Geospatial Deep Learning for Large Scale Naturalness Mapping



Burak Ekim
Department of Aerospace Engineering
University of the Bundeswehr Munich

A thesis submitted for the degree of
Doctor of Engineering

Supervised by Prof. Dr. Michael Schmitt

Geospatial Deep Learning for Large Scale Naturalness Mapping

Burak Ekim

Vollständiger Abdruck der von der Fakultät für Luft- und Raumfahrttechnik der Universität der Bundeswehr München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

angenommenen Dissertation.

Gutachter:
1. Univ.-Prof. Dr.-Ing. Michael Schmitt
2. Prof. Dr. Jan Dirk Wegner

Die Dissertation wurde am 17.12.2025 bei der Universität der Bundeswehr München eingereicht und durch die Fakultät für Luft- und Raumfahrttechnik am 03.03.2026 angenommen. Die mündliche Prüfung fand am 23.03.2026 statt.

Abstract

Human pressure on nature has reached unprecedented levels, creating an urgent need to measure and monitor its impact. Accurately mapping this pressure is crucial for environmental monitoring and conservation. Traditional methods rely on heuristic combinations of static geo-indicators (e.g., land cover, population), but they are limited by their dependence on potentially inaccurate and low-resolution data, leading to inconsistent assessments.

Earth Observation offers a wealth of satellite data to address this task, yet it introduces its own set of challenges. It is often unclear which data modalities are most effective for a given task. Furthermore, the deep learning models applied to this data must be tailored to understand the unique intricacies of satellite imagery and the specific demands of naturalness mapping—a fundamentally ambiguous task, as the concept can only be approximated through imperfect geo-proxies. Moreover, ensuring these models generalize reliably and remain robust when encountering the inevitable distribution shifts inherent in real-world satellite data presents a significant hurdle. Addressing these challenges is key to producing machine learning-derived map products at a large spatial scale, where they can be most impactful.

This thesis presents a complete, end-to-end methodology for mapping human influence from space—a task termed here *naturalness mapping*, representing the absence of modern human influence—from dataset curation to the deployment of a machine learning-derived large scale map product. The work first establishes the task by introducing a rule-based *Naturalness Index* as an annotation source, accompanying a globally-sampled dataset named *MapInWild*. It then details an interpretable machine learning methodology for identifying the most influential input modalities for the task. Subsequently, it tailors a model architecture suited to the unique demands of naturalness mapping from satellite imagery by incorporating spatial-contextual awareness and geographic location priors. The resulting model is then deployed at a continental scale. Acknowledging the ambiguity inherent in mapping naturalness, alongside the critical challenge of model robustness and generalization, the deployment pipeline integrates a distribution shift detector capable of identifying near-distribution covariate and semantic shifts. This tool produces a companion *Reliability Map* product that provides end-users with a crucial layer of confidence for the continental-scale, machine learning-derived *Naturalness Map* product.

Kurzfassung

Der anthropogene Einfluss auf die Natur hat ein beispielloses Ausmaß erreicht, was die präzise Messung und Überwachung seiner Auswirkungen zwingend erforderlich macht. Herkömmliche Methoden zur Kartierung dieses Einflusses stützen sich auf heuristische Kombinationen statischer Geoindikatoren. Sie sind jedoch durch ihre Abhängigkeit von potenziell ungenauen und niedrig aufgelösten Daten begrenzt, was zu inkonsistenten Bewertungen führt.

Die Erdbeobachtung bietet eine Fülle von Satellitendaten für diese Aufgabe, bringt jedoch eigene Herausforderungen mit sich. Es ist oft unklar, welche Datenmodalitäten am effektivsten sind. Zudem müssen die darauf angewandten Deep-Learning-Modelle auf die spezifischen Eigenheiten von Satellitenbildern und die fundamentale Mehrdeutigkeit der Natürlichkeitskartierung zugeschnitten werden. Eine weitere erhebliche Hürde ist die Gewährleistung der Generalisierungsfähigkeit und Robustheit der Modelle gegenüber unvermeidlichen Verteilungsverschiebungen (*distribution shifts*) in realen Satellitendaten. Die Bewältigung dieser Herausforderungen ist essenziell für die Erstellung großskaliger, KI-basierter Kartenprodukte.

Diese Dissertation präsentiert eine vollständige End-to-End-Methodik zur Kartierung der Abwesenheit moderner anthropogener Einflüsse aus dem Weltraum – hier als *Naturalness Mapping* bezeichnet. Die Arbeit definiert die Aufgabe zunächst durch einen regelbasierten *Naturalness Index* als Annotationsquelle und den globalen Datensatz *MapInWild*. Anschließend wird mithilfe von Methoden des interpretierbaren maschinellen Lernens ermittelt, welche Eingabemodalitäten am einflussreichsten sind. Darauf aufbauend wird eine Modellarchitektur entwickelt, die durch räumlich-kontextuelle Informationen und geografische A-priori-Lageinformationen speziell für die Natürlichkeitskartierung aus Satellitenbildern optimiert ist. Dieses Modell wird auf kontinentaler Ebene eingesetzt. Um der Mehrdeutigkeit der Aufgabe sowie der Herausforderung der Modellrobustheit zu begegnen, integriert die Deployment-Pipeline einen Detektor zur Identifikation von Covariate- und Semantic-Shifts. Dieses Werkzeug erzeugt eine begleitende *Reliability Map*, die Endnutzern eine entscheidende Verlässlichkeitsebene für die resultierende *Naturalness Map* bietet.

