle, the average segmentation map obtained from multiple MC-Dropout runs; and on the right, the corresponding uncertainty-aware segmentation map, where red intensity represents the magnitude of predictive uncertainty for each pixel.

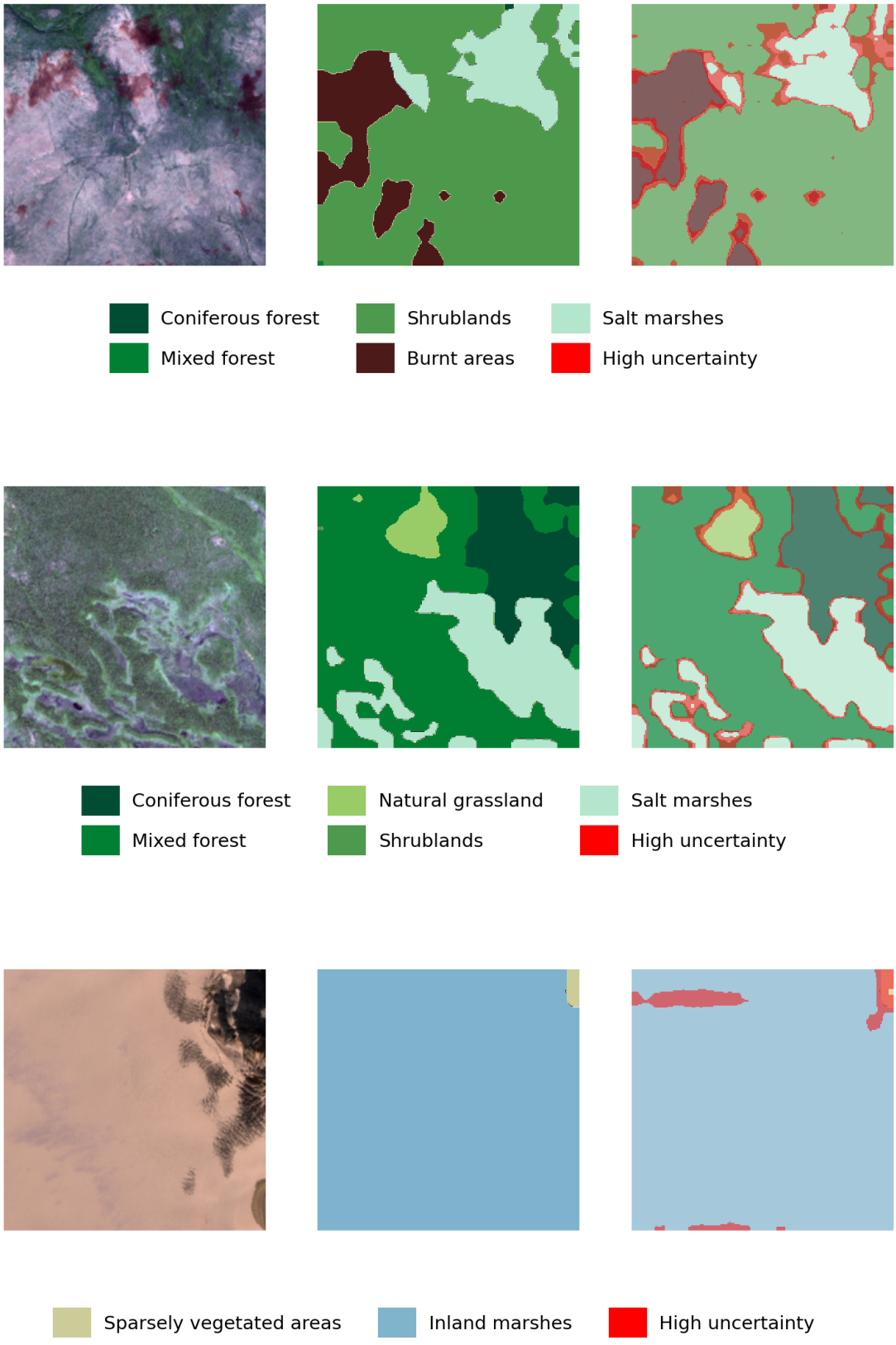


Figure 6.7: Example visualizations of qualitative results from the CNE framework (Part 3 of 3). Each row presents three images: on the left, the original Sentinel-2 image from areas of varying naturalness; in the middle, the average segmentation map obtained from multiple MC-Dropout runs; and on the right, the corresponding uncertainty-aware segmentation map, where red intensity represents the magnitude of predictive uncertainty for each pixel.

Chapter 7

Naturalness Assessment with Transformer-Based Reliable Explainability

In the previous two chapters, we introduced two frameworks to explain naturalness in a data-driven manner. The first, AM-GANs for Naturalness framework [1], focused on XAI but did not address the associated uncertainty in the model's decisions. The second, the Confident Naturalness Explanation (CNE) framework [4], integrated UQ with XAI but required vectorizing segmentation maps into one-dimensional representations, thereby losing the spatial structure of satellite imagery.

In this chapter, we present the NaT-ReX framework [7], which overcomes these limitations and extends the analysis to transformer-based architectures. Vision Transformers (ViTs) [8] have been extensively adopted in classical computer vision tasks, such as image classification, regression, semantic segmentation, and object detection. ViTs have also emerged as the state-of-the-art backbone for Earth observation foundation models [146, 147].

NaT-ReX combines pixel-wise explainability using Layer-wise Relevance Propagation (LRP) attention rollout [9] with UQ via MC-Dropout [6] in a transformer-based model. This yields the ReX score, a spatially fine-grained uncertainty-aware naturalness relevance index.

This chapter builds on the research gaps identified in the related work chapter 3, advancing a strategy for explanations of the appearance of naturalness in satellite imagery. We extend and refine the frameworks introduced in the previous two chapters to address their limitations, presenting the NaT-ReX architecture together with its training and inference strategies and experimental setup. The proposed approach is evaluated qualitatively and quantitatively on the AnthroProtect [107] and MapInWild [33] datasets, and in chapter 8, we compare

NaT-ReX to established naturalness assessment indices to examine the alignment of its results with existing frameworks.

To the best of my knowledge, NaT-ReX is the first framework to combine LRP attention rollout [9] with MC-Dropout [6], enabling class-specific, pixel-level explanations with the associated model's uncertainty estimates.

The key contributions of this chapter are:

1. Propose the NaT-ReX framework and the ReX score for uncertainty-aware naturalness assessment.
2. Provide pixel-wise and class-level explanations of the appearance of naturalness by combining relevance attribution with model uncertainty.
3. Explain naturalness through qualitative and quantitative results on the uncertainty-aware relevance of land cover classes.

7.1 State of the Art

7.1.1 Explainability Approaches for ViT-based architectures

Methods to explain Vision Transformers (ViTs) [8] have evolved from simple visualizations of raw attention to more advanced attribution techniques specifically adapted to transformer architectures. The mathematical foundations of ViTs, including their attention mechanisms, were described in detail in the basic technique chapter 2.4.

A first line of work focuses on raw attention maps. Since ViTs split an image into patches and process them as tokens, the attention weights can be visualized to show how strongly each patch attends to the classification token or other patches. While informative, such visualizations are not always interpretable and tend to become less clear in deeper layers, capturing broad contextual interactions rather than clear indicators of feature importance [8].

Attention rollout, introduced by Abnar and Zuidema [148], improves on this by composing attention across layers. Attention heads are averaged, residual connections are incorporated by adding the identity matrix, and the resulting matrices are recursively multiplied across layers to aggregate patch-to-input relevance. A related method, attention flow, formulates the same problem as a flow graph and applies max-flow algorithms to trace information paths [148]. Both approaches provide more structured token-to-input attributions than raw attention. However, these methods remain class-agnostic: they highlight features the model considers relevant in general, but not specifically for a given class, which negatively affects the interpretability for a specific class.

In parallel, intrinsically interpretable models have been proposed, such as eX-ViT [149], which integrates explainability into the architecture through explainable multi-head attention and an attribute-guided explainer. These models aim to provide explanations as part of the prediction process rather than relying purely on post-hoc methods. However, such approaches are constrained by their reliance on a specific architectural design and therefore cannot be readily transferred across different model architectures, making them model-specific solutions.

In the context of remote sensing, attention rollout and related attribution methods have been applied to land cover classification, scene understanding, and change detection, providing spatially resolved explanations in Earth observation models [150]. Recent surveys highlight the growing role of explainability in Earth observation, where fine-grained insights are crucial for trustworthy analysis [151, 20].

Chefer et al. [9] extended these efforts by proposing a transformer-specific adaptation of Layer-wise Relevance Propagation (LRP), often referred to as LRP attention rollout. Their method propagates relevance conditioned on a target class by weighting attention with input gradients and redistributing relevance through attention and residual connections. In contrast to class-agnostic rollout or raw attention, the resulting relevance maps are class-specific and highlight only those input patterns that contribute to the prediction of the selected class. This distinction is crucial for naturalness assessment, since we are not interested in all image features deemed relevant by the model, but specifically in those that explain the prediction of the naturalness class. Moreover, the approach is not limited to a single model design and can be applied across different transformer-based architectures, making it more broadly applicable than architecture-specific explainability methods. Table 7.1 summarizes the main approaches of explanation methods for Vision Transformers.

7.1.2 Quantifying Uncertainty for ViT-based architectures

UQ methods for ViTs [8] encompass several approaches. Bayesian and stochastic techniques, exemplified by Rad et al.'s PoViT-UQ [152], use MC-Dropout [6] to approximate Bayesian posterior uncertainty, in the context of geological subsurface simulation with uncertainty awareness. Lopez et al. [153] enhance medical image segmentation by employing stochastic ViT encoders with Gaussian embeddings and an uncertainty-aware regularization, explicitly modeling predictive uncertainty in a probabilistic embedding framework. Deterministic methods, such as Zhao et al.'s work [154], introduce innovative self-attention

Table 7.1: Overview of explanation methods for Vision Transformers [8].

Method	Mechanism	Specificity of the attribution maps
Raw attentions [8]	Visualize attention weights from patches to tokens	Class-agnostic
Attention rollout [148]	Propagate attention across layers using matrix multiplications or flow algorithms	Class-agnostic
LRP attributions [9]	Propagate class-conditioned relevance scores through attention and residual blocks	Class-aware
eX-ViT [149]	Intrinsically interpretable architecture with explainable attention and attribute-guided explanations	Class-aware

mechanisms that replace dot-product similarity with Banach space distance combined with Gaussian Process output layers, providing reliable uncertainty estimates without the need for multiple stochastic inferences. Additionally, the Confidence-Filtered Relevance (CFR) framework [28] used the Deep Deterministic Uncertainty estimation (DDU) [155] to provide uncertainty-aware explanations in remote sensing tasks, enhancing the interpretability of Vision Transformer models under uncertainty.

Ensemble approaches like Wang et al.'s Masksemble-aided Cross-ViT [156] leverage model diversity to improve uncertainty estimation for skin cancer diagnosis, enhancing predictive confidence through ensemble aggregation.

Sayer et al. [157] emphasized the critical need for rigorous uncertainty handling in pixel-level aerosol retrievals. They provided a comprehensive review of existing uncertainty estimation methods and proposed a robust evaluation framework aimed at improving the reliability and validation of pixel-level uncertainty estimates in satellite aerosol optical depth products. Given the central focus of this thesis on explainability, the discussion of uncertainty quantification techniques for ViT is therefore kept concise to maintain relevance and clarity.

7.1.3 Integration of XAI and UQ techniques

The recently published work of Essenfelder et al. [29] integrates UQ with XAI to enhance transparency in climate hazard detection. An ensemble of XGBoost classifiers [158] produces probabilistic outputs for multiple hazards, where ensemble disagreement and probability spread serve as measures of predictive uncertainty. This allows the distinction between high-confidence predictions (low spread, high agreement) and uncertain regions where models disagree.

XAI is incorporated through SHAP-based feature attributions [59], providing local and global insights into the drivers behind predictions. These explanations enable experts to verify whether model reasoning is consistent with established domain knowledge. By jointly presenting prediction probabilities, uncertainty measures, and explanatory drivers, the framework ensures that outputs are interpretable and auditable, thereby fostering more informed and trustworthy decision-making in operational early-warning contexts.

This approach shows conceptual overlap with our proposed CNE framework, discussed in the previous chapter, as both emphasize combining UQ and XAI for transparent and actionable outputs.

7.2 Methodology

In this section, we detail how the approach works and outline its main components: a classification head for relevance mapping, a reconstruction head for uncertainty estimation, and the combined ReX score.

7.2.1 NaT-ReX Architecture and Training Strategy

NaT-ReX is built on a Vision Transformer (ViT-B/16) encoder [8] with two output heads trained jointly in a multitask setup: a classification head for predicting the class naturalness versus non-naturalness, and a reconstruction head for image reconstruction. The transformer encoder captures long-range dependencies and global context, essential for scene-level naturalness assessment. A simplified version of the NaT-ReX architecture is shown in Figure 7.1.

The classification head is a fully connected layer that predicts the class image. During inference, we apply Layer-wise Relevance Propagation attention rollout [9] to the classification head to identify the regions of the image that are most relevant to the prediction. The reconstruction head transforms the encoder's output back into the original image layout. Tokens are reshaped into image patches and passed through three convolutional layers with ReLU activations [43, 40].

Training is performed end-to-end using a combined loss: a cross-entropy loss for the classification head [43] and a mean squared error loss for the reconstruction head [40]. This multitask formulation [159] encourages the encoder to learn spatial features that support both attribution and uncertainty quantification [6, 80] in the inference stage.

Classification Head in inference: Relevance to Naturalness via LRP Attention Rollout

LRP attention rollout [9] propagates relevance scores from the classification head back through the attention layers. Figure 7.3 highlights the image regions contributing to both the naturalness and non-naturalness classes.

Let $A^{(b)}$ denote the attention matrix in transformer block b, and $R^{(n_b)}$ the relevance at that layer's output. The adjusted attention map is defined as:

$$\bar{\mathsf{A}}^{(b)} = \mathsf{I} + \frac{1}{H} \sum_{h=1}^{H} \max\left(\nabla \mathsf{A}_h^{(b)} \odot \mathsf{R}_h^{(n_b)}, 0\right) \tag{7.1}$$

Equation 7.1 updates the attention in block b by averaging the gradient-based relevance across all attention heads, keeping only positive contributions. The

identity matrix I ensures residual connections are preserved, allowing each token to retain part of its relevance.

The final relevance map is computed as:

$$\mathsf{M} = \bar{\mathsf{A}}^{(1)} \cdot \bar{\mathsf{A}}^{(2)} \cdot \ldots \cdot \bar{\mathsf{A}}^{(B)} \tag{7.2}$$

Equation 7.2 accumulates relevance across all transformer blocks, effectively tracing how importance propagates from the model's output back to the input tokens. For a detailed derivation, we refer the reader to the basic techniques chapter 2.4.

7.2.2 Reconstruction Head in inference: Uncertainty Quantification via MC-Dropout

To estimate epistemic uncertainty, MC-Dropout [6] is used in the reconstruction head. During inference, J stochastic forward passes are performed with dropout enabled, producing a set of slightly varied reconstructions.

Let $\mathsf{Y}_{h,w,c,j}$ be the reconstructed pixel value at position (h, w, c) from the j^{th} forward pass. The pixel-wise mean is:

$$\bar{\mathsf{Y}}_{h,w,c} = \frac{1}{J} \sum_{j=1}^{J} \mathsf{Y}_{h,w,c,j} \tag{7.3}$$

And the corresponding uncertainty estimate is the standard deviation:

$$\mathsf{S}_{h,w,c} = \sqrt{\frac{1}{J} \sum_{j=1}^{J} \left(\mathsf{Y}_{h,w,c,j} - \bar{\mathsf{Y}}_{h,w,c}\right)^2} \tag{7.4}$$

The resulting uncertainty map $\mathsf{S} \in \mathbb{R}^{H \times W \times C}$ reflects the model's epistemic uncertainty in reconstructing each pixel, with higher values indicating lower confidence.

7.2.3 ReX Score: Combining Naturalness, Relevance, and Uncertainty

The ReX score R_n measures how confidently a pixel n contributes to the class `naturalness`, combining its relevance and associated uncertainty:

$$R_n = \left[\mathsf{M}_{\text{norm},n} \log\left(1 + \alpha \cdot \frac{1}{\mathsf{S}_{\text{norm},n} + c}\right)\right]^{\beta} \tag{7.5}$$

Here, $\mathsf{M}_{\text{norm},n}$ denotes the normalized LRP relevance for patch n, and $\mathsf{S}_{\text{norm},n}$ represents the normalized pixel-wise uncertainty. The relevance map M_{norm} is

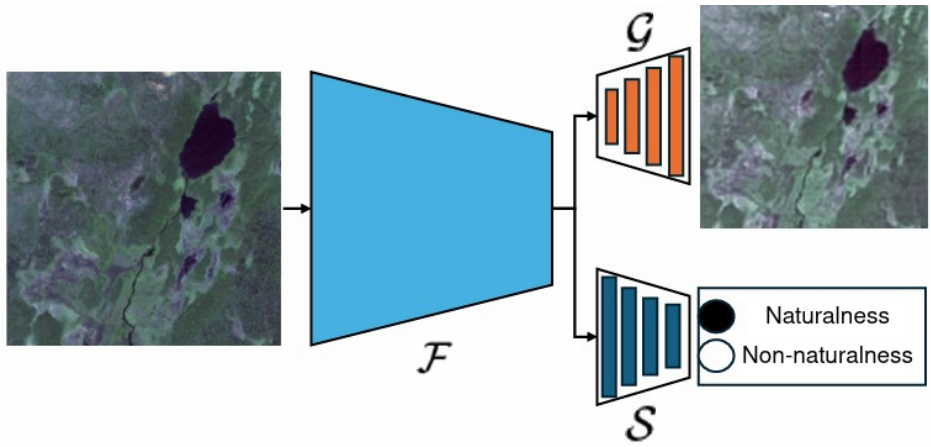


Figure 7.1: Training architecture of the NaT-ReX framework. The input satellite image is processed by a Vision Transformer-based encoder $\mathcal{F}$, which feeds into two parallel heads: a reconstruction head $\mathcal{G}$ that learns to reconstruct the input image, and a classification head $\mathcal{S}$ that predicts naturalness versus non-naturalness. Both heads are trained simultaneously in a multi-task learning setup to optimize both classification and reconstruction objectives.

normalized using min-max scaling based on the minimum and maximum relevance values across the image. The logarithmic term penalizes high uncertainty, ensuring that only patches with both high relevance and low uncertainty receive high ReX scores. The parameters are set as follows: $\alpha = 0.5$ (controls the impact of uncertainty), $c = 0.2$ (prevents division by near-zero values by setting a lower bound), and $\beta = 0.8$ (adjusts the sharpness of the score distribution). These hyperparameters were chosen heuristically, yielding good separation of ReX scores across different land cover classes while avoiding extremely large or small values for those contributing to a specific class, such as naturalness.

This formulation results in pixel-wise uncertainty-aware contribution to naturalness, improving the robustness and interpretability of naturalness assessments. Figure 7.2 illustrates the inference pipeline and score calculation.

7.2.4 Experimental Setup

The NaT-ReX framework uses a multitask learning strategy. The classification head was trained for 10 epochs and achieved 100% accuracy on the training dataset and 98% on the test dataset. The reconstruction head was trained for 30 epochs with a batch size of 16, reaching a best reconstruction loss of 0.129 and 0.133 for the training and test datasets, respectively. During the inference phase, 25 stochastic runs are used through the reconstruction head to produce the uncertainty maps. Overall training was performed with a batch size of 64 using the AdamW optimizer enhanced with momentum [160]. The model training was accelerated using an NVIDIA A100 GPU.

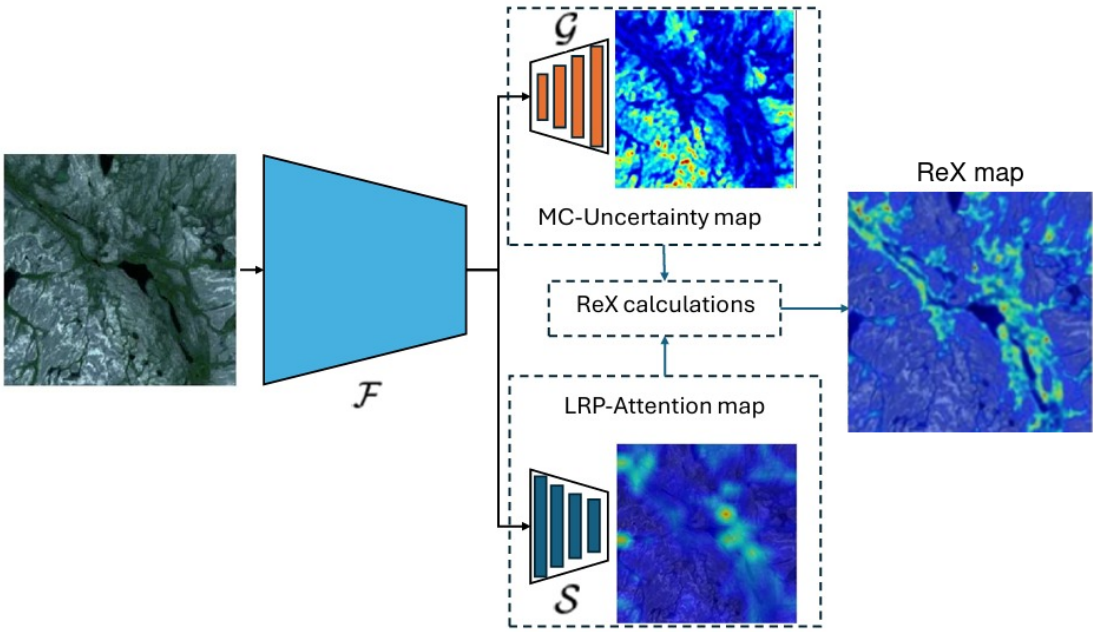


Figure 7.2: Inference pipeline of the NaT-ReX framework. The input image is passed through the shared encoder $\mathcal{F}$, producing two outputs: an uncertainty map from the reconstruction head $\mathcal{G}$ using MC-Dropout, and an LRP attention map from the classification head $\mathcal{S}$ via attention rollout. These two outputs are integrated per pixel using the ReX formulation to generate the final ReX map, which highlights spatial contributions to the class naturalness while accounting for uncertainty.

7.3 Results

7.3.1 Qualitative Evaluation

The qualitative evaluation focuses on the interpretability of the NaT-ReX predictions, emphasizing how the model identifies and localizes naturalness-relevant features while accounting for uncertainty.

The top subplot in Figure 7.3 presents the original input images, followed by class-specific relevance maps (using LRP attention rollout), uncertainty maps (from MC-Dropout), and the resulting ReX score maps. These ReX maps highlight areas with both high relevance and low uncertainty, aligning with regions considered natural in the input images. The bottom subplot shows CutMix experiments, where images blending natural and urban regions are analyzed to test the consistency of relevance scores. Columns 2 and 4 visualize the LRP relevance maps for `non-naturalness` and `naturalness`, respectively. As expected, croplands and urban areas exhibit high relevance for the `non-naturalness` class, whereas vegetated regions and wetlands show strong relevance for `naturalness`. The ReX maps further emphasize regions where the model is confident in its naturalness predictions. Figure 7.5 shows more results in different geographical regions

Figures 10.7 in the appendix present additional qualitative examples, illustrating the original input, relevance, uncertainty, and ReX score maps across various land cover types. Figure 10.5 provides further examples, showing how relevance is distributed in CutMix images—highlighting that urban areas contribute to *non-naturalness*, whereas protected areas support the classification of *naturalness*.

7.3.2 Quantitative Evaluation

Table 7.2 provides a detailed overview of the relevance scores [9], uncertainty estimates, and the resulting ReX scores for each land cover class in both AnthroProtect [12] and MapInWild [33] datasets. These results validate the robustness of NaT-ReX across different regions and classification systems, showing consistently high scores for classes such as transitional woodland-shrub, shrubland, peat bogs, and broad-leaved forest, and low scores for snow and ice, open water, and bare rock.

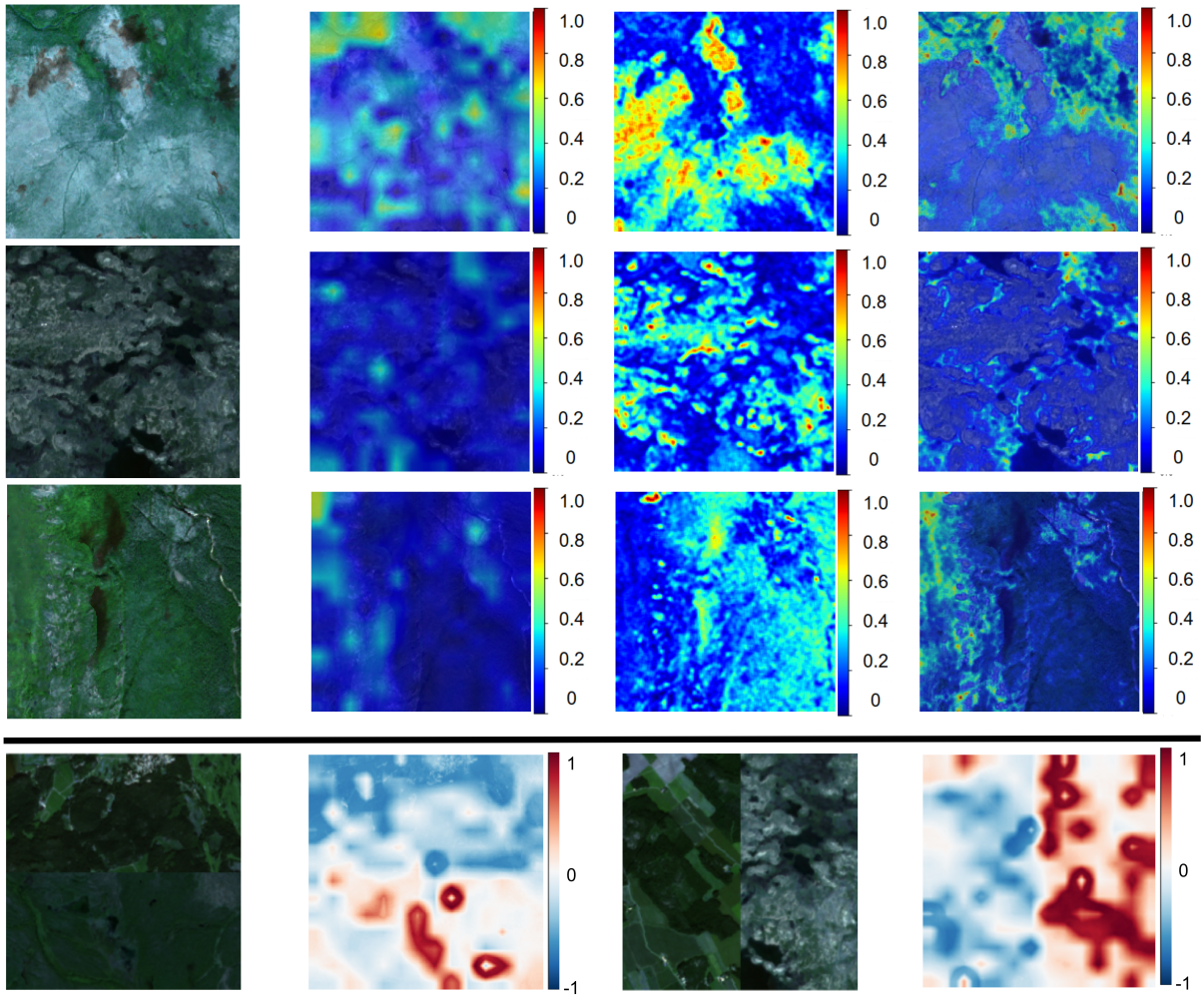


Figure 7.3: top: NaT-ReX visualizations showing the original input, LRP attention maps for *naturalness*, MC-Dropout uncertainty maps, and resulting ReX maps. bottom: CutMix visualizations for evaluating class-specific relevance under mixed patch compositions. Columns 1 and 3 show CutMix images blending natural and urban regions. Columns 2 and 4 show LRP attention maps for non-naturalness (negative relevance) and naturalness (positive relevance), respectively. Additional examples of original images, attention maps, uncertainty estimates, and resulting ReX maps are shown in the appendix.

Table 7.2: Mean relevance scores [9] and MC-Dropout uncertainties [6] for each land cover class in the AnthroProtect and MapInWild datasets. The final ReX score reflects both semantic relevance (via LRP attention rollout) and the model's uncertainty.

Class Name	**Relv. μ**	**Relv. σ**	**Unc μ**	**ReX Score**
AnthroProtect Dataset				
transitional woodland-shrub	0.915	0.072	0.029	1.000
moors and heathland	0.802	0.073	0.000	0.949
peat bogs	0.855	0.070	0.043	0.919
broad-leaved forest	0.892	0.071	0.105	0.853
mixed forest	0.886	0.071	0.173	0.763
coniferous forest	0.853	0.069	0.187	0.723
inland marshes	0.692	0.069	0.120	0.659
natural grassland	0.627	0.082	0.114	0.607
water courses	0.742	0.068	0.236	0.595
sparsely vegetated areas	0.563	0.071	0.171	0.510
beaches, dunes, and sand plains	0.501	0.054	0.183	0.438
water bodies	0.487	0.065	0.295	0.369
bare rock	0.515	0.072	0.351	0.367
glaciers and perpetual snow	0.578	0.049	1.000	0.259
sea and ocean	0.121	0.031	0.299	0.000
MapInWild Dataset				
Shrubland	0.903	0.063	0.077	1.000
Trees	0.474	0.069	0.000	0.630
Bare/Sparse vegetation	0.631	0.066	0.677	0.370
Grassland	0.451	0.068	0.319	0.360
Herbaceous wetland	0.220	0.054	0.127	0.200
Snow and ice	0.277	0.051	1.000	0.120
Open water	0.101	0.052	0.488	0.000

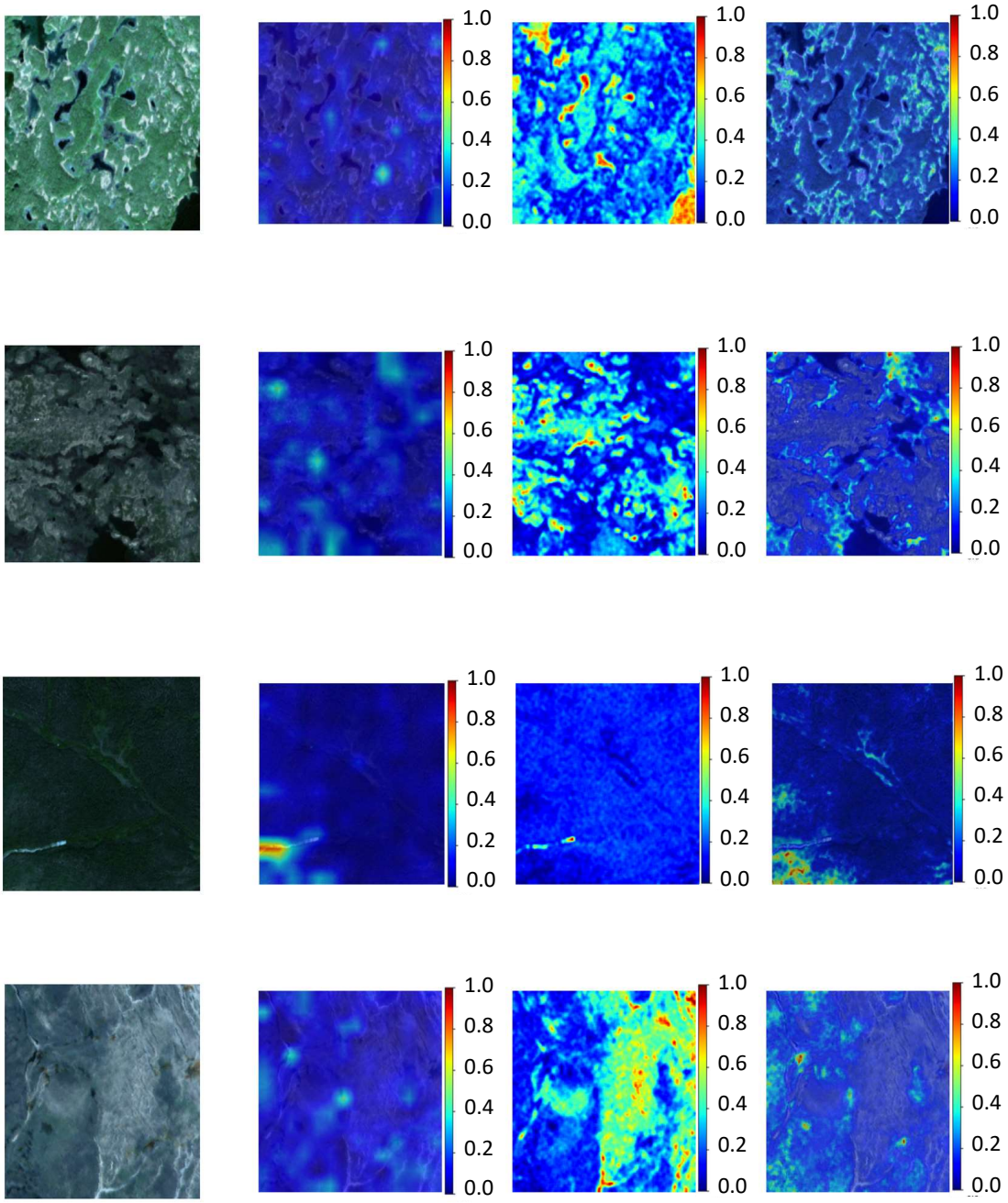


Figure 7.4: Additional qualitative examples (part 1 out of 2). Each row shows: original Sentinel-2 image, LRP attention map for naturalness, MC-Dropout uncertainty map, and ReX score map. These support the spatial interpretation of confidence-weighted naturalness across land cover types. The caption applies to both this figure and the previous one.

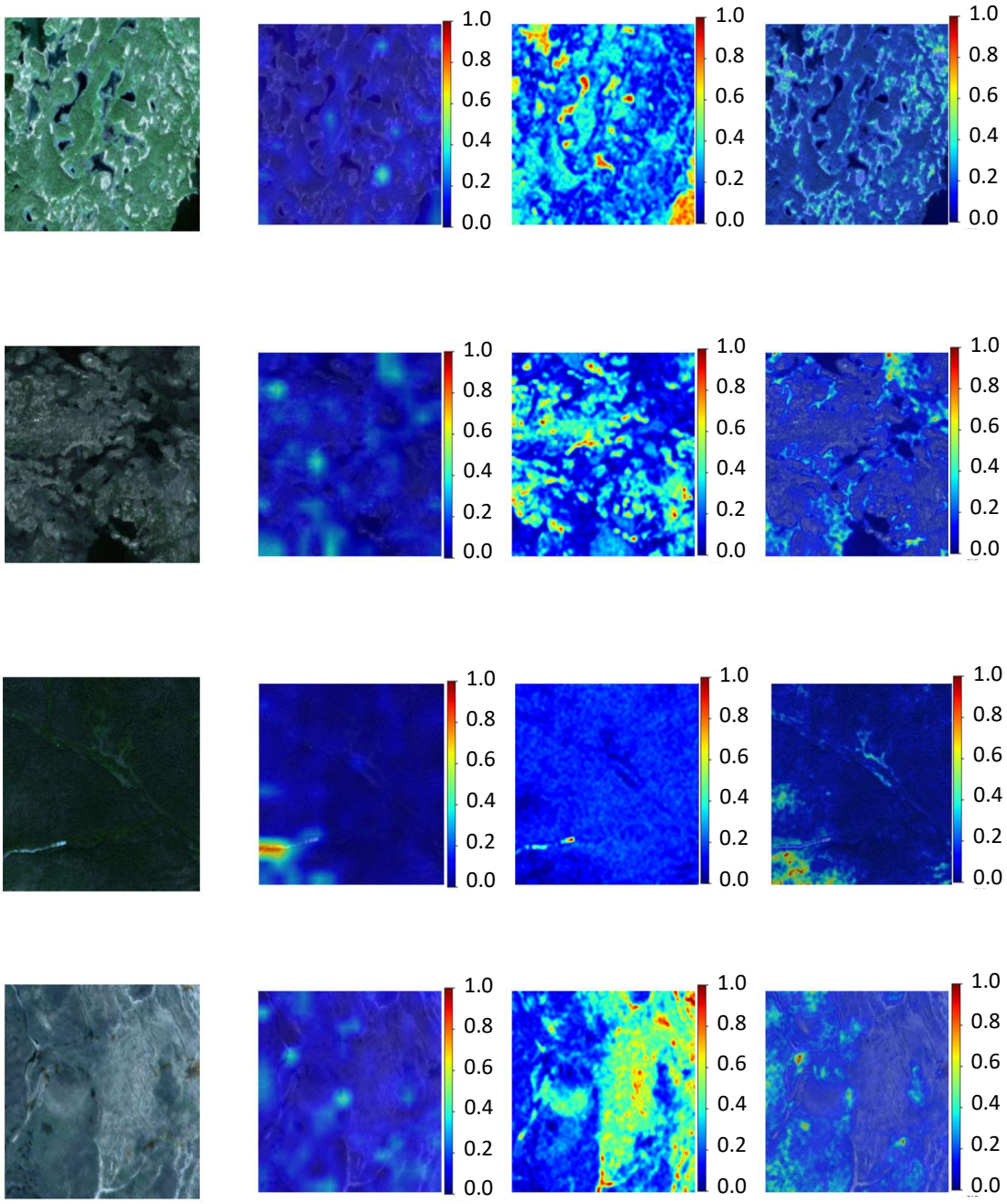


Figure 7.5: Additional qualitative examples (part 2 out of 2). Each row shows: original Sentinel-2 image, LRP attention map for naturalness, MC-Dropout uncertainty map, and ReX score map. These support the spatial interpretation of confidence-weighted naturalness across land cover types. The caption applies to both this figure and the previous one.