Contents

Acknowledgements **ix**

1 Introduction **1**
1.1 Motivation and Objective 1
1.2 Structure of the Thesis 3

2 Theoretical Background **6**
2.1 Sensors and Geospatial Data Sources 6
2.1.1 Sentinel-1 6
2.1.2 Sentinel-2 8
2.1.3 NASA VIIRS 9
2.1.4 ESA WorldCover 9
2.1.5 Elevation Models 10
2.1.6 OpenStreetMap 11
2.2 Machine Learning and Deep Neural Networks 11
2.2.1 Deep Learning and Convolutional Neural Networks 12
2.2.2 Geospatial Deep Learning 14
2.3 Human Influence 19

3 State of the Art **23**
3.1 Paradigms in Human Influence Mapping from Space 23
3.1.1 Rule-Based Mapping 23
3.1.2 Deep Learning-Derived Mapping 26
3.2 Challenges in Large Scale Mapping with Deep Learning 29
3.3 Feature Attribution in Geospatial Deep Learning 31
3.4 Making Deep Learning Geospatial 33
3.4.1 Modelling Spatial Correlation 33
3.4.2 Location Encoding 34
3.5 Distribution Shift in Earth Observation 36
3.6 Contribution of the Thesis 37

4 Dataset Curation for Naturalness Mapping **39**
4.1 From Human Influence to a Naturalness Proxy . . . 39
4.2 Conceptual Framework and Data Sources . . . 40
4.3 Creating the Naturalness Index . . . 40
4.4 Validation as an Annotation Source . . . 41
4.5 Construction of the *MapInWild-NI* Dataset . . . 46
4.5.1 Design Principles and Sampling Strategy . . . 46
4.5.2 Data Acquisition and Quality Assurance . . . 47
4.5.3 Final Dataset . . . 51

5 Deep Learning Methodology for Naturalness Mapping **56**
5.1 Influential Input Modality Selection . . . 56
5.2 Incorporating Geospatial Priors in Deep Learning . . . 59
5.3 Distribution Shift for Large Scale Deployment . . . 62

6 Experiments and Results **64**
6.1 Setting the Stage: Experimental Setup and Foundational Analysis . . . 64
6.1.1 Evaluation Metrics . . . 64
6.1.2 Assessing the Need for Deep Learning . . . 65
6.1.3 Data Centric Machine Learning: Core-set Selection . . . 67
6.2 Influential Modality Selection . . . 67
6.2.1 Model and Training Details . . . 68
6.2.2 Post-Hoc Analysis of Sample-Level Modality Influence . . . 68
6.2.3 Validation via Ablation Study and Aggregated Influence . . . 69
6.2.4 Robustness Check via Occlusion Strategy Investigation . . . 71
6.3 Incorporating Geospatial Priors . . . 72
6.4 Final Model . . . 74
6.5 Out-of-Distribution Detection . . . 77
6.5.1 Preparing the Distribution Shift Detector . . . 77
6.5.2 Detecting Controlled Shifts . . . 79
6.6 Discussion . . . 84
6.6.1 Influential Modality Selection . . . 84
6.6.2 Incorporating Geospatial Priors . . . 84
6.6.3 Analysis of Final Model Performance and Trade-offs . . . 84
6.6.4 Out-of-Distribution Detection . . . 85

7 Naturalness Map of Europe **86**
7.1 Large Scale Deployment: Naturalness Map . . . 86
7.2 Comparison to Existing Human Influence Maps . . . 88
7.3 Large Scale Deployment: Prediction Reliability Map . . . 106

8 Discussion **110**

9 Conclusion and Perspectives **115**

9.1 Summary and Conclusion 115

9.2 Future Work 116

9.3 Ethical Considerations 117

Publications **119**

Bibliography **121**

List of Figures

1.1 Overview of the proposed framework for data-driven naturalness mapping 4

2.1 Multi-sensor satellite and basemap data for Istanbul, Türkiye 7
2.2 The LeNet-5 architecture 14
2.3 Sentinel-2 image pyramid simulation over Ebersberger Forest, Germany . 15
2.4 Seasonal variation over the United Kingdom from Sentinel-2 16
2.5 Contextual input scales around Nairobi National Park, Kenya 18
2.6 Human impact drivers and combined index for the Puget Sound region, US 21

4.1 Naturalness Index calculation workflow 42
4.2 Accesibility and Tobler's hiking function for Naturalness Index 43
4.3 Final Naturalness Index maps for case study regions 43
4.4 Comparison of HM and HF over study areas 44
4.5 Distribution of the NI values across study areas 44
4.6 Comparison of average NI values for different area categories 45
4.7 Close-look comparison of NI, HM, and HF values over protected areas . . 46
4.8 Guided stratified sampling workflow for WDPA polygons 48
4.9 MapInWild-NI dataset curation and geodata querying workflow 49
4.10 Mean NI values per ESA WorldCover class 50
4.11 Percentage breakdown of ESA WorldCover classes in the annotation . . . 51
4.12 Mean NI values across VIIRS nighttime brightness bins 52
4.13 Geographical distribution of AOIs and dataset splits across continents . . 52
4.14 Distribution of NI values in training, validation, and test sets 53
4.15 A sample from MapInWild-NI over Ile-Alatau State National Nature Park 53
4.16 A sample from MapInWild-NI over Kaaterskill 54
4.17 A negative sample from MapInWild-NI over Chennai, India 54

5.1 Framework for land naturalness assessment using occlusion strategy . . . 57
5.2 Latent space fusion framework for naturalness prediction 60
5.3 Overview of the OOD detection method that assigns surrogate labels . . 62
5.4 Four-step workflow of the OOD detection framework 63

6.1 Sample-level modality scores, highlighting VIIRS influence 69

6.2 Sample-level modality scores highlighting Sentinel-2 influence 70
6.3 Performance deviation and embedding distance comparison per modality 71
6.4 Comparison of embedding distances across different occlusion strategies . 72
6.5 Visual inference results over Cape Coral and Copenhagen 73
6.6 Final model: Binned metrics 75
6.7 Final model: Distribution comparison 75
6.8 DL-NI Correlation with existing indices 76
6.9 Final Model: Qualitative prediction maps 78
6.10 TARDIS hyperparameter selection strategy 83