Additional qualitative visualizations are provided in the Appendix to support deeper visual interpretation of the results.

7.4 Discussion

The results further illustrate naturalness as follows: in the AnthroProtect dataset, highly structured natural classes such as transitional woodland-shrub, peat bogs, and broad-leaved forest achieved high ReX scores (1.00, 0.92, and 0.85, respectively), consistent with their strong relevance and relatively low uncertainty. In contrast, classes with inherently higher variability or extremely high or low signals, such as glaciers and perpetual snow or open waters, showed low ReX scores (0.26 and 0.0), reflecting both elevated uncertainty and weaker relevance signals. Similarly, in the MapInWild dataset, shrubland yielded the highest ReX score (1.00), while snow and ice, as well as open water, received very low scores (0.12 and 0.00), again demonstrating sensitivity to uncertainty in areas with low to no distinctive features [28]. A more detailed discussion of the implications and interpretive meaning of these quantitative values is provided in the next chapter.

A key limitation of NaT-ReX lies in its current approach to uncertainty estimation. By relying solely on MC-Dropout [6], the framework approximates only epistemic uncertainty, reflecting limitations in model knowledge and representation capacity. However, in satellite imagery and ecological applications, aleatoric uncertainty—stemming from sensor noise, atmospheric effects, and natural variability—is equally critical [145]. Ignoring this component may lead to overconfident predictions, especially in heterogeneous or dynamically changing landscapes. Furthermore, MC-Dropout requires multiple forward passes at test time, substantially increasing inference cost compared to deterministic models, which poses challenges for large-scale or real-time monitoring systems.

7.5 Conclusion

In this chapter, we introduced NaT-ReX, a transformer-based framework designed to quantify naturalness in satellite imagery through the integration of explainability and uncertainty estimation. Central to the framework is the ReX score, which fuses relevance maps from Layer-wise Relevance Propagation [9] with pixel-level uncertainty maps from MC-Dropout [6]. This yields spatially explicit, confidence-weighted naturalness predictions that are robust and interpretable. Unlike traditional indices based on predefined assumptions [10, 11], NaT-ReX learns semantic relevance directly from the data while preserving spatial structure.

The results on the AnthroProtect [12] and MapInWild [33] datasets highlight consistent naturalness patterns across different ecosystems. Classes such as transitional woodland-shrub, shrubland, and peat bogs rank highest in naturalness, whereas snow and ice, open water, and bare rock receive lower ReX scores due to higher uncertainty or lower semantic relevance. These findings align well with expert-based assessments [10, 11, 16] and confirm the potential of the framework for conservation applications.

By combining spatial relevance and uncertainty, NaT-ReX moves beyond pixel-wise classification toward a more nuanced, data-driven evaluation of naturalness. This can support transparent monitoring and guide conservation priorities. Future work should address scalability, explore richer uncertainty modeling [80, 155], and incorporate more diverse ecological datasets.

Chapter 8

Quantitative and Ecological Analysis of Patterns Forming Naturalness in Fennoscandia

The frameworks developed in this dissertation follow a fully data-driven approach, without relying on predefined assumptions about naturalness. Since no absolute ground truth exists for quantifying naturalness, NI [11] and HII [10] are employed as external benchmarks rather than target values. Positive correlations between the framework outputs and these indices indicate alignment with established ecological assessments, supporting the broader validity of the proposed approach.

8.1 Comparing Frameworks to Naturalness Indices

To investigate the appearance of naturalness in satellite imagery, we compare the outputs of three developed frameworks—AM-GANs for Naturalness [1], CNE [4], and NaT-ReX [7]. The analysis focuses on the Fennoscandian region, which represents the overlapping area between the AnthroProtect and MapInWild datasets. As naturalness is geographically dependent and varies across biomes, restricting the evaluation to this single ecological context enables controlled and meaningful comparisons. Only the test dataset is used for evaluation to ensure consistent cross-framework assessment on ecologically coherent imagery.

For comparability of land cover classes' contribution to naturalness across different indices, we adopt the ESA WorldCover land cover higher hierarchy [93]. This choice is motivated by its close alignment with the coarse land cover categories introduced in the HII [10], enabling semantically meaningful comparisons between model-derived naturalness scores and established measures of ecological impact. Since the HII [10] was originally designed to assess patterns

of urbanization and anthropogenic influence, we use its inverted form in this chapter, where a value of 0 denotes low contribution to naturalness and a value of 1 indicates the highest contribution to naturalness. In the following evaluations, we compare the positive contributors to naturalness from the three proposed frameworks, after normalization to the range [0, 1], with the contributors to naturalness in the NI [11] and inverted HII [10], also normalized to the same range. A value of 0 in these comparisons represents the least relative contribution to naturalness compared to other land cover classes present in the comparison and should not be interpreted as a complete or absolute lack of contribution to naturalness.

8.1.1 Alignment of AM-GANs Scores with NI and HII

Table 8.1 presents the normalized scores alongside the Naturalness Index (NI) [11] and the inverted Human Influence Index (HII) [10]. The comparison indicates strong consistency across several land cover classes, while notable deviations appear in others.

Wetlands achieve the highest AM-GANs score (1.0), closely matching NI (1.0) and HII (1.0). Bare or sparsely vegetated areas also show relatively consistent values across all three measures, indicating that the framework can capture naturalness in ecologically distinct and less heterogeneous classes.

By contrast, shrubland and trees are underestimated by AM-GANs (0.0 and 0.218) compared to NI (0.952 and 0.698) and HII (0.91 and 0.672). Grassland, snow, and ice fall in between, showing weaker alignment. These deviations arise because AM-GANs do not downweight land cover classes with higher uncertainty, which can lead to discrepancies with established indices as well as with the proposed frameworks [4, 7].

Figure 8.1 illustrates these patterns: strong alignment for wetlands, moderate consistency for bare areas, and notable divergence in vegetation-dominated classes.

Table 8.1: Comparison of AM-GANs naturalness scores with the Naturalness Index (NI) and reversed Human Influence Index (HII) across land cover classes.

Land Cover Class	**AM-GANs**	**NI**	**HII**
Shrubland	0.000	0.952	0.910
Trees	0.218	0.698	0.672
Grassland	0.175	0.524	0.493
Bare/Sparse Vegetation	0.292	0.206	0.149
Wetland	1.000	1.000	1.000
Snow and Ice	0.087	0.000	0.000

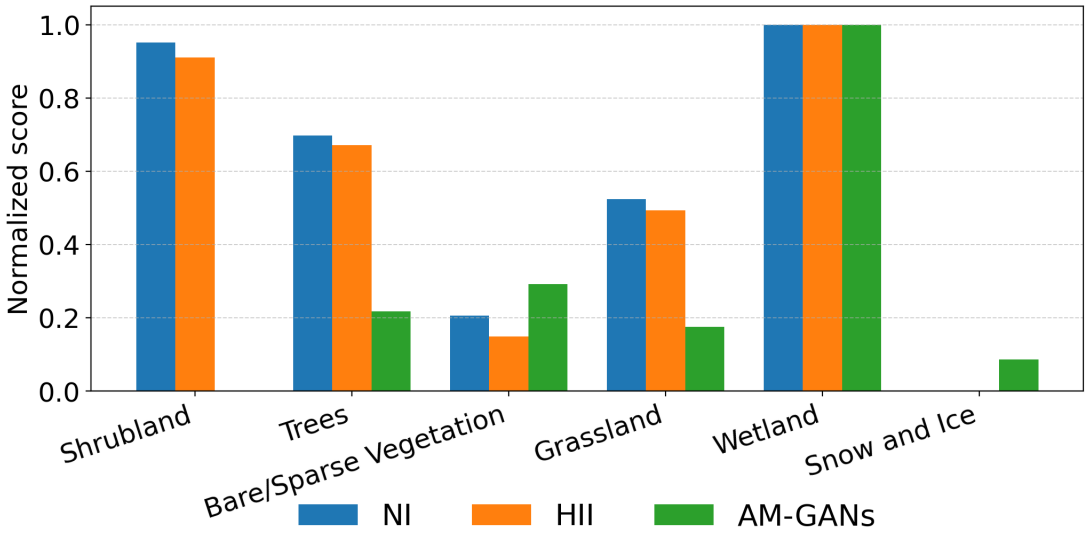


Figure 8.1: Comparison of naturalness score derived from our proposed AM-GANs for Naturalness [1] with the Naturalness Index (NI) [11] and the inverted Human Influence Index (HII) [10] across different land cover classes.

8.1.2 Alignment of CNE Scores with NI and HII

Table 8.2 shows the CNE scores with the naturalness scores derived from the Naturalness Index (NI) [11] and the inverted Human Influence Index (HII) [10].

Shrubland shows a high consistency across all three measures (CNE = 1.000; NI = 0.952; HII = 0.910). Wetlands also align closely, with CNE yielding 0.817 compared to 1.000 (NI) and 1.000 (HII). In contrast, tree cover is underestimated (CNE = 0.417 vs. NI = 0.698; HII = 0.672), suggesting limitations in capturing structural complexity within this class. These results highlight the advantages of integrating UQ methods into explainability techniques. By down-weighting low-confidence predictions, XAI explanations become more valid and consistent.Figure 8.2 visualizes these comparisons. These conclusions are also in line with recent findings by Emam et al. and Essenfelder[28, 29]

Table 8.2: Comparison of Confident Naturalness Explanation (CNE) scores with the Naturalness Index (NI) [11] and the inverted Human Influence Index (HII) [10].

Land Cover Class	**CNE**	**NI**	**HII**
Shrubland	1.000	0.952	0.910
Trees	0.417	0.698	0.672
Wetlands	0.817	1.000	1.000
Bare/Sparse Vegetation	0.500	0.206	0.149
Grassland	0.000	0.524	0.493
Snow and Ice	0.033	0.000	0.000

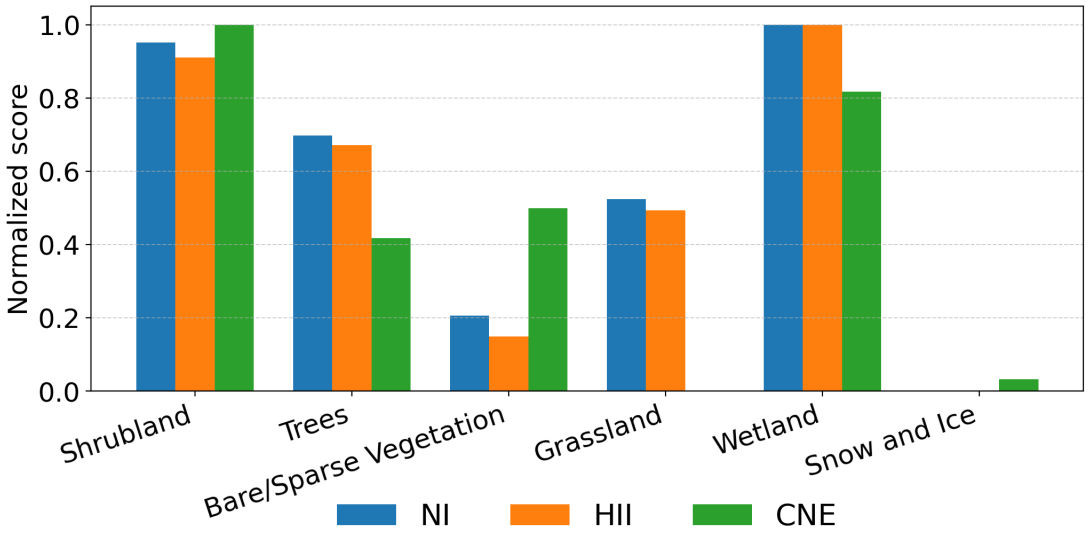


Figure 8.2: Comparison of CNE-derived naturalness scores [4] with the Naturalness Index (NI) [11] and the inverted Human Influence Index (HII) [10] across land cover classes.

8.1.3 Alignment of ReX Scores with NI and HII

Shrublands exhibit the highest alignment across all measures (ReX = 1.000; NI = 0.952; HII = 0.910), whereas wetlands show the lowest alignment (ReX = 0.067; NI = 1.000; HII = 1.000). The corresponding values for all land cover classes are provided in Table 8.3.

Figure 8.3 illustrates these comparisons, showing that ReX reproduces broad naturalness patterns while systematically underestimating wetlands and, to a lesser extent, grasslands and trees. This suggests that ReX can reliably capture stable ecological structures but struggles with classes characterized by higher ecological variability. We believe that increasing the model's confidence in these underestimated land cover classes could further improve the alignment of our Nat-ReX framework with NI and HII, leading to more consistent and ecologically valid results [29, 28].

Table 8.3: Comparison of ReX naturalness scores with the Naturalness Index (NI) and reversed Human Influence Index (HII) across land cover classes.

Land Cover Class	**ReX**	**NI**	**HII**
Shrublands	1.000	0.952	0.910
Trees	0.550	0.698	0.672
Bare/Sparse vegetation	0.283	0.206	0.149
Grassland	0.200	0.524	0.493
Wetland	0.067	1.000	1.000
Snow and ice	0.000	0.000	0.000

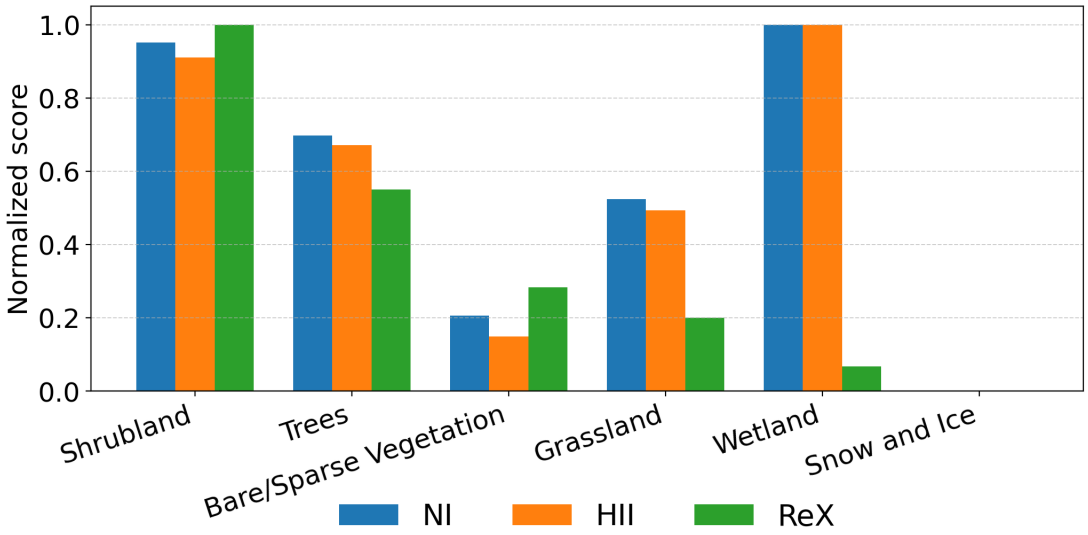


Figure 8.3: ReX scores compared with the Naturalness Index (NI) and reversed Human Influence Index (HII) across land cover classes.

8.1.4 Correlation between the Proposed and the Established Frameworks

To quantify alignment across frameworks, we computed Pearson (r) and Spearman (ρ) correlations with NI [11] and HII[10]. CNE shows the highest alignment with both NI and HII ($r_{\mathrm{NI}} = 0.74$; $r_{\mathrm{HII}} = 0.73$), indicating the strongest consistency with the reference indices. In contrast, AM-GANs exhibit the lowest alignment ($r_{\mathrm{NI}} = 0.42$; $r_{\mathrm{HII}} = 0.45$), reflecting weaker correspondence with NI and HII. The detailed correlation results are presented in Table 8.4, and the overall framework alignment is illustrated in Figure 8.3.

Table 8.4: Alignment of each framework with NI [11] and HII [10]: Pearson r and Spearman ρ across six land-cover classes.

Framework	NI		HII	
	r	ρ	r	ρ
CNE	0.74	0.66	0.73	0.66
ReX	0.51	0.37	0.48	0.37
AM-GANs	0.42	0.26	0.45	0.26

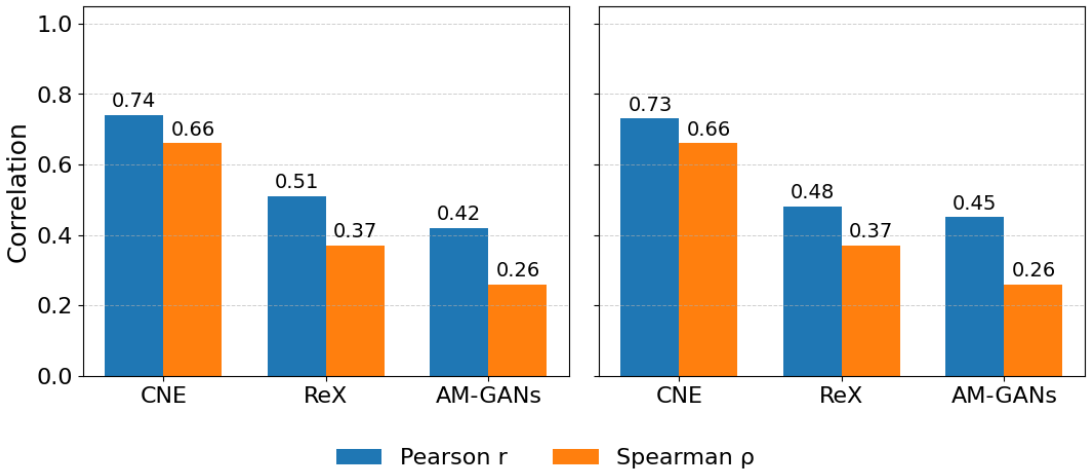


Figure 8.4: Alignment of framework naturalness scores with NI [11] (left) and reversed HII [10] (right). Bars show Pearson r (blue) and Spearman ρ (orange) across six land-cover classes (scores min–max normalized to $[0, 1]$). CNE exhibits the strongest agreement, followed by ReX; AM-GANs is weaker.

8.1.5 Sources of Variations in The Proposed Frameworks

The differences in naturalness scores across our proposed AM-GANs[1], CNE [4], and NaT-ReX [7] frameworks arise from their fundamentally different strategies for leveraging the latent knowledge of trained models for explainability. Each method employs distinct explainable machine learning techniques, with varying degrees of integration of uncertainty quantification, which shapes both the interpretation and reliability of their outputs [28].

AM-GANs for Naturalness [1] combine a trained regressor with activation maximization and generative adversarial training to produce attribution maps highlighting features associated with naturalness [2]. This approach enables interpretable sample generation but suffers from notable limitations. Generated images can contain artifacts [41, 2], and uncertainty arises jointly from the regressor, the generators, and the data. Moreover, the framework does not explicitly address or weigh this uncertainty, leaving the compounded sources of uncertainty unaccounted for. These shortcomings contribute to the divergence of AM-GANs for Naturalness results from our other proposed frameworks, where uncertainty is explicitly modeled and integrated into the explanation process.

The CNE framework uses a semantic segmentation model as a pixel-level, multi-class classifier and distills its latent knowledge into a simpler surrogate model [59]. While this strategy provides class-level naturalness patterns with reduced computational cost, it does not preserve spatial relations between pixels, limiting the expressiveness of explanations. Uncertainty is estimated using Monte Carlo dropout [6], which highlights class-level variance but cannot capture fine-grained spatial uncertainty.

The Nat-ReX framework [7], in contrast, is a Vision Transformer (ViT)-based architecture trained in a multitask fashion. Instead of regressing continuous scores, it employs a binary classification head to distinguish naturalness from non-naturalness, alongside a reconstruction head that is used to estimate the model's uncertainty in the inference phase. Explanations are derived using Layer-wise Relevance Propagation (LRP) with attention rollout [9], while uncertainty is quantified via Monte Carlo dropout [6] applied to the reconstruction head. This dual mechanism produces pixel-wise attribution maps complemented by uncertainty pixel and class-level estimates, addressing the limitations observed in AM-GANs for Naturalness and CNE [1, 4].

A shared limitation across all frameworks is their dependence on land cover segmentation datasets such as CORINE [115] and ESA WorldCover [93] as domain knowledge to contextualize numerical outputs. While these products are widely accepted, they remain approximate: for example, areas labeled as "wetland" may contain diverse ecological elements such as ponds, mixed vegetation, or transitional zones. This mismatch becomes especially critical for the NaT-ReX framework, which generates fine-grained pixel-level relevance maps, thereby amplifying the ecological interpretation limits imposed by approximate categorical inputs.

In sum, the methodological distinctions, the extent to which the appearance of naturalness is incorporated, and the approaches to uncertainty quantification collectively explain the discrepancies observed among the three frameworks presented in this thesis. Each of the proposed frameworks is built on a different backbone with distinct strategies to integrate knowledge about naturalness, ranging from a regressor to a pixel-level segmentation model to a ViT-based scene classifier. These designs incorporate varying levels of uncertainty about land cover classes. Beyond these structural differences, our aim was to explore naturalness through multiple perspectives, employing different approaches and learning strategies [4, 1, 7]. However, despite differences in the ordering of land cover class contributions to naturalness, all proposed frameworks consistently identify the same six land cover classes as the primary positive contributors to the appearance of naturalness.

8.2 Ecological Significance of Key Contributors to Naturalness

To contextualize our naturalness assessments, we draw on the ecological characteristics of land cover classes that contribute to naturalness. These include shrublands, wetlands, sparsely vegetated areas, forests, grasslands, and glaciers.

8.2.1 Shrublands

Shrublands—including moors, heathlands, and transitional communities—are widespread across upland boreal zones of Fennoscandia. Dominated by low woody vegetation such as heather (*Calluna vulgaris*), bilberry (*Vaccinium myrtillus*), and dwarf birch (*Betula nana*), they are adapted to nutrient-poor soils, short growing seasons, and harsh climatic conditions [161].

These communities typically establish in areas where tree growth is limited by shallow soils, cold climate, or exposure, and they persist without intensive land management. Low-intensity grazing or practices such as rotational burning may occur, but generally maintain, rather than transform, their ecological structure. Coastal heathlands, for example, host specialized fungal and invertebrate communities shaped by centuries of low-level human activity with minimal ecosystem disruption [162].

Long-term ecological studies demonstrate that shrubland vegetation structure and species composition remain remarkably stable under climatic variability and moderate land-use pressures [161, 15, 16], highlighting their resilience and value as ecological baselines for naturalness assessment.

Shrublands contribute strongly to the perception of naturalness because they typically occur in regions unsuitable for agriculture or urbanization due to climatic and edaphic constraints, are primarily maintained by natural processes such as succession, disturbance, and climatic filtering, and provide extensive habitats with limited fragmentation or artificial modification. They therefore serve as reliable indicators of minimal human impact and ecological continuity in high-latitude European landscapes [15, 16, 163].

8.2.2 Wetlands

Wetlands—including inland marshes, peat bogs, and fens—are widely recognized as some of the most pristine land cover types due to their hydrological complexity, long-term stability, and limited accessibility. They play a crucial role in biodiversity conservation and carbon storage, making them ecologically and climatologically valuable [164, 165].

Peat bogs are formed over centuries or millennia as decaying sphagnum moss and other moisture-adapted plants accumulate under waterlogged conditions. The anaerobic environment slows decomposition, allowing thick layers of peat to develop [166]. Typical vegetation includes sphagnum mosses, cotton grass (*Eriophorum*), dwarf birch (*Betula nana*), and sedges (*Carex* spp.), all adapted to acidic and nutrient-poor soils.

Fens and marshes, though richer in minerals than bogs, remain permanently saturated and support diverse wetland plant communities. In boreal Fennoscan-

dia, many wetlands remain ecologically intact, particularly in remote regions and protected areas such as Oulanka and Muddus National Parks [10].

Because wetlands are difficult to drain or cultivate, they are rarely converted to agriculture or settlement, making them strong indicators of minimal human impact. Intact peatlands also serve as paleoclimatic archives and reservoirs for specialized and endemic species [166, 167].

Figure 8.5: Example of a bald cypress swamp ecosystem as part of a wetland. Wetlands are defined by saturated soils that support specialized plant and animal life in low-oxygen conditions. While this image depicts a southern U.S. swamp, similar hydrological and ecological features occur in Fennoscandian peat bogs and marshes, making such wetlands important indicators of high naturalness due to limited human influence. Source: Encyclopædia Britannica [168]

8.2.3 Bare and Sparsely Vegetated Areas

This category includes landscapes with little or no vegetative cover, such as bare rock, scree slopes, glacial outwash plains, alpine tundra, and dunes. Their structure is shaped mainly by abiotic constraints, including extreme temperatures, shallow or absent soils, frost-heave, wind erosion, and seasonal snow or ice cover [133]. In Fennoscandia, these environments are concentrated in Arctic and alpine zones, particularly in Norway's Scandes Mountains and northern Lapland, where climatic and geomorphological conditions restrict primary productivity.

Where vegetation occurs, it is sparse and dominated by lichens, mosses, and stress-tolerant pioneer species such as *Dryas octopetala*, *Saxifraga oppositifolia*,

Figure 8.6: Blanket peat bog moorland on Kinder Scout, UK. This type of landscape represents a classic example of a peat-dominated wetland, characterized by persistent waterlogging, sphagnum moss accumulation, and extremely low human impact. In Fennoscandian contexts, similar boreal bogs are widespread and signify high ecological naturalness due to their hydrological isolation and unsuitability for agriculture or development. Source: Martyn Williams [164]

and crustose lichens. These form discontinuous, low-biomass communities adapted to cold, desiccation, and nutrient-poor substrates [165, 16]. The ecological simplicity and fragility of such systems make them sensitive to disturbance, yet their inaccessibility and limited economic value reduce direct anthropogenic pressures [165, 133].

Because they are unsuitable for commercial forestry or intensive grazing, these regions are rarely subject to settlement or infrastructure development. They thus represent valuable ecological baselines where natural geomorphic and successional processes prevail. Their persistence and resistance to land conversion make them strong indicators of naturalness [133, 10].

8.2.4 Forests

Boreal forests, particularly coniferous stands, form some of the most extensive and ecologically significant ecosystems in Fennoscandia. Dominated by Norway spruce (*Picea abies*) and Scots pine (*Pinus sylvestris*), they cover vast areas with relatively limited permanent human alteration. In remote regions of northern Sweden, Finland, and Norway, these woodlands maintain near-natural dynamics shaped by periodic disturbances such as wildfire, windthrow, and insect outbreaks [169].

Natural boreal forests are structurally complex, with multilayered canopies, coarse woody debris, uneven-aged stands, and diverse understories of bilberry (*Vaccinium myrtillus*), lingonberry (*Vaccinium vitis-idaea*), mosses, and lichens [170]. These attributes are key indicators of ecological integrity and can, to some degree, be inferred from high-resolution satellite imagery.

Although intensive forestry has simplified large parts of the Scandinavian forest landscape—through clear-cutting, monocultures, and soil preparation—significant tracts remain semi-natural or undisturbed. Such areas are often located in mountainous zones, peatland–forest ecotones, and protected reserves [171]. They typically lack infrastructure, regenerate slowly, and support species assemblages shaped over centuries without direct human intervention. Forests with these characteristics make strong contributions to both the perception and quantification of naturalness [133, 16, 133, 165].

8.2.5 Natural Grasslands

Natural grasslands are open herbaceous ecosystems that develop in areas with little or no agricultural modification. In Fennoscandia, they typically occur in subarctic valleys, alpine slopes, glacial terraces, and other marginal lands where climatic or edaphic conditions prevent intensive land use [173]. These habitats support long-established plant communities dominated by stress-tolerant species

Figure 8.7: Young Norway spruce (*Picea abies*) sapling growing in a boreal forest environment. Norway spruce is one of the dominant native conifer species in Fennoscandia and forms dense, cold-tolerant woodlands that are characteristic of the region's natural forests. Areas with regenerating native spruce stands are typically minimally disturbed and indicate high ecological value. Source: The Hedge Nursery National Park [172]

such as *Deschampsia flexuosa*, *Festuca ovina*, and *Bistorta vivipara*, adapted to nutrient-poor soils, snow cover, and natural disturbances including frost heave and wild herbivore grazing. In contrast to managed grasslands or pastures, natural grasslands are not fertilized, reseeded, or plowed, and often retain floristic continuity with pre-industrial landscapes [174, 175].

In sum, these land cover classes provide ecologically grounded reference points that help guide the interpretation of naturalness scores. Their physical, biological, and land-use characteristics ensure they remain among the least altered by humans, particularly in the Fennoscandian region. Therefore, their spatial correlation with high naturalness scores in our framework serves as both a qualitative validation of the approach and an ecological anchor for interpreting the results [10, 33, 16, 133, 165].

8.2.6 Glaciers, Snow, and Ice

Glaciers and permanent snowfields in Fennoscandia—especially in the high-altitude regions of the Scandinavian Mountains—are formed and maintained by natural climatic and geological processes [176, 177]. Their existence depends on low temperatures, heavy snowfall, and limited solar radiation, which make these environments some of the most pristine and least modified landscapes in Europe.

Because of the extreme cold, steep terrain, and short growing seasons, human activity in these areas is minimal. They are unsuitable for agriculture, construction, or intensive land use. Most human presence is restricted to scientific monitoring or occasional recreation, such as at the long-term glacier observatory Storglaciären in northern Sweden. Located at high altitudes and in remote northern zones, these snow- and ice-dominated regions remain extensive, continuous, and largely inaccessible. Their natural formation, climatic dependence, and negligible human disturbance make them strong positive indicators of naturalness in Fennoscandia [176, 177, 133].

Chapter 9

Conclusion

This dissertation addressed the challenge of understanding and assessing the appearance of naturalness from satellite imagery using XAI methods, with a strong emphasis on interpretability and uncertainty awareness. Naturalness is important for biodiversity and ecosystem functioning, and its systematic assessment via satellite imagery supports effective environmental monitoring and decision-making. However, existing approaches often lacked interpretability, data-driven pattern attribution, and consideration of uncertainty, which, individually or in combination, reduce trust and applicability in real-world settings and introduce bias in naturalness explanations [28, 10, 27].

To overcome these issues, this dissertation introduced three novel frameworks that combine XAI and UQ. Each framework contributes a novel perspective on how to analyze naturalness spatially and semantically, enabling more transparent insights and supporting better conservation planning. By utilizing uncertainty-aware interpretable results from deep learning models, this dissertation further enhances our current knowledge of the appearance of naturalness in satellite imagery [28]. The dissertation used two satellite datasets, focusing on their overlapping regions to obtain coherent explanations of naturalness appearance within a specific ecosystem [33, 12].

9.1 Short Summary of Key Contributions

In Chapter 5, we developed the AM-GANs for Naturalness framework [1], a generative framework that integrates activation maximization [2] within the CycleGAN-styled [3] objective function to generate satellite images that exaggerate or suppress the patterns contributing to naturalness in satellite imagery. By doing so, we constructed visually meaningful attribution maps that capture domain-relevant features associated with the appearance of naturalness. The framework employs a classifier-based feedback loop to steer the generator

toward producing images that isolate naturalness and non-naturalness visual features. This approach provides visual, semantically valid interpretations of what constitutes ecological naturalness in satellite imagery. Furthermore, we proposed a fully data-driven approach to both quantitatively and qualitatively explain the appearance of naturalness [1].

In Chapter 6, to provide an interpretable and uncertainty-aware representation of naturalness in satellite imagery, we proposed the CNE framework [4]. This method combines deep semantic segmentation [5], surrogate modeling for explainability [71], and Monte Carlo Dropouts [6] to build class-wise uncertainty-aware explanations of naturalness predictions. The key contribution of CNE is its ability to connect land cover classes with naturalness scores through a logistic regression surrogate, where the attribution is weighted by predictive confidence. The proposed CNE index combines the surrogate's feature coefficients and the class-wise uncertainty estimates, yielding confident insights into which land cover types are strongly and reliably associated with naturalness in satellite imagery. The method leverages the model's latent knowledge while providing a data-driven, uncertainty-weighed explanation of the model predictions. To the best of our knowledge, this is the first approach to integrate surrogate modeling as a form of XAI with UQ in remote sensing applications, specifically for explaining the appearance of naturalness.

In chapter 7, to overcome limitations of class-level attribution and lack of spatial awareness in CNE, we introduced the NaT-ReX framework[7]. NaT-ReX is a Transformer-based architecture[8] that combines Layer-wise Relevance Propagation (LRP) attention rollout [9] for spatial explainability with Monte Carlo Dropout-based uncertainty estimation [6] in the reconstruction head. The proposed ReX index captures fine-grained pixel-level relevance while weighing it by the associated model's uncertainty. In addition to its local, high-resolution explanations, NaT-ReX also supports class-wise analysis by aggregating pixel-level scores across land cover types. The framework thus achieves dual-scale interpretability—both pixel and class-level—enabling a detailed understanding of naturalness.

In chapter 8, we compared the naturalness explanations from our three frameworks with established external indices such as the Human Influence Index (HII) [10] and its modernized version, the Naturalness Index (NI) [11]. We analyzed not only the statistical correlation between the methods but also delved into the ecological foundations of high-scoring land cover classes. In particular, we investigated why classes such as wetlands, shrublands, forests, bare/sparse vegetation, and permanent glaciers consistently emerged as strong contributors to naturalness. These land cover types are often located in climatically or geomorphologically inhospitable zones with minimal agricultural or

urban exploitation, and their ecological resilience and isolation from direct human transformation make them highly aligned with the concept of naturalness. The chapter provides a detailed discussion of where and why our methods converge or diverge from existing indices and highlights the ecological validity and limitations of different land cover representations.

Together, these contributions form a coherent set of explainable and uncertainty-aware approaches for naturalness assessment in Earth observation. Each framework targets a unique aspect of the problem: AM-GANs for Naturalness provides generative insights, CNE emphasizes class-wise interpretation with confidence weighting, and NaT-ReX offers spatially structured, uncertainty-aware transformer-based relevance maps.

9.2 Open Source Contributions

To ensure transparency, reproducibility, and accessibility, all methods developed in this thesis are released as open-source code. The repositories provide trained models, scripts, and documentation, enabling reproduction, extension, and application in remote sensing contexts:

- **AM-GANs for Naturalness Framework** [1]: `https://github.com/ahmedemam576/SpacEX`
- **Confident Naturalness Explanation Framework (CNE)** [4]: `https://github.com/ahmedemam576/confident_explanations`
- **NaT-ReX Framework** [7]: `https://github.com/ahmedemam576/vit_rex`

Each repository includes usage instructions to support adoption. By releasing these tools openly, this thesis contributes to advancing explainable Earth observation and naturalness assessment.

9.3 Collaborative Projects and Contributions

Parts of this thesis were conducted in collaboration with an interdisciplinary research team working at the intersection of remote sensing and digital agriculture. These collaborative efforts extended the core principles of XAI and UQ into the context of decision support systems for sustainable crop production. Two key publications emerged from these joint projects:

- **A Framework for Enhanced Decision Support in Digital Agriculture Using Explainable Machine Learning**, presented at ECCV 2024

Workshops. This work adapted principles of XAI approaches and pattern attribution to the agricultural domain, enabling domain experts to better understand how different machine learning models make their decisions in the context of harvest readiness of crops [35].

- **Enhancing Decision Support in Crop Production: Analyzing Conformal Prediction for Uncertainty Quantification**, published in *Computers and Electronics in Agriculture*, 2025. This study explored the integration of conformal prediction techniques into smart agriculture pipelines, providing statistically valid confidence intervals for decision-critical outcomes in precision farming [37].

They demonstrate how explainable and uncertainty-aware machine learning techniques can support decision making beyond environmental monitoring, particularly in domains that require interpretability, trust, and actionable insights for non-technical stakeholders.

9.4 Future Work

The three frameworks developed in this thesis—AM-GANs [1], CNE [4], and NaT-ReX [7], highlight the value of interpretability and uncertainty quantification in naturalness assessment. While effective, they remain task-specific and domain-limited. Future work should move toward scalable, generalizable foundation models for remote sensing that treat naturalness as a semantic objective and integrate uncertainty estimation directly into their design.

9.4.1 Toward Ecology-Oriented Foundation Models

Current frameworks are tailored to specific tasks, limiting transferability across regions and applications. Large-scale EO foundation models pretrained on large global satellite imagery of naturalness could enable robust adaptation and even zero-shot applications to new ecosystems. To advance naturalness assessment, future efforts should develop foundation models explicitly tailored to ecological semantics. Ultimately, we should aim for foundation models that are not only generalizable but also inherently explainable and specialized for naturalness understanding [147, 146].