7.1 Naturalness Map of Europe 88
7.2 Human Footprint Index (HFP) map of Europe 92
7.3 Global Human Modification map (gHM) of Europe 93
7.4 Deep Learning-Human Footprint Index (DL-HFI) of Europe 94
7.5 Density plot: DL-NI against gHM 95
7.6 Density plot: DL-NI against HFP 96
7.7 Density plot: DL-NI against DL-HFI 97
7.8 DL-NI, HFP, gHM, and DL-HFI scores across WDPA and built-up areas 98
7.9 Austria map comparison: Sentinel-2, DL-NI, HFP, and gHM 98
7.10 Switzerland map comparison: Sentinel-2, DL-NI, HFP, and gHM 99
7.11 Kosovo map comparison: Sentinel-2, DL-NI, HFP, and gHM 99
7.12 Gibraltar map comparison: Sentinel-2, DL-NI, HFP, and gHM 100
7.13 Bavaria, Germany map comparison: Sentinel-2, DL-NI, HFP, and gHM . 100
7.14 Munich, Germany map comparison: Sentinel-2, DL-NI, HFP, and gHM . 101
7.15 Vilnius, Lithuania map comparison: Sentinel-2, DL-NI, HFP, and gHM . 101
7.16 Vindelfjällen, Sweden map comparison: Sentinel-2, DL-NI, HFP, and gHM 102
7.17 Gabès, Tunisia map comparison: Sentinel-2, DL-NI, HFP, and gHM . . . 102
7.18 Métlaoui, Tunisia map comparison: Sentinel-2, DL-NI, HFP, and gHM . 103
7.19 Negrine, Algeria map comparison: Sentinel-2, DL-NI, HFP, and gHM . . 103
7.20 Oriental, Morocco map comparison: Sentinel-2, DL-NI, HFP, and gHM . 104
7.21 Taghramt, Morocco map comparison: Sentinel-2, DL-NI, HFP, and gHM 104
7.22 Tetouan, Morocco map comparison: Sentinel-2, DL-NI, HFP, and gHM . 105
7.23 Continental-level Prediction Reliability Map 107
7.24 Kernel density plot of TARDIS scores 108
7.25 WILD predictions and reliability scores 109

List of Tables

3.1 Existing Human Influence indices 25

4.1 Comparison of geospatial data sources for HII and NI 40
4.2 IUCN protected area categories 47
4.3 Seasonal distribution of the dataset across train, validation, and test sets. 50

6.1 Determining the appropriate learning formulation for the task 66
6.2 Ablation Study: Influence metrics per modality 70
6.3 UNet conditioned on geospatial priors 73
6.4 Summary of experimental setups for controlled covariate and semantic shifts 80
6.5 Quantitative comparison of g_{oracle} and g^* classifiers on EuroSAT and xBD 82

Acknowledgements

This thesis stands on the shoulders of many. While any attempt to capture a years-long journey in a few lines will inevitably fall short, I nonetheless feel compelled to try.

I am grateful to my advisor, Michael Schmitt. Michael, I deeply appreciate your unwavering support, your guidance, and the space you provided for me to explore and grow throughout my PhD.

Over the course of this journey, I have crossed paths with countless incredible individuals. To my friends and colleagues in Munich: this chapter of my life was made significantly richer and more enjoyable because I was able to share it with all of you. I was equally fortunate to collaborate with amazing people during my research stays at the University of Bonn, Microsoft, and Google. To avoid the risk of inadvertently omitting anyone from a long list, I have chosen not to name people individually. To all of you: you surely know who you are, and I am deeply grateful. It truly takes a village.

Lastly, my deepest gratitude goes to my family: my sister, my mom, and my father. This milestone is as much yours as it is mine. Sizi seviyorum.

Chapter 1

Introduction

1.1 Motivation and Objective

Recent advances in Earth observation provide widespread availability of satellite imagery at a planetary scale, with ever-improving temporal and spatial resolutions. This technological progress is fundamental for environmental monitoring, as it allows us to observe the terrestrial ecosystems shaped by human activity. For millennia, humans have altered the land to suit our needs—from clearing forests to constructing cities. In the current era, the cumulative effect of these actions has led to significant consequences for our planet. Quantifying this human pressure on nature is therefore essential for guiding conservation planning, ecosystem management, and sustainable development. Over the past two decades, mapping human pressure has become an important tool for research and policy-making, aiding in the assessment of habitat loss, deforestation risks, and species extinction threats. These maps inform spatial planning, land-use decisions, and help identify critical intact habitats, enabling targeted conservation, restoration efforts, and long-term biodiversity protection. In light of global agendas such as the Sustainable Development Goals (SDGs) and the Kunming-Montreal Global Biodiversity Framework, quantifying human pressure on nature is more urgent than ever, and this effort begins with accurate and large scale mapping that is both temporally and spatially consistent.

A substantial body of literature exists on mapping human pressure, with prominent examples including the Human Influence Index [1], the Human Footprint Index [2], and the Global Human Modification Index [3]. These conventional methods rely on integrating various geo-indicators—such as land cover, population, and infrastructure—using heuristic combinations to create proxies for anthropogenic drivers. While foundational, this approach offers only a partial view of the complex, multidimensional nature of environmental disturbance because it relies on a limited set of pre-defined proxy indicators and heuristic rules to represent a phenomenon that is influenced by countless interacting factors. Furthermore, its reliance on these drivers presents a significant challenge, as their

completeness and accuracy vary by region, creating the potential for biased or misleading impact assessments.

These limitations motivate a shift towards machine learning methods capable of learning the complex, associative relationships directly from satellite imagery. By removing the reliance on external geo-indicators and their inherent inaccuracies, such an approach can quantify human pressure using only the image data itself. Realizing this potential for globally consistent, machine learning-derived map products that can be produced at any time and place where satellite data is available is a key motivation for this thesis.

However, adopting a machine learning–driven approach introduces its own set of research frontiers. To date, no prior studies have introduced an end-to-end framework to map human pressure from satellite imagery alone. Developing such a system requires addressing several interconnected challenges: identifying which of the many available features are most informative for the task; tailoring model architectures to the unique characteristics of the satellite data; and scaling the model for large scale deployment. Only by tackling these issues can the full potential of this data-driven paradigm be realized, producing the information-rich map products needed to inform modern conservation and policy initiatives.

To address these challenges, this thesis develops a methodology that builds upon the concept of human influence mapping but reframes the task as *naturalness mapping*—quantifying the *absence* of human influence—culminating in a continental-scale demonstration of the approach. It begins by introducing a rule-based human pressure index, termed *Naturalness*, to generate annotations as a form of supervision. A globally distributed multi-modal dataset, *MapInWild*, is then curated and paired with these annotations to form *MapInWild-NI*, the basis for training a deep learning model. The methodology then introduces a feature attribution method from the field of interpretable machine learning to identify the most salient input data modalities for the task. This analysis reveals that Sentinel-2 imagery is the most influential data source for predicting naturalness. Building on this insight, the subsequent steps proceed using only Sentinel-2 data. Recognizing that an off-the-shelf deep learning model is sub-optimal, the thesis further details how the model architecture is tailored to the unique challenges of the task and satellite data by incorporating mechanisms to model spatial correlation and encode geographic location. The final deep learning model is designed to produce dense, pixel-wise naturalness predictions from Sentinel-2 imagery alone. Finally, acknowledging the challenges inherent in this work—that deep learning-derived maps are never perfect, that "naturalness" is a concept that is difficult to define fundamentally, and that the annotations used are themselves only proxies—the framework proposes a post-hoc, model-agnostic distribution shift detector for real-world deployment. This detector addresses the lack of labeled out-of-distribution examples by quantifying distributional shifts at inference time using surrogate labels derived from the feature space. This measure is

then used to produce a companion *Reliability Map* to accompany the primary *Naturalness Map*. This additional map product provides a layer of transparency about the model's predictive performance, enabling more informed decision-making by end-users.