9.4.2 Efficient Uncertainty Quantification

Current UQ methods (e.g., MC Dropout) are computationally costly and scale poorly to large EO models. Throughout this thesis, we primarily employed Monte Carlo Dropout [6], which captures only model (epistemic) uncertainty. Future work should aim for approaches that capture both model and data (aleatoric) uncertainty, providing a more complete picture of prediction reliability. Promising directions include deterministic methods such as evidential deep learning and deep deterministic uncertainty (DDU) [155], which avoid stochastic sampling while offering calibrated uncertainty estimates. In addition, lightweight fine-tuning techniques (e.g., LoRA) [178] could be explored as a way to efficiently embed UQ into foundation models, enabling uncertainty-aware adaptation at scale with minimal computational overhead.

9.4.3 Better on-the-ground data and better segmentation maps

Future naturalness datasets should pair satellite patches with on-the-ground observations (in situ), such as vegetation plots, species lists, bioacoustic recordings, camera-trap counts, trail counters, and simple checks of road and path access. These observations should match the satellite date as closely as possible and include notes about location accuracy. Labels should combine several public sources so the model learns naturalness together with habitat and management status: WDPA protection [109], IUCN categories [179], ESA WorldCover [93], CORINE Land Cover [115], the Human Influence Index [10], and nighttime lights [114]. Practical clean-up steps include using superpixels, propagating coarse labels to finer grids, and training with loss functions that are robust to noisy labels. Extra context layers such as slope, ruggedness, rivers, peatland likelihood, and recent disturbance (for example, clear-cuts or fires) can be added as extra input channels or predicted as small side tasks. For sensitive sites, blur or coarsen locations and publish a short data card that explains choices and limits. AnthroProtect [107] and MapInWild [33] are good starting points for building such datasets.

9.4.4 Integrate knowledge from text

A lot of what we know about naturalness is written in words, not stored in pixels. Examples are park management plans, habitat guides, WDPA notes [109], and research papers. We can link these texts to satellite images so the model learns both what an area looks like and what people say about it. One simple approach is to train the model to match short sentences to image patches, for example, “no roads, subalpine birch forest,” and to utilize multiple languages so that the model transfers more effectively across regions. Clear rules found in text can also guide learning without many manual labels, for example, “motorized access allowed” usually means lower naturalness, while “wilderness area” usually means higher naturalness. Knowledge graphs that link each image patch to its protection status, ecoregion, biome, and management regime add context and help carry consistent information to nearby unlabeled areas. This makes the system easier to trust because predictions come with plain-language reasons, not only numbers.

Chapter 10

Appendix

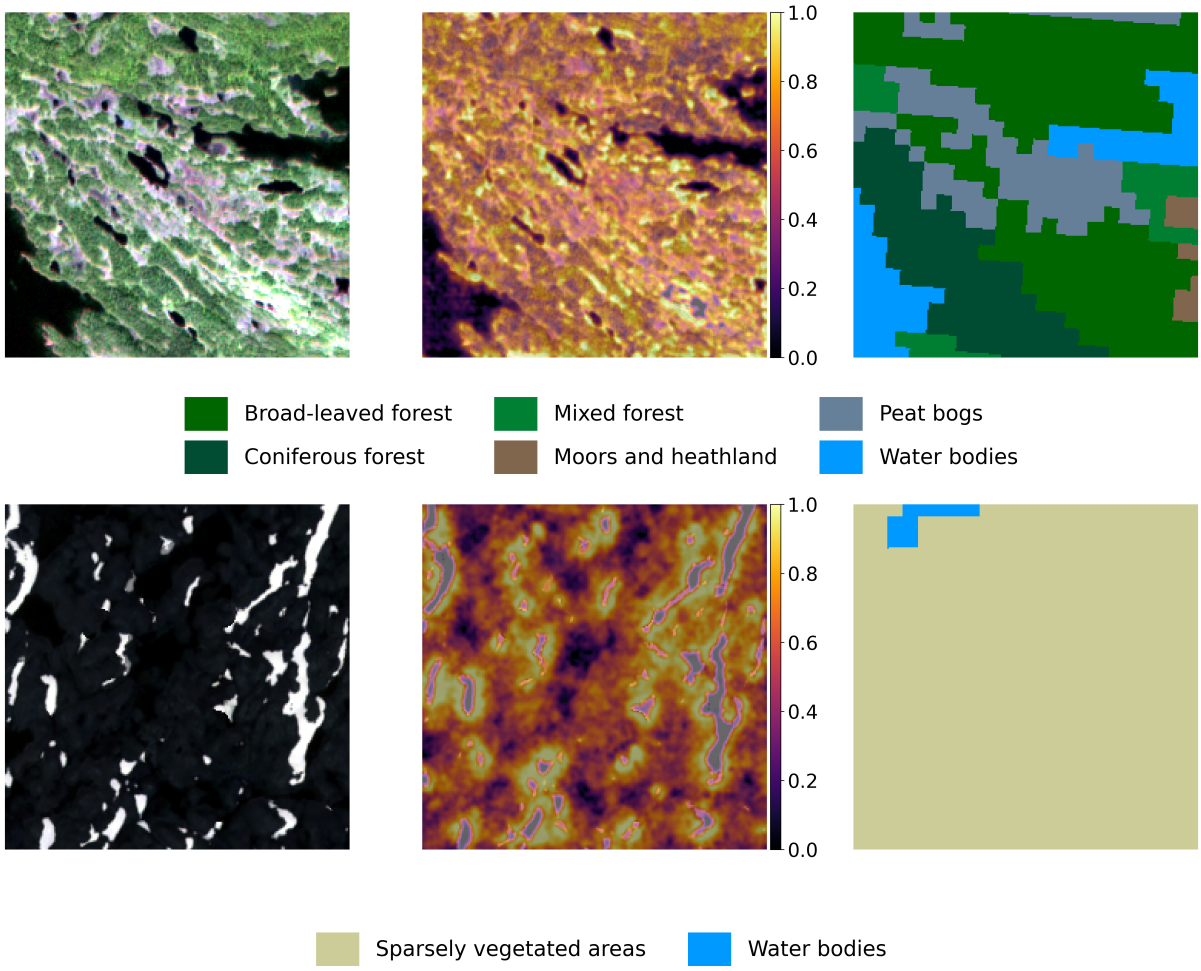


Figure 10.1: Qualitative examples illustrating three columns per row, from the AM-GANs for Naturalness framework (part 1 of 2): the original image (left), the attribution map generated by AM-GANs for naturalness (center), and the segmentation map corresponding to the original image (right).

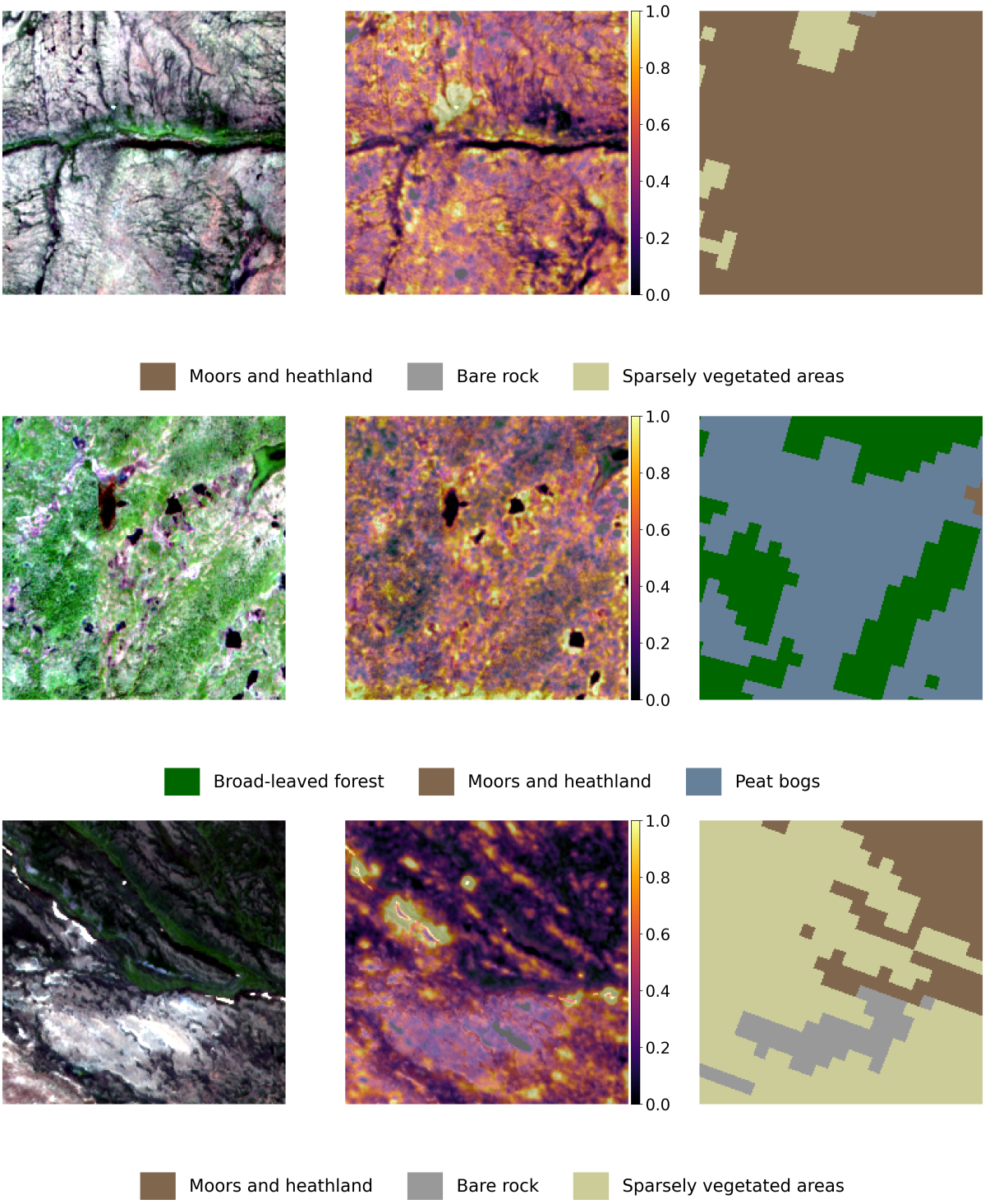


Figure 10.2: Qualitative examples illustrating three columns per row, from the AM-GANs for Naturalness framework (part 2 of 2): the original image (left), the attribution map generated by AM-GANs for naturalness (center), and the segmentation map corresponding to the original image (right).

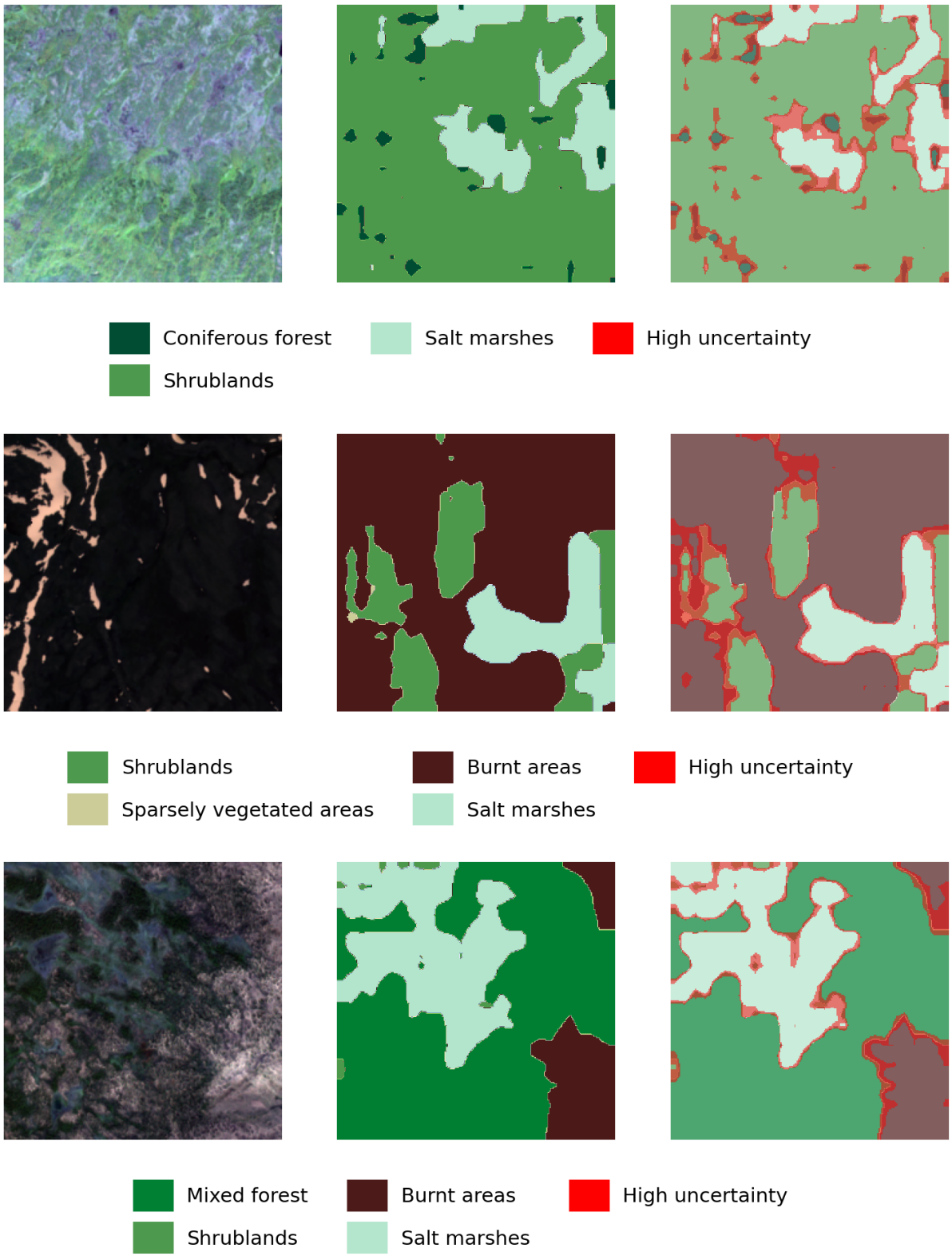


Figure 10.3: Qualitative results of the CNE framework (part 1 of 2). The left column shows the original input images, the middle column displays the mean predicted segmentation maps, and the right column presents the predicted segmentation maps with overlaid uncertainty maps, where red indicates regions of higher uncertainty.

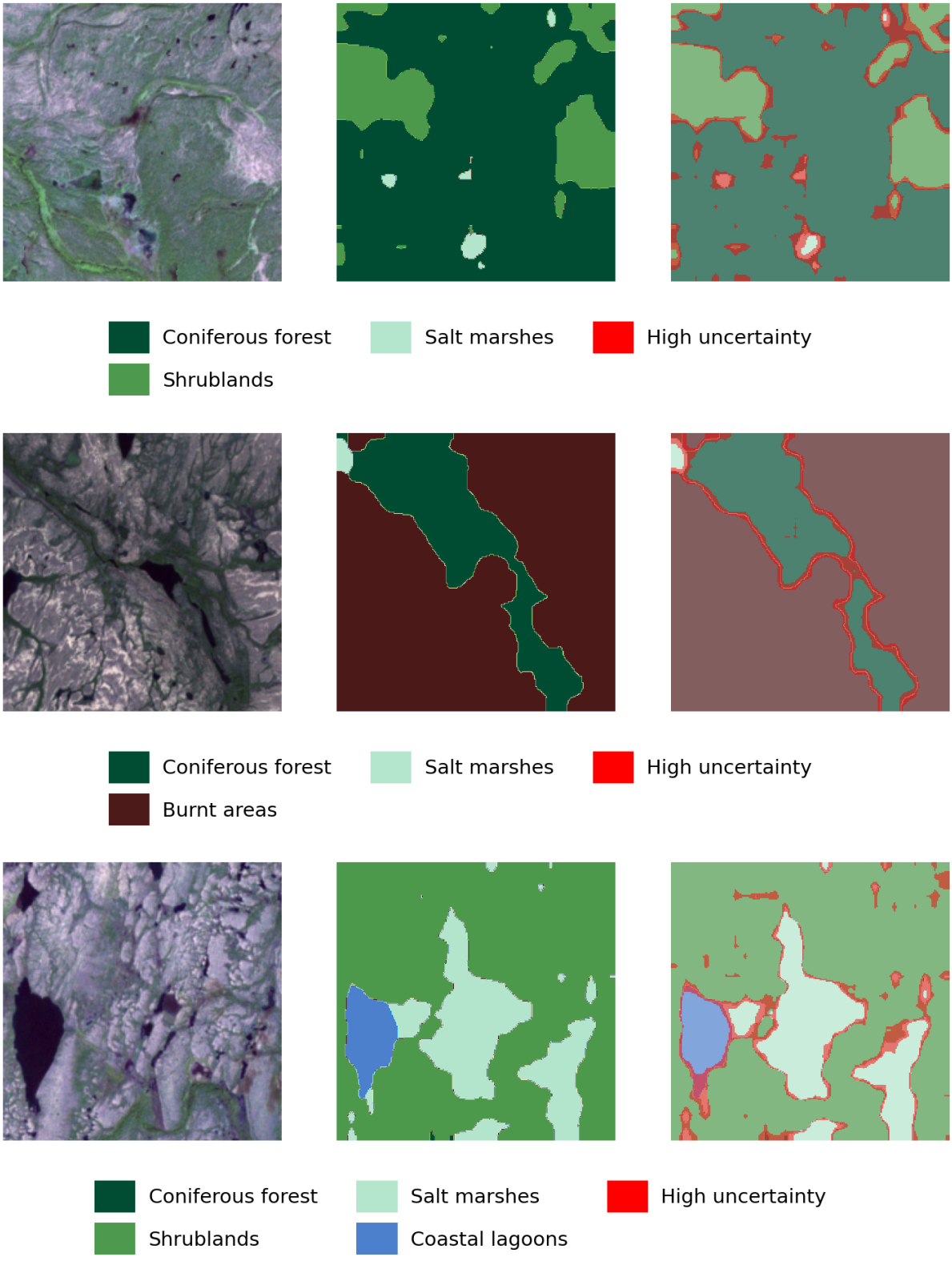


Figure 10.4: Qualitative results of the CNE framework (part 2 of 2). The left column shows the original input images, the middle column displays the mean predicted segmentation maps, and the right column presents the predicted segmentation maps with overlaid uncertainty maps, where red indicates regions of higher uncertainty.