This thesis follows a progression from data curation to model design and large scale deployment. Figure 1.1 provides an overview of this end-to-end framework. The process begins by forming annotation sources and curating a dataset from globally sampled satellite imagery for the task of naturalness mapping. Following this, the framework proceeds by estimating the relative importance of available input modalities to determine which data sources are most relevant for naturalness prediction. Based on these insights, a geospatially aware deep learning model is designed to integrate spatial and geographical context. Finally, the deep learning model, wrapped with the distribution shift detection tool tailored for large scale inference, jointly predicts the naturalness map and the prediction reliability map.

1.2 Structure of the Thesis

The structure of this thesis is as follows:

- Chapter 1 (Introduction): Motivates the research, outlines the core challenges in mapping human influence, and provides a high-level overview of the proposed methodology and contributions.

- Chapter 2 (Theoretical Background): Covers the foundational knowledge required for the subsequent work. It details the geospatial data sources central to this thesis and introduces the fundamental principles of deep learning, culminating in a discussion of the *Geospatial Deep Learning* domain.

- Chapter 3 (State of the Art): Reviews the relevant literature. It explores the historical and current paradigms of human influence mapping, discusses the challenges of large scale deployment, and examines the literature on feature attribution, geospatial-aware deep learning, and distribution shift detection. The chapter concludes by identifying the critical research gaps that motivate the contributions of this thesis.

- Chapter 4 (Dataset Curation for Naturalness Mapping): Details the creation of the foundational data for this thesis. It introduces the rule-based *Naturalness Index* developed to serve as the annotation source and describes the curation of the *MapInWild-NI* dataset (including sampling, data acquisition, and assembly), which is used for the supervised training of the deep learning model.

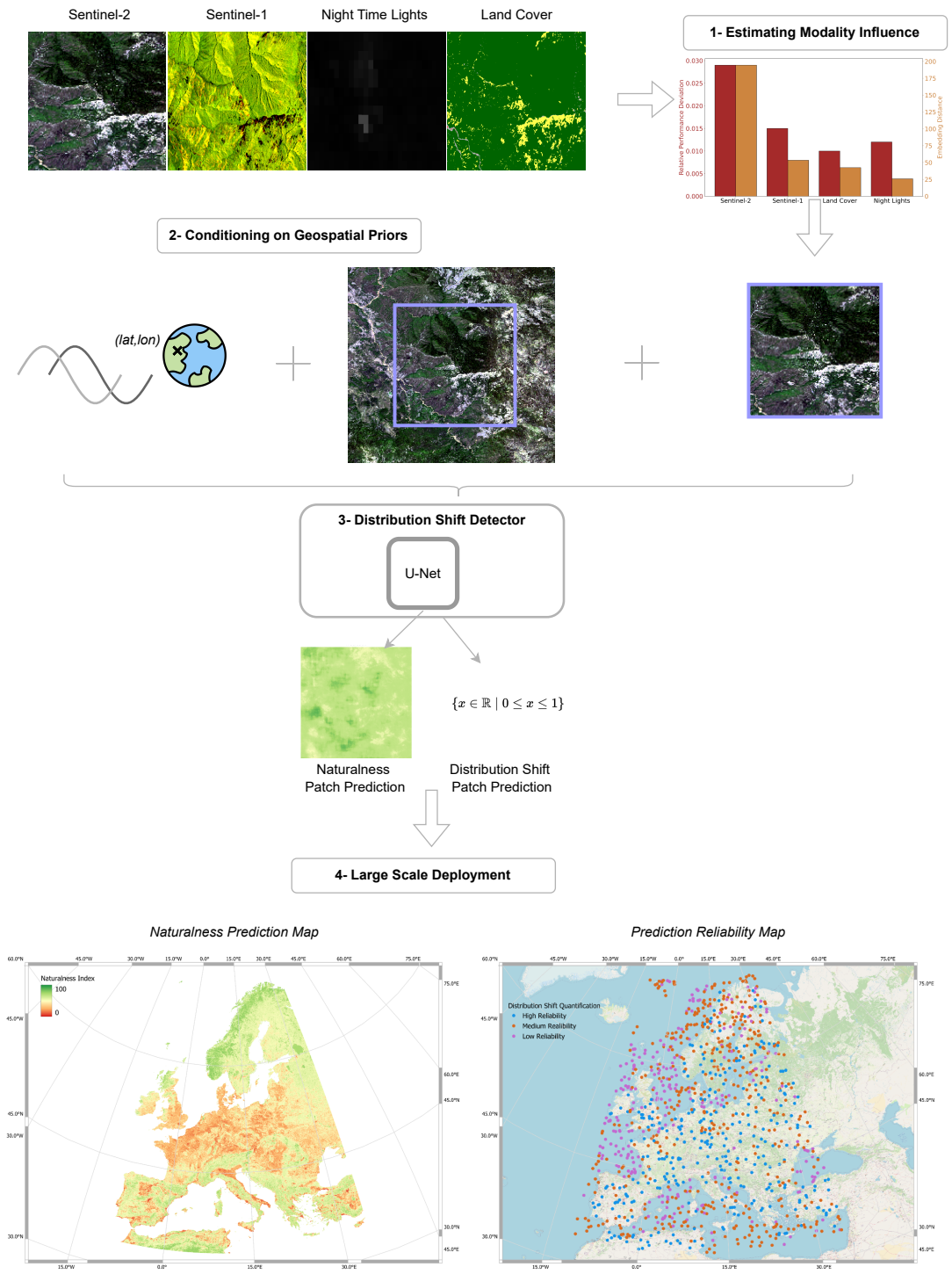


Figure 1.1: The workflow starts with the relative influence estimation of each modality on naturalness prediction. The most informative modality then guides the design of a geospatially aware model integrating coordinates and long-range context. The model is then coupled with an out-of-distribution detector that quantifies distribution shift for each inference input. Deployed at a continental scale, presenting the *Naturalness Map of Europe* and its companion *Prediction Reliability Map*. The dataset sample (ID: 11115171) is a patch covering the Yuki Wilderness area, California, and can be viewed on the WDPA via www.protectedplanet.net/11115171.

- Chapter 5 (Deep Learning Methodology for Naturalness Mapping): Presents the core deep learning framework developed in this thesis. It details the three main methodological components: first, an interpretable machine learning method for influential input modality selection; second, the design of a geospatially-aware model architecture incorporating spatial correlation modeling and location encoding; and finally, a post-hoc method for out-of-distribution detection.

- Chapter 6 (Experiments and Results): Presents the experimental validation for the methodologies introduced in the previous chapter. It details the experimental setups, presents the quantitative and qualitative results, and provides a detailed discussion of the findings for each contribution.

- Chapter 7 (Naturalness Map of Europe): Showcases the practical application of the complete framework. It details the large scale deployment pipeline and presents the two final map products: the continental-scale *Naturalness Map of Europe* and its accompanying *Reliability Map*. The chapter concludes with comprehensive regional evaluations of the framework's performance.

- Chapter 8 (Discussion): Provides a broader, holistic discussion of the thesis. It synthesizes the key findings, reflects on the implications and limitations of the developed methods and the produced map, and discusses the overarching themes of the work.

- Chapter 9 (Conclusion and Perspectives): Concludes the thesis. It provides a summary of the work, suggests promising directions for future research, and includes a statement on ethical considerations.

Chapter 2

Theoretical Background

This chapter establishes the theoretical foundation for the methods and data central to this thesis. As the thesis exists at the intersection of Earth Observation and Machine Learning, the chapter is accordingly organized into two primary sections, providing the necessary background in each domain. The first section provides an overview of the primary data types used in modern environmental monitoring. It details the operational principles and key characteristics of satellite missions such as Sentinel-1 and Sentinel-2, as well as foundational geospatial products like land cover maps and digital elevation models. The second section introduces the fundamental computational concepts. It begins with the principles of supervised learning and regression, then narrows its focus to Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs). The section culminates in a discussion of Geospatial Deep Learning, a subfield dedicated to adapting these models to the unique properties of geographic data, such as location and spatial context.

2.1 Sensors and Geospatial Data Sources

This section introduces the key sensors and geospatial data sources used in this work, all of which are illustrated in Figure 2.1.

2.1.1 Sentinel-1

The Sentinel-1 mission, operated by the European Space Agency (ESA) as part of the Copernicus programme, is designed as a multi-satellite constellation in a near-polar, sun-synchronous orbit [5]. Each satellite is equipped with a C-band ($\sim$5.405 GHz) Synthetic Aperture Radar (SAR) instrument. The primary advantage of SAR is its ability to acquire imagery day and night and under all weather conditions, as the radar signal penetrates cloud cover, making it exceptionally reliable for consistent environmental monitoring. In its nominal two-satellite configuration, the mission achieves a revisit time of 6 days at the equator, with more frequent coverage at higher latitudes.