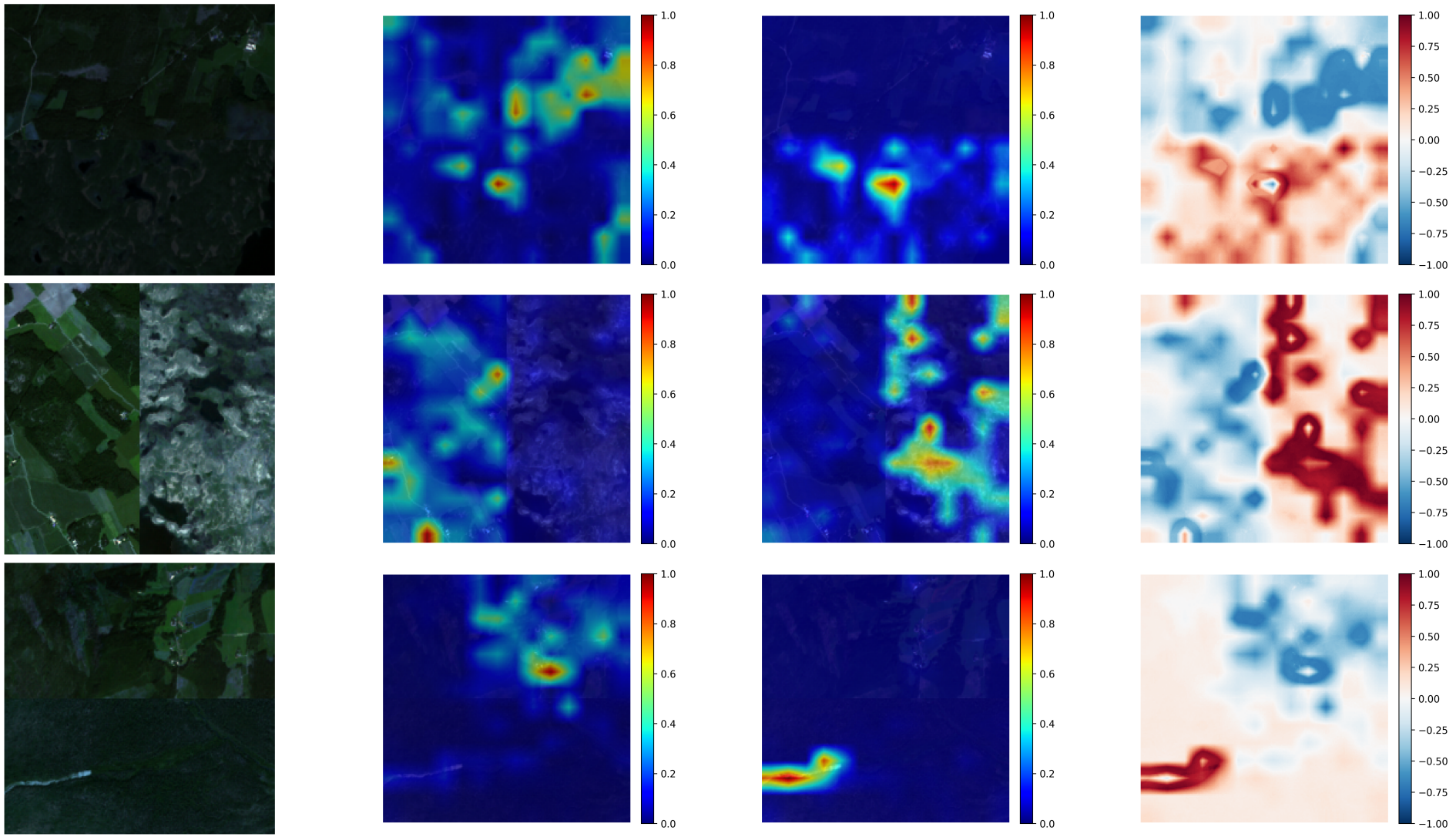


Figure 10.5: Four columns show: input Sentinel-2 image; relevance patterns for the non-naturalness class; relevance for the naturalness class; and an attribution map combining both, with red indicating naturalness and blue indicating non-naturalness relevance. Relevance maps are generated using LRP attention rollout [9].

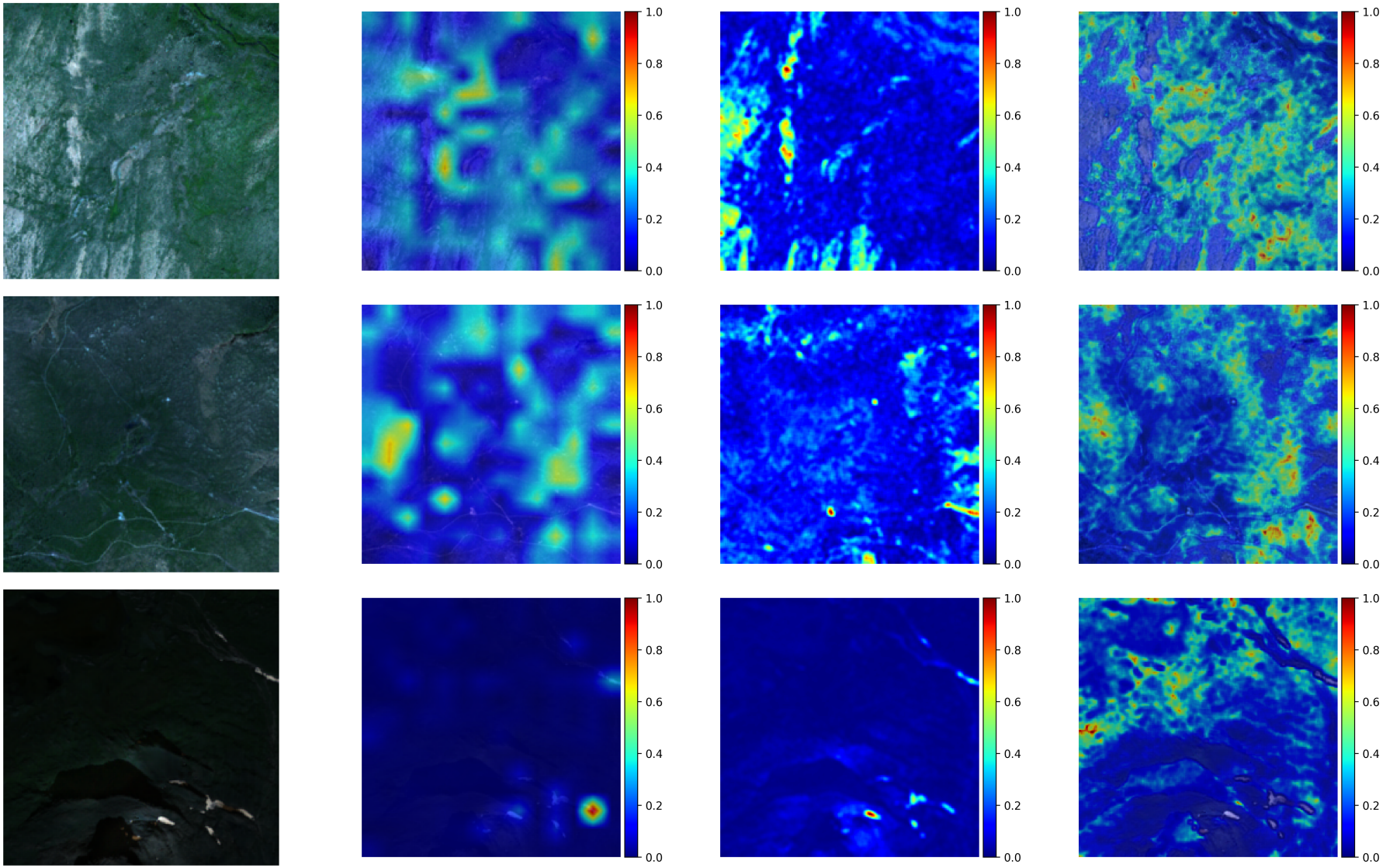


Figure 10.6: Extended visual outputs from the NaT-ReX framework (1 of 2): input images, relevance maps, uncertainty maps, and resulting ReX score maps for various protected areas in Fennoscandia.

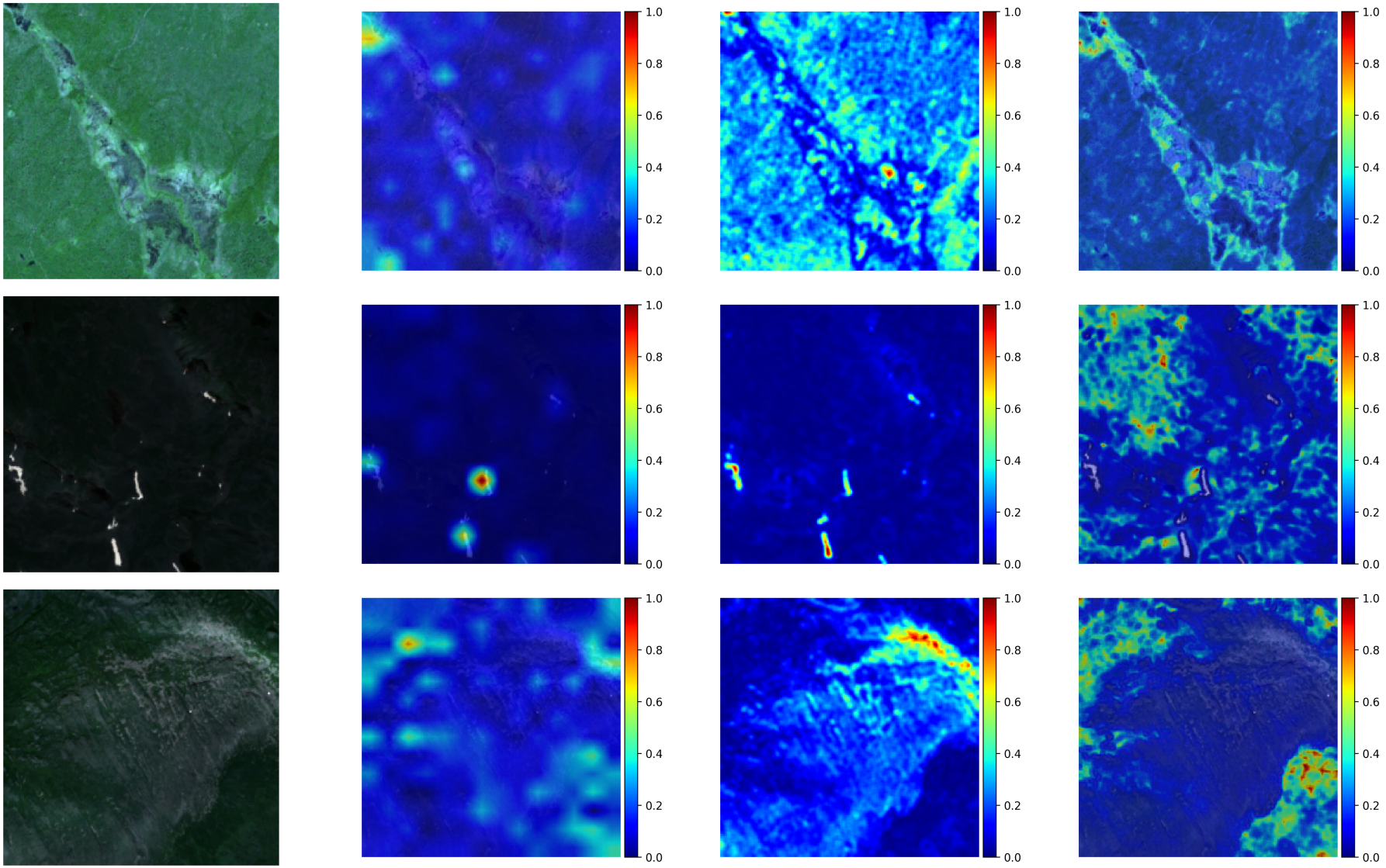


Figure 10.7: Extended visual outputs from the NaT-ReX framework(2 of 2): input images, relevance maps, uncertainty maps, and resulting ReX score maps for various protected areas in Fennoscandia.

Bibliography

[1] A. Emam, T. T. Stomberg, and R. Roscher, "Leveraging Activation Maximization and Generative Adversarial Training to Recognize and Explain Patterns in Natural Areas in Satellite Imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[2] A. Nguyen, J. Yosinski, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] A. Emam, M. Farag, and R. Roscher, "Confident Naturalness Explanation (CNE): A Framework to Explain and Assess Patterns Forming Naturalness," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[6] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 2015.

[7] A. Emam, M. Farag, M. Rußwurm, and R. Roscher, "NaT-ReX: Naturalness Assessment with Transformer-Based Reliable Explainability," in *Pattern Recognition* (M. Keuper and F. Locatello, eds.), (Springer, Cham), pp. 571–585, Springer Nature Switzerland, 2026. Conference name: DAGM - German Conference for Pattern Recognition (GCPR), 2025.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly,

J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, arXiv, June 2021.

[9] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Nashville, TN, USA), pp. 782–791, IEEE, June 2021.

[10] E. W. Sanderson, M. Jaiteh, M. A. Levy, K. H. Redford, A. V. Wannebo, and G. Woolmer, "The Human Footprint and the Last of the Wild: The human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not," *BioScience*, vol. 52, pp. 891–904, Oct. 2002.

[11] B. Ekim, Z. Dong, D. Rashkovetsky, and M. Schmitt, "The naturalness index for the identification of natural areas on regional scale," *International Journal of Applied Earth Observation and Geoinformation*, vol. 105, p. 102622, Dec. 2021.

[12] T. T. Stomberg, J. Leonhardt, I. Weber, and R. Roscher, "Recognizing protected and anthropogenic patterns in landscapes using interpretable machine learning and satellite imagery," *Frontiers in Artificial Intelligence*, vol. 6, Dec. 2023. Publisher: Frontiers.

[13] P. B. Landres, M. W. Brunson, L. Merigliano, C. Sydoriak, and S. Morton, "Naturalness and wildness: the dilemma and irony of managing wilderness," in *Wilderness science in a time of change conference*, vol. 5, pp. 377–390, USDA Forest Service, 2000.

[14] R. A. Mittermeier, C. G. Mittermeier, T. M. Brooks, J. D. Pilgrim, W. R. Konstant, G. A. B. da Fonseca, and C. Kormos, "Wilderness and biodiversity conservation," *Proceedings of the National Academy of Sciences*, vol. 100, pp. 10309–10313, Sept. 2003.

[15] S. Winter, "Ermittlung von strukturellen indikatoren zur abschätzung des einflusses forstlicher bewirtschaftung auf die biozönosen von tieflandbuchenwäldern," *Dissertation, TU Dresden*, 2005.

[16] S. Winter, H. S. Fischer, and A. Fischer, "Relative quantitative reference approach for naturalness assessments of forests," *Forest Ecology and Management*, vol. 259, no. 8, pp. 1624–1632, 2010.

[17] S. Outdoors, "Hermannsdalstinden hike – at the roof of Moskenesoya island, Lofoten," Aug. 2019.

[18] A. Vali, S. Comai, and M. Matteucci, "Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review," *Remote Sensing*, vol. 12, no. 15, 2020.

[19] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, Á. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, A. Mosavi, and G. Camps-Valls, "Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources," *CoRR*, vol. abs/2012.05795, 2020.

[20] G. Taşkın, E. Aptoula, and A. Ertürk, "Explainable ai for earth observation: Current methods, open challenges, and opportunities," *arXiv preprint arXiv:2311.04491*, 2023.

[21] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 680–688, 2016.

[22] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, P. VanValkenburgh, S. A. Wernke, and Y. Huo, "AI Foundation Models in Remote Sensing: A Survey," Aug. 2024.

[23] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[24] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable Machine Learning for Scientific Insights and Discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020. Conference Name: IEEE Access.

[25] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

[26] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, 2018.

[27] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, L. Nachman, R. Chunara, M. Srikumar, A. Weller, and A. Xiang, "Uncertainty as a form of

transparency: Measuring, communicating, and using uncertainty," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 401–413, 2021.

[28] A. Emam and R. Roscher, "Confidence-filtered relevance (cfr): An interpretable and uncertainty-aware machine learning framework for naturalness assessment in satellite imagery," 2025.

[29] A. Essenfelder *et al.*, "Expert-driven explainable ai models for agriculture-related climate hazard detection," *Communications Earth & Environment*, vol. 6, p. 123, 2025.

[30] H. Aggarwal, P. Lohia, A. Agarwal, B. Panda, V. Ramesh, and M. Shrivastava, "Generative models for explainability: Survey and framework," *arXiv preprint arXiv:2102.03046*, 2021.

[31] Z.-H. Zhou and et al., "Activation maximization generative adversarial nets," *arXiv preprint arXiv:1703.02000*, 2017.

[32] B. Hall *et al.*, "Interpreting a land cover classifier with shap and global surrogates," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 8, pp. 1355–1359, 2020.

[33] B. Ekim, T. T. Stomberg, R. Roscher, and M. Schmitt, "MapInWild: A remote sensing dataset to address the question of what makes nature wild [Software and Data Sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, pp. 103–114, Mar. 2023.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

[35] A. Emam, M. Farag, J. Kierdorf, L. Klingbeil, U. Rascher, and R. Roscher, "A Framework for Enhanced Decision Support in Digital Agriculture Using Explainable Machine Learning," in *Computer Vision – ECCV 2024 Workshops* (A. Del Bue, C. Canton, J. Pont-Tuset, and T. Tommasi, eds.), pp. 31–45, Springer Nature Switzerland.

[36] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *ArXiv*, vol. abs/2107.07511, 2021.

[37] M. Farag, A. Emam, J. Leonhardt, and R. Roscher, "Enhancing decision support in crop production: Analyzing conformal prediction for uncertainty

quantification," *Computers and Electronics in Agriculture*, vol. 237, p. 110559, 2025.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1–9, 2015.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015.

[40] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, arXiv, June 2014.

[42] L. Sick, D. Engel, P. Hermosilla, and T. Ropinski, "Attention-Guided Masked Autoencoders For Learning Image Representations," Feb. 2024.

[43] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[44] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," in *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, (Cambridge, MA, USA), pp. 507–513, MIT Press, 1998.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[46] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3431–3440, 2015.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pp. 234–241, Springer, 2015.

[48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[50] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *ECCV*, 2018.

[51] M. Mirza and S. Osindero, "Conditional generative adversarial nets," in *arXiv preprint arXiv:1411.1784*, 2014.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[53] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[54] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[55] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[56] B. D. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 279–288, Association for Computing Machinery, 2019.

[57] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *Digital Signal Processing*, vol. 73, pp. 1–15, 2017.

[58] M. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (J. DeNero, M. Finlayson, and S. Reddy, eds.), (San Diego, California), pp. 97–101, Association for Computational Linguistics, June 2016.

[59] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, Curran Associates, Inc., 2017.

[60] R. Caruana, Y. Lou, J. Gehrke, P. Koch, N. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015.

[61] C. Molnar, *Interpretable Machine Learning*. Lulu.com, 2020.

[62] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[63] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," Oct. 2019.