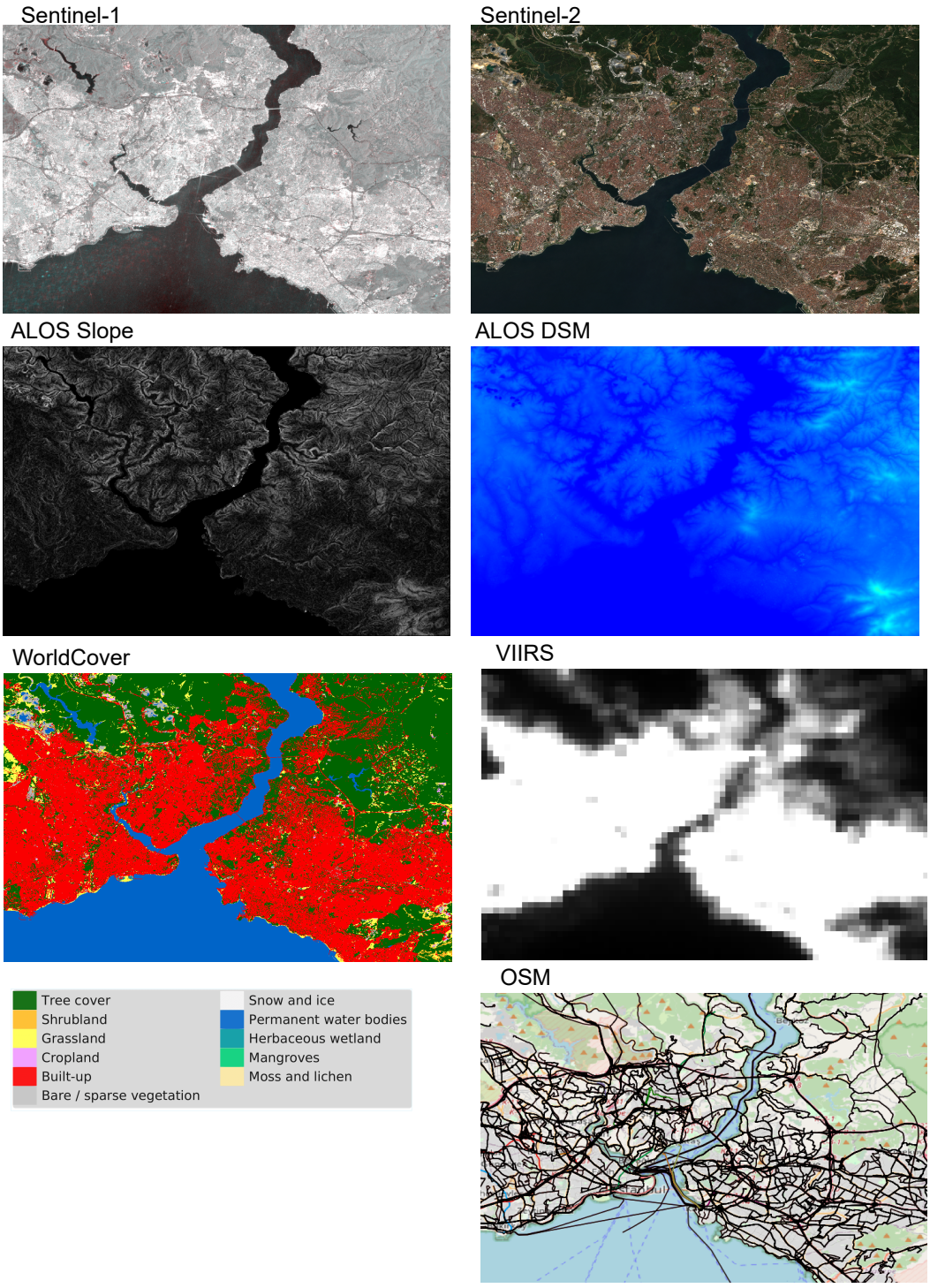


Figure 2.1: A visualization of various geospatial data over Istanbul metropolitan area, Türkiye. Panels display: **Sentinel-1** C-band Synthetic Aperture Radar (SAR) mean intensity (VV polarization); **Sentinel-2** Multi-Spectral Instrument (MSI) true-color composite (Bands 4, 3, 2); **ALOS Slope** derived from the ALOS DSM; **ALOS DSM** (JAXA ALOS World 3D - AW3D30); **WorldCover** 10m land cover classification (ESA WorldCover 2021); **VIIRS** Day/Night Band (DNB) average monthly nighttime radiance (May 2023); and the **OSM** roads and railways. *Base map data copyrighted OpenStreetMap [4] contributors and available from* `https://www.openstreetmap.org`.

The sensor can transmit and receive in both horizontal (H) and vertical (V) polarizations, enabling different acquisition schemes. These include single polarization (VV or HH) and dual polarization (VV+VH or HH+HV). The choice of polarization affects the information captured about the physical properties of the surface; for instance, dual-polarization data can enhance the discrimination of different land cover types. Sentinel-1 data are distributed as two main product types. Specifically, these are Single Look Complex (SLC) products, which preserve both phase and amplitude information, and Ground Range Detected (GRD) products, which provide multi-looked intensity data projected to the ground range.

2.1.2 Sentinel-2

The Copernicus Sentinel-2 mission is a multi-satellite constellation in a sun-synchronous orbit, designed to provide high-resolution, multi-spectral optical imagery for land monitoring [6]. With two satellites operating in tandem, the mission achieves a high revisit frequency of five days at the equator, which increases at mid-latitudes. Each satellite carries a MultiSpectral Instrument (MSI) that acquires data in a push-broom configuration over a wide 290 km swath. The MSI sensor captures information across 13 spectral bands provided at different native spatial resolutions:

- 10-meter resolution: Four bands, including the conventional Red, Green, and Blue (RGB) bands, as well as a broad Near-Infrared band. These are primarily used for land cover classification and vegetation analysis.
- 20-meter resolution: Six bands, including four narrow bands in the red-edge region of the Near-Infrared spectrum, which are particularly sensitive to the state of vegetation health, and two wider Shortwave Infrared bands, useful for detecting snow, ice, and assessing vegetation moisture content.
- 60-meter resolution: Three bands primarily intended for atmospheric correction and cloud screening (e.g., aerosol, water vapour, and cirrus detection).

Sentinel-2 data is processed and distributed at different levels. The Level-1C (L1C) product provides Top-Of-Atmosphere reflectance, ortho-rectified and geometrically corrected. The Level-2A (L2A) product provides Bottom-Of-Atmosphere surface reflectance, having been atmospherically corrected from the L1C data. L2A products are considered analysis-ready data and are the standard input for most terrestrial applications. The data is organized into a global tiling grid system based on the Military Grid Reference System, with each tile covering an area of approximately 100 $\times$ 100 km.

2.1.3 NASA VIIRS

The Visible Infrared Imaging Radiometer Suite (VIIRS) is a key instrument aboard the Suomi National Polar-orbiting Partnership and subsequent Joint Polar Satellite System satellites, operated by NASA and NOAA [7]. As a whisk-broom scanning radiometer, VIIRS provides global daily observations by acquiring data over a very wide swath of approximately 3000 km. The instrument collects data across 22 spectral bands, spanning the visible to the longwave infrared spectrum (0.41 to 12.5 µm). These bands are designed to support a wide range of applications, including the monitoring of clouds, aerosols, sea surface temperature, vegetation health, and active fires. The data is provided at two native spatial resolutions at nadir:

- Imagery Bands (I-bands): Five high-resolution bands providing approximately 375-meter resolution, primarily used for detailed imaging of land and cloud features.
- Moderate-Resolution Bands (M-bands): Sixteen bands providing approximately 750-meter resolution, used for quantitative measurements of various environmental parameters.

A unique and widely utilized component of VIIRS is its panchromatic Day/Night Band (DNB). This highly sensitive sensor is capable of capturing detailed imagery of the Earth's surface at night, detecting low levels of visible light radiance from sources such as city lights, gas flares, and wildfires. A key feature of the DNB is its near-constant spatial resolution of approximately 742 meters across the entire scan swath, which avoids the significant pixel distortion common in other wide-swath sensors. Processed DNB data, such as NASA's Black Marble product suite, provide robust, high-quality nighttime lights imagery that serves as a powerful proxy for human activity, infrastructure, and socioeconomic patterns.

2.1.4 ESA WorldCover

The European Space Agency (ESA) WorldCover product is a global land cover map produced at a 10-meter spatial resolution for a specific reference year (e.g., 2020, 2021). It was developed as part of the ESA WorldCover project to provide a new baseline global land cover product that is more accurate and detailed than previous generations of maps [8]. The product is derived from a combination of both Sentinel-1 C-band Synthetic Aperture Radar (SAR) data and Sentinel-2 multispectral optical imagery. The integration of these two distinct sensor types ensures that the classification is robust across different geographic regions and seasons, leveraging the structural information from SAR and the spectral information from optical data. The classification algorithm is based on a large scale machine learning system trained on a massive volume of reference

data. The WorldCover map provides a classification of the Earth's land surface into 11 distinct, discrete classes.

Each pixel in the product is assigned one of these land cover labels. It is designed to be easily usable and consistent, providing a valuable resource for large scale environmental monitoring, land management, food security, and climate modeling. However, like all large scale, automatically generated map products, it contains inherent classification errors and uncertainties. Comparative studies have assessed its performance; for instance, Venter et al. (2022) [9] report a global overall accuracy of approximately 65.5% for the WorldCover map.

2.1.5 Elevation Models

Digital Elevation Models (DEMs) are fundamental geospatial datasets that represent the Earth's topography. Within the literature, a distinction is strictly made between a Digital Surface Model (DSM), which captures the elevation of the topmost surfaces including vegetation canopies and human-made structures, and a Digital Terrain Model (DTM), which represents the bare-earth topography. These grids are essential for a wide range of environmental analyses, allowing for the quantification of terrain characteristics such as slope, aspect, and ruggedness. These derived metrics are crucial for applications in hydrology, geology, land use planning, modeling physical accessibility, and characterizing landscapes. Here are two widely used global DEM options, both of which predominantly function as surface models:

NASA SRTM Digital Elevation Model (30 m) The Shuttle Radar Topography Mission (SRTM), a joint project between NASA and the National Geospatial-Intelligence Agency, collected data over 11 days in February 2000. It used C-band radar interferometry to generate the most complete, high-resolution digital topographic database of the Earth at the time [10]. Because C-band radar only partially penetrates dense vegetation and reflects off built structures, the SRTM product essentially acts as a DSM in forested and urban areas. The most commonly used version has a spatial resolution of 1 arc-second (approximately 30 meters at the equator) and covers most of the Earth's land surface between 60° North and 56° South latitude. Due to its consistency and broad availability, the SRTM dataset has become a benchmark for global-scale topographic analysis.

JAXA ALOS AW3D30 Version 3.2 The AW3D30 is a global digital surface model produced by the Japan Aerospace Exploration Agency (JAXA). It is derived from stereo-pairs of optical imagery captured by the Panchromatic Remote-sensing Instrument for Stereo Mapping aboard the Advanced Land Observing Satellite (ALOS) [11]. The AW3D30 product provides a spatial resolution of 1 arc-second (approximately 30 meters). A key

advantage of the ALOS dataset is its full global coverage, including the high-latitude regions and mountainous areas where the SRTM mission had data voids or lower quality.

2.1.6 OpenStreetMap

OpenStreetMap (OSM) is a global, collaborative project to create a free, editable map of the world [4]. It is a prominent example of Volunteered Geographic Information, where a global community of contributors, ranging from hobbyists to GIS professionals, collectively build and maintain the dataset. This crowd-sourced approach allows for the creation of a highly detailed and often more up-to-date map than many commercial or official alternatives, particularly in regions that may be overlooked by traditional mapping agencies. The data within OSM is structured using three fundamental elements: nodes, ways, and relations. The semantic meaning of these elements is defined by a flexible, community-driven tagging system. Users attach key-value pairs (e.g., `highway=motorway` or `building=yes`) to describe the attributes of a feature. A key characteristic of OSM is its spatial heterogeneity. The completeness, accuracy, and level of detail can vary significantly across the globe, often correlating with population density and the presence of an active local mapping community [12, 13].

2.2 Machine Learning and Deep Neural Networks

Machine learning (ML) is a field of artificial intelligence focused on developing algorithms that enable computers to learn patterns and make predictions from data, rather than being explicitly programmed for a specific task. These algorithms are typically categorized into several learning paradigms, primarily supervised, unsupervised, and reinforcement learning. Traditional machine learning approaches often work well with smaller, structured datasets and typically require feature selection, where domain expertise is used to engineer informative input variables. In contrast, the subfield of deep learning utilizes layered neural networks to automatically learn feature hierarchies from large, unstructured data like satellite imagery, albeit with greater computational requirements. A key advantage of many traditional machine learning models, however, is their interpretability. This refers to the degree to which a human can understand the cause of a decision made by the model. Interpretability is a desirable property as it allows for the identification of potential biases and spurious correlations, while also revealing underlying data relationships that can lead to new insights.

The supervised learning framework is central to many data-driven applications. In this paradigm, the goal is to learn a mapping function $f : X \to Y$ from an input space X to an output space Y, based on a training dataset of input-output pairs $\mathcal{D} = \{(x^{(k)}, y^{(k)})\}_{k=1}^{M}$. The theoretical objective is to determine the optimal function f from a hypothesis space $\mathcal{F}$ by minimizing the true risk (or expected generalization error) $\mathcal{R}(f)$ over the true,

unknown data distribution. Because this true distribution is inaccessible, the objective is approximated by minimizing the empirical risk (or average loss) $\mathcal{L}(\theta)$ over the finite training data, where the function f is parameterized by the weights θ:

$$\min_{f \in \mathcal{F}} \mathcal{R}(f) \approx \min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \frac{1}{M} \sum_{k=1}^{M} \ell(f(x^{(k)}; \theta), y^{(k)})$$

where $\ell(\cdot, \cdot)$ is a chosen loss function quantifying the discrepancy between the prediction and the true label. Common supervised tasks in computer vision include:

- Classification: The output is a discrete category, such as assigning a single class label to an entire image.
- Object Detection: The goal is to identify and localize multiple objects within an image by predicting bounding boxes and class labels for each.
- Semantic Segmentation: This involves assigning a class label to every pixel in an image, creating a dense, pixel-wise classification map. For an input image $I \in \mathbb{R}^{H \times W \times C}$, a neural network typically outputs an intermediate probability tensor of size $H \times W \times K$. The final output map $\hat{Y}_{seg}$ is obtained by taking the argmax across the class dimension, resulting in a single-channel matrix where each pixel is assigned a discrete integer representing one of the K classes: $\hat{Y}_{seg} \in \{1, \ldots, K\}^{H \times W \times 1}$.
- Regression: The output is a continuous value. Conceptually, regression seeks to model the associative relationship between the input features and a continuous target variable. In a computer vision context, this can involve predicting a single value for an entire image or, more relevant to dense prediction, performing image-to-image regression. In this dense format, the model learns a mapping from an input image to an output image where each pixel contains a continuous value, producing a quantitative map: $\hat{Y}_{reg} \in \mathbb{R}^{H \times W \times 1}$.

2.2.1 Deep Learning and Convolutional Neural Networks

Deep Learning (DL) is a subfield of machine learning that utilizes Deep Neural Networks (DNNs)—hereafter referred to simply as *models* when the context is clear—to learn complex patterns from large datasets. DNNs consist of multiple layers of interconnected nodes, or artificial neurons, arranged in a hierarchy. Each layer transforms the input from the preceding layer using a set of learned weights and a non-linear activation function. This layered architecture allows the network to learn hierarchical feature representations, from simple patterns in the initial layers to abstract, high-level concepts in the deeper layers.