[64] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), vol. abs/1311.2901 of *Lecture Notes in Computer Science*, (Cham), pp. 818–833, Springer International Publishing, Nov. 2014.

[65] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, 2009. Technical Report.

[66] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," in *Deep Learning Workshop, International Conference on Machine Learning (ICML)*, 2015.

[67] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2014.

[68] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, 2015.

[69] R. Roscher, S. Wenzel, and B. Waske, "Discriminative archetypal self-taught learning for multispectral landcover classification," in *Proc. of Pattern Recogniton in Remote Sensing 2016 (PRRS)*, 2016.

[70] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[71] H. Tan and H. Kotthaus, "Surrogate Model-Based Explainability Methods for Point Cloud NNs," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (Waikoloa, HI, USA), pp. 2927–2936, IEEE, Jan. 2022.

[72] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, p. e0130140, 2015.

[73] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE*, vol. 109, pp. 247–278, Mar. 2021. Conference Name: Proceedings of the IEEE.

[74] G. Taskin, E. Aptoula, and A. Ertürk, "Chapter 7 - explainable ai for earth observation: current methods, open challenges, and opportunities," in *Advances in Machine Learning and Image Analysis for GeoAI* (S. Prasad, J. Chanussot, and J. Li, eds.), pp. 115–152, Elsevier, 2024.

[75] D. A. Freedman, *Statistical models: theory and practice*. Cambridge University Press, 2009.

[76] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *International Conference on Machine Learning*, pp. 1613–1622, 2015.

[77] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," *CoRR*, vol. abs/1706.04599, 2017.

[78] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, ACM, 2002.

[79] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, pp. 61–74, MIT Press, 1999.

[80] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[81] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, and P. K. Dokania, "Calibrating deep neural networks using focal loss," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 15288–15299, 2020.

[82] K. Bykov, M. M.-C. Höhne, K.-R. Müller, S. Nakajima, and M. Kloft, "How Much Can I Trust You? – Quantifying Uncertainties in Explaining Neural Networks," June 2020.

[83] G. Grabherr, G. Koch, H. Kirchmeir, and K. Reiter, *Hemerobie österreichischer Waldökosysteme*, vol. 17. Innsbruck: Veröffentlichungen des Österreichischen MaB-Programms, 1998.

[84] United States Congress, "Wilderness act of 1964." Public Law 88-577, 16 U.S.C. §§1131–1136, 1964. Signed September 3, 1964.

[85] L. J. Cookson, "A definition for wildness," *Ecopsychology*, vol. 3, no. 3, pp. 186–192, 2011.

[86] Norwegian Ministry of Climate and Environment, "Act relating to the management of biological, geological and landscape diversity (nature diversity act)." Unofficial English translation, 2009. Includes provisions for protected areas and restrictions on harmful activities.

[87] Government Offices of Sweden, "The swedish environmental code (1998:808)." Official English version (Ds 2000:61), 1998. Framework law for protected areas and nature protection measures.

[88] Finland, Ministry of the Environment, "Nature conservation act (1096/1996)." Finlex; English translation, 1996. Establishes protected areas; prohibits actions that jeopardize conservation objectives.

[89] C. for International Earth Science Information Network (CIESIN), "Gridded population of the world, version 4.11 (gpwv4): Population density," 2020.

[90] N. Imagery and M. Agency, "Vector map level 0 (vmap0)," tech. rep., United States Department of Defense, 1997.

[91] C. D. Elvidge, K. E. Baugh, J. B. Dietz, T. Bland, P. C. Sutton, and H. W. Kroehl, "Radiance calibration of dmsp-ols low-light imaging data of human settlements," *Remote Sensing of Environment*, vol. 68, pp. 77–88, 1997.

[92] C. D. Elvidge, M. L. Imhoff, K. E. Baugh, V. R. Hobson, I. Nelson, and J. Safran, "Night-time lights of the world: 1994–1995," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 56, pp. 81–99, 1997.

[93] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, N.-E. Tsendbazar, P. Xu, F. Ramoino, and O. Arino, "ESA WorldCover 10 m 2021 v200," Oct. 2022.

[94] W. Schroeder, P. Oliva, L. Giglio, and I. Csiszar, "The new viirs 375 m active fire detection data product: Algorithm description and initial assessment," *Remote Sensing of Environment*, vol. 143, pp. 85–96, 2014.

[95] W. R. Tobler, "Three presentations on geographical analysis and modeling: Non-isotropic geographic modeling, speculations on the geometry of geography, global spatial analysis," Tech. Rep. 93-1, National Center for Geographic Information and Analysis, 1993.

[96] M. Haklay, "How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets," *Environment and Planning B: Planning and Design*, vol. 37, no. 4, pp. 682–703, 2010.

[97] H. Leibundgut, *Europäische Urwälder der Bergstufe*. Haupt Verlag, 1982.

[98] J. Parviainen, "Virgin and natural forests in the temperate zone of europe," *Forest Snow and Landscape Research*, vol. 79, no. 1–2, pp. 9–18, 2005.

[99] M. Christensen, K. Hahn, E. Mountford, P. Odor, D. Rozenberger, J. Diaci, T. Standovar, S. Wijdeven, S. Winter, T. Vrska, and P. Meyer, "Dead wood in european beech (fagus sylvatica) forest reserves," *Forest Ecology and Management*, vol. 210, pp. 267–282, 2005.

[100] B. G. Jonsson, N. Kruys, and T. Ranius, "Ecology of species living on dead wood—lessons for dead wood management," *Silva Fennica*, vol. 39, no. 2, pp. 289–309, 2005.

[101] M. Dietrich, "The lichen flora of the merli forest, giswil ow (central switzerland)," *Botanica Helvetica*, vol. 101, no. 2, pp. 167–182, 1991.

[102] H. Rheault, P. Drapeau, Y. Bergeron, and P. A. Esseen, "Edge effects on epiphytic lichens in managed black spruce forests of eastern north america," *Canadian Journal of Forest Research*, vol. 33, no. 1, pp. 23–32, 2003.

[103] H. S. Fischer, "Simulating the distribution of plant communities in an alpine landscape," *Coenoses*, vol. 5, no. 1, pp. 37–43, 1990.

[104] J. A. G. Jaeger, H. Esswein, H.-G. Schwarz-von Raumer, and M. Müller, "Landschaftszerschneidung in baden-württemberg," *Naturschutz und Landschaftsplanung*, vol. 33, no. 10, pp. 1–13, 2001.

[105] J. M. Anderson, "A conceptual framework for evaluating and quantifying naturalness," *Conservation Biology*, vol. 5, no. 3, pp. 347–352, 1991.

[106] B. Ekim and M. Schmitt, "Mapping land naturalness from sentinel-2 using deep contextual and geographical priors," 2024.

[107] T. T. Stomberg, T. Stone, J. Leonhardt, I. Weber, and R. Roscher, "Exploring wilderness characteristics using explainable machine learning in satellite imagery," *arXiv preprint arXiv:2203.00379*, 2022.

[108] B. Ekim and M. Schmitt, "Deep Occlusion Framework for Multimodal Earth Observation Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[109] UNEP-WCMC and IUCN, "World database on protected areas (wdpa)," 2021.

[110] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ." `https://www.openstreetmap.org`, 2017.

[111] T. Uchiyama, N. Sogi, K. Niinuma, and K. Fukui, "Visually explaining 3D-CNN predictions for video classification with an adaptive occlusion sensitivity analysis," July 2022.

[112] Copernicus Open Access Hub, "Copernicus sentinel-1 data." European Space Agency (ESA), 2014. Available at the Copernicus Open Access Hub.

[113] Copernicus Open Access Hub, "Copernicus sentinel-2 data." European Space Agency (ESA), 2015. Available at the Copernicus Open Access Hub.

[114] M. O. Román, Z. Wang, Q. Sun, V. Kalb, S. D. Miller, A. Molthan, L. Schultz, J. Bell, E. C. Stokes, B. Pandey, K. C. Seto, *et al.*, "Nasa's black marble nighttime lights product suite," *Remote Sensing of Environment*, vol. 210, pp. 113–143, 2018.

[115] European Environment Agency, "CORINE Land Cover 2018, Europe, 6-yearly - version 2020_20u1, May 2020," 2019.

[116] "Use of copernicus sentinel data in research." Copernicus Data Space Ecosystem FAQ, cite format: "Modified Copernicus Sentinel data [Year] processed in Copernicus Browser". Accessed: March 11, 2026.

[117] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD*, vol. 96, pp. 226–231, 1996.

[118] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[119] M. C. Peel, B. L. Finlayson, and T. A. McMahon, "Updated world map of the köppen–geiger climate classification," *Hydrology and Earth System Sciences*, vol. 11, no. 5, pp. 1633–1644, 2007.

[120] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5188–5196, 2015.

[121] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017.

[122] H. Tan, "Visualizing global explanations of point cloud dnns," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 451–461, 2023.

[123] S. Mertes, T. Huber, K. Weitz, A. Heimerl, and E. André, "GANterfactual—Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning," *Frontiers in Artificial Intelligence*, vol. 5, p. 825565, Apr. 2022.

[124] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features," Aug. 2019.

[125] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[126] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, p. 1501–1510, Oct 2017.

[127] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *2017 IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976, July 2017.

[128] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *CVPR*, 2018.

[129] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[130] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[131] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369, 2010.

[132] U. Gunnarsson, M. Löfroth, S. Sandring, and U. Gunnarsson, *The Swedish wetland survey: compiled excerpts from the national final report.* No. 6618 in Rapport / Naturvårdsverket, Stockholm: Swedish Environmental Protection Agency, 2014.

[133] M. Pisaric and J. P. Smol, "Arctic Ecology – A Paleoenvironmental Perspective," in *Arctic Ecology*, pp. 23–55, John Wiley & Sons, Ltd, 2021. Section: 2 _eprint:.

[134] R. Gipiškis, C.-W. Tsai, and O. Kurasova, "Explainable AI (XAI) in Image Segmentation in Medicine, Industry, and Beyond: A Survey," May 2024.

[135] C. Schorr, P. Goodarzi, F. Chen, and T. Dahmen, "Neuroscope: An Explainable AI Toolbox for Semantic Segmentation and Image Classification of Convolutional Neural Nets," *Applied Sciences*, vol. 11, p. 2199, Jan. 2021. Publisher: Multidisciplinary Digital Publishing Institute.

[136] K. Vinogradova, A. Dibrov, and G. Myers, "Towards interpretable semantic segmentation via gradient-weighted class activation mapping," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[137] P. Knab, S. Marton, and C. Bartelt, "Beyond pixels: Enhancing lime with hierarchical features and segmentation foundation models," *arXiv preprint arXiv:2403.07733*, 2024. preprint.

[138] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

[139] M. Rottmann, P. Colling, T.-P. Hack, R. Chan, F. Hüger, P. Schlicht, and H. Gottschalk, "Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities," *arXiv preprint arXiv:1811.00648*, 2018.

[140] T. Brunschwiler, "Sustainable inference of remote sensing data by recursive semantic segmentation – a flood extent mapping study," in *AAAI Spring Symposium Series*, 2023. Demonstrates Monte Carlo dropout–based uncertainty estimation within a recursive segmentation pipeline.

[141] F. Westphal, W. Lidberg, M. Dos Santos Toledo Busarello, and A. M. Ågren, "Uncertainty quantification for lidar-based maps of ditches and natural streams," *Environmental Modelling & Software*, vol. 191, p. 106488, 2025. Compares MC dropout, predictive entropy, mutual information, feature conformal prediction, and conformal regression for LiDAR-based segmentation.

[142] A. Bennetot, G. Franchi, J. Del Ser, R. Chatila, and N. Diaz-Rodriguez, "Greybox XAI: a Neural-Symbolic learning framework to produce interpretable predictions for image classification," Sept. 2022.

[143] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," Dec. 2017. _eprint: 1706.05587.

[144] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," *CoRR*, vol. abs/2002.08791, 2020.

[145] A. Kendall and R. Cipolla, "Modelling Uncertainty in Deep Learning for Camera Relocalization," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2016.

[146] D. Szwarcman, S. Roy, P. Fraccaro, . E. Gíslason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. Almeida, R. Sedona, Y.-J. Kang, S. D. Chakraborty, S. Wang, C. Gomes, A. Kumar, M. Truong, C. Godwin, H.-C. Lee, C.-Y. Hsu, A. A. Asanjan, B. Mujeci, D. Shidham, T. Keenan, P. Arevalo, W. Li, H. Alemohammad, P. Olofsson, C. Hain, R. Kennedy, B. Zadrozny, D. Bell, G. Cavallaro, C. Watson, M. Maskey, R. Ramachandran, and J. Bernabe-Moreno, "Prithvi-eo-2.0: A versatile

multi-temporal foundation model for earth observation applications," *arXiv preprint arXiv:2412.02732*, 2024.

[147] J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp, R. Ramachandran, P. Fraccaro, T. Brunschwiler, G. Cavallaro, J. Bernabe-Moreno, and N. Longépé, "Terramind: Large-scale generative multimodality for earth observation," *arXiv preprint arXiv:2504.11171*, 2025.

[148] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4190–4197, 2020.

[149] L. Yu, W. Xiang, J. Fang, Y.-P. P. Chen, and L. Chi, "ex-vit: A novel explainable vision transformer for weakly supervised semantic segmentation," *Pattern Recognition*, vol. 142, p. 109666, 2023.

[150] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia, and F. S. Khan, "Transformers in remote sensing: A survey," *Remote Sensing*, vol. 15, no. 7, 2023.

[151] A. Höhl, I. Obadić, M. Á. Fernández Torres, H. Najjar, D. Oliveira, Z. Akata, A. Dengel, and X. X. Zhu, "Opening the black-box: A systematic review on explainable ai in remote sensing," *arXiv preprint arXiv:2402.13791*, Feb. 2024.

[152] A. N. Rad *et al.*, "Povit-uq: P-wave polarity and arrival time determination with uncertainty estimation," *Geophysical Journal International*, vol. 243, no. 1, p. ggaf324, 2025.

[153] A. N. Lopez *et al.*, "Uncertainty-aware vision transformers for medical image segmentation," *OpenReview*, 2025.

[154] Y. Zhao *et al.*, "Uncertainty estimation in deterministic vision transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[155] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deep deterministic uncertainty: A new simple baseline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24384–24394, June 2023.

[156] A. N. Wang *et al.*, "Masksemble layer aided cross-vit for uncertainty quantification in skin cancer diagnosis," *Preprint on Figshare*, 2024.

[157] A. M. Sayer, Y. Govaerts, P. Kolmonen, A. Lipponen, M. Luffarelli, T. Mielonen, F. Patadia, T. Popp, A. C. Povey, K. Stebel, and M. L. Witek, "A review and framework for the evaluation of pixel-level uncertainty estimates in satellite aerosol remote sensing," *Atmospheric Measurement Techniques*, vol. 13, pp. 373–404, 2020.

[158] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 2016.

[159] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017. arXiv preprint arXiv:1706.05098.

[160] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2019.

[161] E. D. McArthur and S. G. Kitchen, "Shrubland ecosystems: Importance, distinguishing characteristics, and dynamics," in *USDA Forest Service Proceedings RMRS-P-47*, U.S. Department of Agriculture, Forest Service, 2007.

[162] M. Kerdoncuff, I. E. Måren, and A. E. Eycott, "Traditional prescribed burning of coastal heathland provides niches for xerophilous and sun-loving beetles," *Biodiversity and Conservation*, vol. 31, p. 2521–2543, 2022.

[163] D. J. Forester and G. E. Machlis, "Modeling human factors that affect the loss of biodiversity," *Conservation Biology*, vol. 10, pp. 1253–1263, 1996.

[164] C. B. Staff, "Guest post: How human activity threatens the world's carbon-rich peatlands," Dec. 2020.

[165] K. A. Bråthen, M. Tuomi, J. Kapfer, H. Böhner, and T. Maliniemi, "Changing species dominance patterns of Boreal-Arctic heathlands: evidence of biotic homogenization," *Ecography*, vol. 2024, no. 6, p. e07116, 2024. _eprint:.

[166] S. E. Page and A. J. Baird, "Peatlands and Global Change: Response and Resilience," Oct. 2016.

[167] F. Worrall, P. Chapman, J. Holden, C. Evans, R. Artz, P. Smith, and R. Grayson, "Peatlands and Climate Change,"

[168] Kathryn8, "Bald cypress swamp." Image from Encyclopædia Britannica, 2025. Stately bald cypresses (*Taxodium distichum*) thriving in a Louisiana swamp. © Kathryn8—E+/Getty Images. Accessed: October 4, 2025.

[169] P.-A. Esseen, B. Ehnström, L. Ericson, and K. Sjöberg, "Boreal forests," *Ecological Bulletins*, no. 46, pp. 16–47, 1997.

[170] M. Korkmaz, A. Akyol, T. Turkoglu, A. Bergner, N. Jansson, and A. Tolunay, "Perspective on forest biodiversity indicators for protected areas: A comparison of Turkish and Swedish forest expert opinions," *Applied Ecology and Environmental Research*, vol. 16, pp. 3595–3609, Jan. 2018.

[171] H. Tømmervik, B. Johansen, I. Tombre, D. Thannheiser, K. A. Høgda, E. Gaare, and F. E. Wielgolaski, "Vegetation Changes in the Nordic Mountain Birch Forest: the Influence of Grazing and Climate Change," *Arctic, Antarctic, and Alpine Research*, vol. 36, pp. 323–332, Aug. 2004. Publisher: Taylor & Francis _eprint:

[172] "Norway spruce tree saplings (*Picea abies*)." `https://www.hedgenursery.co.uk/norway-spruce-tree-saplings-picea-abies-p56`. Accessed: October 4, 2025.

[173] R. Virtanen, L. Nagy, J. Jeník, J. Štursa, P. Ozenda, J.-L. Borel, G. Coldea, F. Pedrotti, D. Gafta, D. Gómez, J. A. Sesé, L. Villar, G. Nakhutsrishvili, J. Gamisans, A. Strid, A. Andonoski, and V. Andonovski, "The regional accounts," in *Alpine Biodiversity in Europe* (L. Nagy, G. Grabherr, C. Körner, and D. B. A. Thompson, eds.), pp. 29–121, Berlin, Heidelberg: Springer, 2003.

[174] L. Kullman and L. Kjällgren, "Holocene pine tree-line evolution in the Swedish Scandes: Recent tree-line rise and climate change in a long-term perspective," *Boreas*, vol. 35, no. 1, pp. 159–168, 2006. _eprint:.

[175] M. Eronen, M. Lindholm, S. Saastamoinen, and P. Zetterberg, "Variable Holocene climate, treeline dynamics and changes in natural environments in northern Finnish Lapland," *Chemosphere - Global Change Science*, vol. 1, pp. 377–387, Nov. 1999.

[176] A. P. Stroeven *et al.*, "Deglaciation of fennoscandia," *Quaternary Science Reviews*, vol. 147, pp. 91–121, 2016.

[177] K. Ploeg and A. P. Stroeven, "History and dynamics of fennoscandian ice sheet retreat, contemporary ice-dammed lake evolution, and faulting in the torneträsk area, northwestern sweden," *The Cryosphere*, vol. 19, pp. 347–373, 2025.

[178] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, W. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[179] "Iucn protected area categories." `https://en.wikipedia.org/wiki/IUCN_protected_area_categories`. Accessed on March 11, 2026.

List of Figures

1.1 High-naturalness landscape in Fennoscandia with minimum human influence: glaciated mountains, alpine lakes, and intact tundra–rock vegetation with no visible infrastructure [17]. 2

2.1 The plot illustrates a negative linear relationship between the amount of human activity and the naturalness score, showing that higher levels of human activity are associated with lower naturalness. The underlying data were generated by the author solely for illustrative purposes to demonstrate how linear regression works. 11

2.2 The plot illustrates logistic regression modeling the inverse relationship between human activity and the probability of naturalness. The orange curve shows the predicted probability, while the blue points represent class labels, assigned as 1 (naturalness) when the probability exceeds 0.5 and 0 (non-naturalness) otherwise. The underlying data were generated by the author solely for illustrative purposes to demonstrate how logistic regression works. 13

2.3 Schematic of a single-hidden-layer multilayer perceptron (MLP). The input features $\boldsymbol{x}^{(i)}$ (e.g., human activity, green space, noise level) are transformed through a hidden layer into intermediate representations $\boldsymbol{h}^{(i)}$, which are subsequently mapped to the output $\hat{y}$, representing the predicted naturalness score. 14

2.4 Convolutional neural network to classify areas of naturalness. Convolution extracts local features, ReLU activation function adds nonlinearity, pooling downsamples, and the flattened features feed a fully connected layer that outputs probabilities for the classes Naturalness and Non-naturalness. 16

2.5 Residual block in ResNet. Deep networks learn a residual function $\mathcal{F}(\boldsymbol{x}; \mathsf{W})$ that is added to an identity shortcut to preserve signal and gradients, addressing degradation in very deep models. The block output is $\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \mathsf{W}) + \boldsymbol{x}$; if dimensions differ, the shortcut uses a projection W_s so that $\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \mathsf{W}) + \mathsf{W}_s\boldsymbol{x}$ [39]. 17

2.6 Concept of atrous (dilated) convolution in DeepLab for preserving spatial resolution. top: In a standard CNN, the output stride increases with network depth (e.g., 4, 8, 16, 32), progressively reducing spatial resolution. bottom: DeepLab replaces the later downsampling operations with atrous convolutions using increasing dilation rates (e.g., 2, 4, 8, 16), maintaining a constant output stride (e.g., 16) while enlarging the receptive field for dense prediction [5]. 19

2.7 Effect of the dilation rate in atrous convolution. With a fixed 3×3 kernel, increasing the dilation rate (e.g., 1, 6, 24) enlarges the spacing between sampled pixels, thereby expanding the receptive field while preserving the feature-map resolution without additional downsampling [5] 20

2.8 Basic GAN schematic. The generator $\mathcal{G}$ maps a latent vector $\mathbf{z} \sim p_z(\mathbf{z})$ to a fake sample $\mathbf{x}' = \mathcal{G}(\mathbf{z})$. The discriminator $\mathcal{D}$ receives real data $\mathbf{x} \sim p_r(\mathbf{x})$ and fake data $\mathbf{x}'$ and outputs $\mathcal{D}(\cdot) \in [0, 1]$. Gradients update $\mathcal{D}$ directly and update $\mathcal{G}$ by backpropagating through $\mathcal{D}$ to increase $\mathcal{D}(\mathcal{G}(\mathbf{z}))$ [41]. 21

2.9 Vision Transformer (ViT) with self-attention notation. An image $\mathsf{X} \in \mathbb{R}^{H \times W \times C}$ is split into $N = \frac{HW}{P^2}$ patches; each vectorized patch $\mathsf{x}[i] \in \mathbb{R}^{P^2 C}$ is linearly embedded $\boldsymbol{z}_p^i = \mathsf{x}[i]\mathsf{E} \in \mathbb{R}^D$. A learnable token $\boldsymbol{x}_{\text{class}}$ is prepended and positional embeddings E_{pos} are added to form $\mathsf{z}_0 = [\boldsymbol{x}_{\text{class}}; \boldsymbol{z}_p^1; \ldots; \boldsymbol{z}_p^N] + \mathsf{E}_{\text{pos}}$. The sequence passes through encoder layers using self-attention $\text{Attention}(\mathsf{Q}, \mathsf{K}, \mathsf{V}) = \text{softmax}\big(\mathsf{Q}\mathsf{K}^\mathsf{T}/\sqrt{d_k}\big)\,\mathsf{V}$ (applied in multi-head form). The final class token representation $\boldsymbol{z}_{\text{class}}$ is fed to an MLP head to produce $\hat{y}$ [8]. 23

2.10 Grad-CAM pipeline in our notation. A forward pass provides activation maps A^k at a chosen layer and the class logit $\mathscr{y}^c$. Backpropagation yields gradients $\partial \mathscr{y}^c / \partial \mathsf{A}^k$, from which weights are computed as $\alpha_k^c = \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} \partial \mathscr{y}^c / \partial \mathsf{A}_{u,v}^k$. The class-specific heatmap is $\mathsf{L}_{\text{Grad-CAM}}^c = \text{ReLU}\big(\sum_k \alpha_k^c \, \mathsf{A}^k\big)$, then upsampled and overlaid on the input image. 30

2.11 Monte Carlo Dropout (MC-Dropout). Dropout remains active at inference, randomly closing a subset of hidden units (filled black; dashed connections ignored). For an input X, run T stochastic forward passes $\{\hat{y}^{(t)}\}_{t=1}^T$ with $\hat{y}^{(t)} = f_{\text{drop}}^{(t)}(\mathsf{X})$. The predictive mean and variance are estimated by $\mathbb{E}[\hat{y}] \approx \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)}$ and $\text{Var}[\hat{y}] \approx \frac{1}{T} \sum_{t=1}^{T} \big(\hat{y}^{(t)} - \mathbb{E}[\hat{y}]\big)^2$, providing an approximation to epistemic uncertainty. 37

4.1 Example tiles from the AnthroProtect dataset showing protected areas in Fennoscandia. The satellite image patches depict landscapes with minimal direct human influence [107]. 49

4.2 Collage of Sentinel-2 tiles from the MapInWild dataset, illustrating diverse land-use patterns and degrees of human activity across regions [33]. 51

5.1 Pipeline overview consisting of three sequential phases: Regressor training, which learns class-specific representations; GANs training, which enhances these representations via CycleGAN-guided activation maximization to generate maximized and minimized image pairs; and Attribution mapping, which explains patterns contributing to naturalness by comparing the generated image pairs. 57

5.2 GAN pretraining phase (Stage 1): generators learn to produce realistic satellite imagery while discriminators enforce realism and perceptual similarity. Activation maximization is excluded at this stage. 58

5.3 The AM-GANs for Naturalness in inference: the trained maximizer and minimizer generators map the input image into generated images with higher or lower naturalness scores, enabling attribution map generation based on comparing the generated image pairs. 60

5.4 The qualitative results of the AM-GANs for the Naturalness framework (Part 1 of 2). Each triplet of images includes: (1) the original RGB input, (2) the corresponding attribution map highlighting spatial patterns contributing to naturalness estimation, and (3) the associated semantic segmentation mask for land cover association. The examples illustrate how the framework selectively identifies the ecological patterns that contribute to the appearance of naturalness. 65

5.5 The qualitative results of the AM-GANs for the Naturalness framework (Part 2 of 2). Each triplet of images includes: (1) the original RGB input, (2) the corresponding attribution map highlighting spatial patterns contributing to naturalness estimation, and (3) the associated semantic segmentation mask for land cover association. The examples illustrate how the framework selectively identifies the ecological patterns that contribute to the appearance of naturalness. 66

6.1 Overview of the CNE framework. In the explainability path (top), input images are processed by the segmentation model to produce predicted segmentation masks Y, which are vectorized and paired with naturalness labels to train a logistic regression surrogate $\mathcal{L}$. Only positive coefficients are retained to compute attribution. In the uncertainty path (bottom-left), MC-Dropout generates multiple segmentation outputs, allowing estimation of class-wise uncertainty maps S. Both paths are then combined (bottom-right) to compute the CNE index, assigning confidence-weighted relevance scores to land cover classes. The uncertainty estimation process is detailed in Figure 6.3. 68

6.2 Flowchart of the CNE framework. It consists of two parallel paths operating on segmentation model outputs. The left path performs the explainability task: predicted segmentation masks are vectorized and used to train a logistic regression model with naturalness labels, resulting in attribution scores for each land cover class. The right path estimates uncertainty via MC-Dropout: multiple stochastic forward passes yield class-wise uncertainty maps. Both paths are combined to compute the final CNE index, which assigns confidence-weighted relevance scores to each land cover pattern. Details of the uncertainty estimation process are shown in Figure 6.3. 74

6.3 Illustration of the uncertainty estimation process. Input images are passed through the segmentation model with MC-Dropout. The output tensor Y is indexed by batch b and sample j. On the right, two outputs are shown: the upper image is the mean prediction over J runs (argmax per pixel), while the lower image shows the standard deviation map, where bright pixels indicate high uncertainty. 75

6.4 Uncertainty-aware segmentation maps. Two examples showing Sentinel-2 RGB images, predicted segmentation masks, and grayscale overlaid uncertainty maps. Brighter pixels indicate higher uncertainty. Segmentation colors correspond to different land cover patterns. Dashed and solid red circles show the areas with the lowest and highest confidence, respectively. 76

6.5 Example visualizations of qualitative results from the CNE framework (Part 1 of 3). Each row presents three images: on the left, the original Sentinel-2 image from areas of varying naturalness; in the middle, the average segmentation map obtained from multiple MC-Dropout runs; and on the right, the corresponding uncertainty-aware segmentation map, where red intensity represents the magnitude of predictive uncertainty for each pixel. 80

6.6 Example visualizations of qualitative results from the CNE framework (Part 2 of 3). Each row presents three images: on the left, the original Sentinel-2 image from areas of varying naturalness; in the middle, the average segmentation map obtained from multiple MC-Dropout runs; and on the right, the corresponding uncertainty-aware segmentation map, where red intensity represents the magnitude of predictive uncertainty for each pixel. 81

6.7 Example visualizations of qualitative results from the CNE framework (Part 3 of 3). Each row presents three images: on the left, the original Sentinel-2 image from areas of varying naturalness; in the middle, the average segmentation map obtained from multiple MC-Dropout runs; and on the right, the corresponding uncertainty-aware segmentation map, where red intensity represents the magnitude of predictive uncertainty for each pixel. 82

7.1 Training architecture of the NaT-ReX framework. The input satellite image is processed by a Vision Transformer-based encoder $\mathcal{F}$, which feeds into two parallel heads: a reconstruction head $\mathcal{G}$ that learns to reconstruct the input image, and a classification head $\mathcal{S}$ that predicts naturalness versus non-naturalness. Both heads are trained simultaneously in a multi-task learning setup to optimize both classification and reconstruction objectives. 90

7.2 Inference pipeline of the NaT-ReX framework. The input image is passed through the shared encoder $\mathcal{F}$, producing two outputs: an uncertainty map from the reconstruction head $\mathcal{G}$ using MC-Dropout, and an LRP attention map from the classification head $\mathcal{S}$ via attention rollout. These two outputs are integrated per pixel using the ReX formulation to generate the final ReX map, which highlights spatial contributions to the class naturalness while accounting for uncertainty. 91

7.3 top: NaT-ReX visualizations showing the original input, LRP attention maps for *naturalness*, MC-Dropout uncertainty maps, and resulting ReX maps. bottom: CutMix visualizations for evaluating class-specific relevance under mixed patch compositions. Columns 1 and 3 show CutMix images blending natural and urban regions. Columns 2 and 4 show LRP attention maps for non-naturalness (negative relevance) and naturalness (positive relevance), respectively. Additional examples of original images, attention maps, uncertainty estimates, and resulting ReX maps are shown in the appendix. 93

7.4 Additional qualitative examples (part 1 out of 2). Each row shows: original Sentinel-2 image, LRP attention map for naturalness, MC-Dropout uncertainty map, and ReX score map. These support the spatial interpretation of confidence-weighted naturalness across land cover types. The caption applies to both this figure and the previous one. 95

7.5 Additional qualitative examples (part 2 out of 2). Each row shows: original Sentinel-2 image, LRP attention map for naturalness, MC-Dropout uncertainty map, and ReX score map. These support the spatial interpretation of confidence-weighted naturalness across land cover types. The caption applies to both this figure and the previous one. 96

8.1 Comparison of naturalness score derived from our proposed AM-GANs for Naturalness [1] with the Naturalness Index (NI) [11] and the inverted Human Influence Index (HII) [10] across different land cover classes. 101

8.2 Comparison of CNE-derived naturalness scores [4] with the Naturalness Index (NI) [11] and the inverted Human Influence Index (HII) [10] across land cover classes. 102

8.3 ReX scores compared with the Naturalness Index (NI) and reversed Human Influence Index (HII) across land cover classes. . 103

8.4 Alignment of framework naturalness scores with NI [11] (left) and reversed HII [10] (right). Bars show Pearson r (blue) and Spearman ρ (orange) across six land-cover classes (scores min–max normalized to $[0, 1]$). CNE exhibits the strongest agreement, followed by ReX; AM-GANs is weaker. 104

8.5 Example of a bald cypress swamp ecosystem as part of a wetland. Wetlands are defined by saturated soils that support specialized plant and animal life in low-oxygen conditions. While this image depicts a southern U.S. swamp, similar hydrological and ecological features occur in Fennoscandian peat bogs and marshes, making such wetlands important indicators of high naturalness due to limited human influence. Source: Encyclopædia Britannica [168] . 107

8.6 Blanket peat bog moorland on Kinder Scout, UK. This type of landscape represents a classic example of a peat-dominated wetland, characterized by persistent waterlogging, sphagnum moss accumulation, and extremely low human impact. In Fennoscandian contexts, similar boreal bogs are widespread and signify high ecological naturalness due to their hydrological isolation and unsuitability for agriculture or development. Source: Martyn Williams [164] 108

8.7 Young Norway spruce (*Picea abies*) sapling growing in a boreal forest environment. Norway spruce is one of the dominant native conifer species in Fennoscandia and forms dense, cold-tolerant woodlands that are characteristic of the region's natural forests. Areas with regenerating native spruce stands are typically minimally disturbed and indicate high ecological value. Source: The Hedge Nursery National Park [172] 110

10.1 Qualitative examples illustrating three columns per row, from the AM-GANs for Naturalness framework (part 1 of 2): the original image (left), the attribution map generated by AM-GANs for naturalness (center), and the segmentation map corresponding to the original image (right). 121

10.2 Qualitative examples illustrating three columns per row, from the AM-GANs for Naturalness framework (part 2 of 2): the original image (left), the attribution map generated by AM-GANs for naturalness (center), and the segmentation map corresponding to the original image (right). 122

10.3 Qualitative results of the CNE framework (part 1 of 2). The left column shows the original input images, the middle column displays the mean predicted segmentation maps, and the right column presents the predicted segmentation maps with overlaid uncertainty maps, where red indicates regions of higher uncertainty. 123

10.4 Qualitative results of the CNE framework (part 2 of 2). The left column shows the original input images, the middle column displays the mean predicted segmentation maps, and the right column presents the predicted segmentation maps with overlaid uncertainty maps, where red indicates regions of higher uncertainty. 124
10.5 Four columns show: input Sentinel-2 image; relevance patterns for the non-naturalness class; relevance for the naturalness class; and an attribution map combining both, with red indicating naturalness and blue indicating non-naturalness relevance. Relevance maps are generated using LRP attention rollout [9]. 125
10.6 Extended visual outputs from the NaT-ReX framework (1 of 2): input images, relevance maps, uncertainty maps, and resulting ReX score maps for various protected areas in Fennoscandia. . . 126
10.7 Extended visual outputs from the NaT-ReX framework(2 of 2): input images, relevance maps, uncertainty maps, and resulting ReX score maps for various protected areas in Fennoscandia. . . 127

List of Tables

1.1 Comparison of the three developed frameworks forming the main components of this thesis. The AM-GANs for Naturalness framework [1] leverages XAI, the CNE framework [4] adds UQ and provides land-cover class-level insights, while the NaT-ReX framework [7] integrates both XAI and UQ for both class- and pixel-level insights. 8

2.1 Classification of Explanation Methods in XAI 26

3.1 Tick matrix comparing six methods for naturalness mapping or assessment across key dimensions. Columns indicate whether the method is rule-based (uses manually defined scoring logic), XAI-based (employs explainable AI techniques), and provides high-resolution output (e.g., 10 m maps). Methods include Human Influence Index (HII) [10], Naturalness Index (NI) [11], Relative Qualitative Naturalness Assessment (RANA) [16], Activation Space Occlusion Sensitivity (ASOS) [12], Modality Occlusion [108], and Multimodal Learning [106]. 46

4.1 Key differences between the AnthroProtect [12] and MapInWild [33] datasets. 52

5.1 Performance metrics of the regressor r. 61

5.2 Hyperparameters used for training the framework. 61

5.3 Reconstruction performance of the trained GAN on training and validation sets. Arrows indicate whether higher (↑) or lower (↓) values are preferable. 62

5.4 Mean and standard deviation of each land cover class contribution to naturalness for the AnthroProtect and MapInWild datasets. . . 63

6.1 CNE metric scores and datasets distribution in the regions representing naturalness. The table presents a comparative analysis across two datasets: the top section details insights derived from the AnthroProtect dataset [12], while the bottom section displays results from the MapInWild dataset [33]. 77

7.1 Overview of explanation methods for Vision Transformers [8]. . . 86
7.2 Mean relevance scores [9] and MC-Dropout uncertainties [6] for each land cover class in the AnthroProtect and MapInWild datasets. The final ReX score reflects both semantic relevance (via LRP attention rollout) and the model's uncertainty. 94

8.1 Comparison of AM-GANs naturalness scores with the Naturalness Index (NI) and reversed Human Influence Index (HII) across land cover classes. 100
8.2 Comparison of Confident Naturalness Explanation (CNE) scores with the Naturalness Index (NI) [11] and the inverted Human Influence Index (HII) [10]. 101
8.3 Comparison of ReX naturalness scores with the Naturalness Index (NI) and reversed Human Influence Index (HII) across land cover classes. 102
8.4 Alignment of each framework with NI [11] and HII [10]: Pearson r and Spearman ρ across six land-cover classes. 103