The process of training a DNN involves optimizing its parameters (weights and biases) to minimize a loss function, which quantifies the discrepancy between the model's

predictions and the ground-truth targets. Specifically, given a training dataset $\mathcal{D} = \{(x^{(k)}, y^{(k)})\}_{k=1}^{M}$ where $x^{(k)}$ is the input and $y^{(k)}$ is the annotation, the goal is to find the optimal set of parameters θ (weights and biases) that minimizes the empirical risk (or average loss) $\mathcal{L}(\theta)$:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ \mathcal{L}(\theta) = \underset{\theta}{\operatorname{argmin}}\ \frac{1}{M}\sum_{k=1}^{M} \ell(f(x^{(k)};\theta), y^{(k)})$$

where $f(x^{(k)};\theta)$ is the model's prediction and $\ell(\cdot,\cdot)$ is the loss function (e.g., Cross-Entropy or Mean Squared Error). This optimization is typically performed using back-propagation to compute the gradient of the loss with respect to the parameters, $\nabla_\theta \mathcal{L}(\theta)$, and a gradient-based optimization algorithm to iteratively update θ.

For processing data with a grid-like topology, such as images, a specialized type of DNN known as a Convolutional Neural Network (CNN) is particularly effective. The design of CNNs is inspired by the human visual cortex. Their core building block is the convolutional layer. The central operation is the discrete 2D operation, which is technically a cross-correlation but is conventionally termed "convolution" in the deep learning literature. The operation of an input feature map I with a kernel K of size $M_K \times N_K$ to produce an output feature map O is formally defined as:

$$O(i,j) = (I \circledast K)(i,j) = \sum_{m=0}^{M_K-1}\sum_{n=0}^{N_K-1} I(i+m, j+n)K(m,n)$$

This operation has two key properties that make it highly efficient for image data: 1) sparse interactions, where each output unit is connected only to a small local patch of the input via the filter, and 2) parameter sharing, where the same filter is used across the entire input, making the network equivariant to translations of features.

The pioneering LeNet-5 architecture (Figure 2.2), introduced by LeCun et al., is a classic example that introduced a sequence of alternating convolutional and pooling layers, followed by fully-connected dense layers for tasks like handwritten digit recognition [14].

CNNs are designed to incorporate spatial representation through mechanisms like local connectivity and weight sharing [15, 16]. However, these same design principles introduce limitations when dealing with the complexities of geospatial data. While a deep CNN's theoretical receptive field may be large, research has shown that its effective receptive field (ERF)—the area that truly influences an output pixel—is significantly smaller and heavily biased towards the center of the patch [17]. The distinction between theoretical and effective receptive fields highlights a broader conceptual tension within convolutional architectures: while they are theoretically capable of integrating information over large spatial extents, in practice they are most responsive to local neighborhoods. This reliance on purely local features hinders the model's ability to capture long-range dependencies and increases the risk of learning spurious correlations [18].

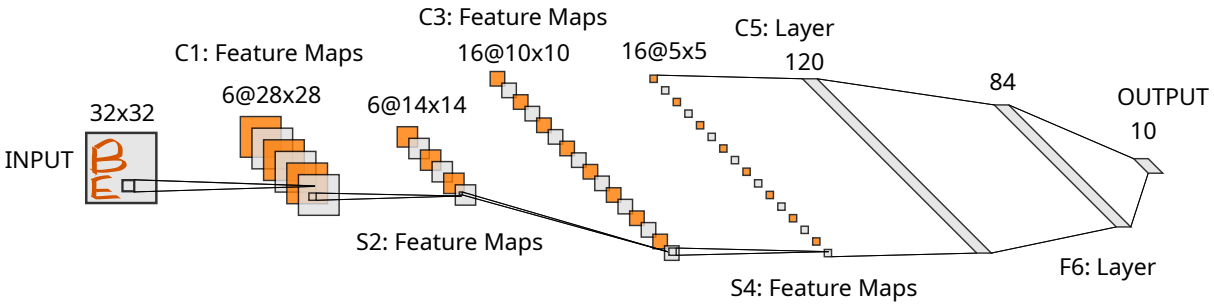


Figure 2.2: The LeNet-5 Architecture [14]. This Convolutional Neural Network (CNN) for digits recognition demonstrates the sequential processing of an input image. The convolution steps illustrate the local connectivity of the filters and the downsampling effect of pooling operations at different stages. Specifically, it shows how a local input region contributes to a feature in the C1 map, and how regions in feature maps are aggregated during pooling. The network progresses through alternating convolutional (C1, C3) and pooling (S2, S4) layers, followed by fully-connected dense layers (C5, F6) and an output layer for classification.

2.2.2 Geospatial Deep Learning

Satellite data constitutes a distinct modality characterized by unique challenges that differentiate it from standard computer vision domains. First, the data captures phenomena across vast spatial and temporal scales, where targets range logarithmically from meter-level objects to kilometer-wide landforms, and temporal patterns evolve from daily fluctuations to decadal climate shifts [19]. This complexity is compounded by the spectral channels available, as satellite sensors record diverse wavelengths beyond the visible spectrum—such as infrared and synthetic aperture radar—often at high radiometric resolutions that standard image formats cannot accommodate. Furthermore, the sheer data volume presents a computational bottleneck, involving petabyte-scale archives of continuous global coverage that far exceed the size of traditional machine learning datasets. Conversely, reliable annotations are scarce; ground-truth labels are often sparse, and geographically biased. Despite these fundamental distinctions, a significant portion of current research continues to apply models designed for natural images to well-resolved remote sensing tasks, rather than addressing the modality's specific constraints [20]. By failing to account for these unique characteristics and deployment demands, such approaches frequently compromise scalability and predictive performance, ultimately stalling the development of specialized methodologies required for this domain [21].

The modeling objective in this domain is to learn from georeferenced data within a continuous spatial field, requiring an understanding of how spatial context and geographic position jointly determine predictive outcomes. To achieve this, the field benefits significantly from methods designed to address two fundamental geographic principles that generic DL models are ill-equipped to handle. First, the principle of Spatial Autocorrelation dictates that the prediction at any location is statistically dependent on its broader

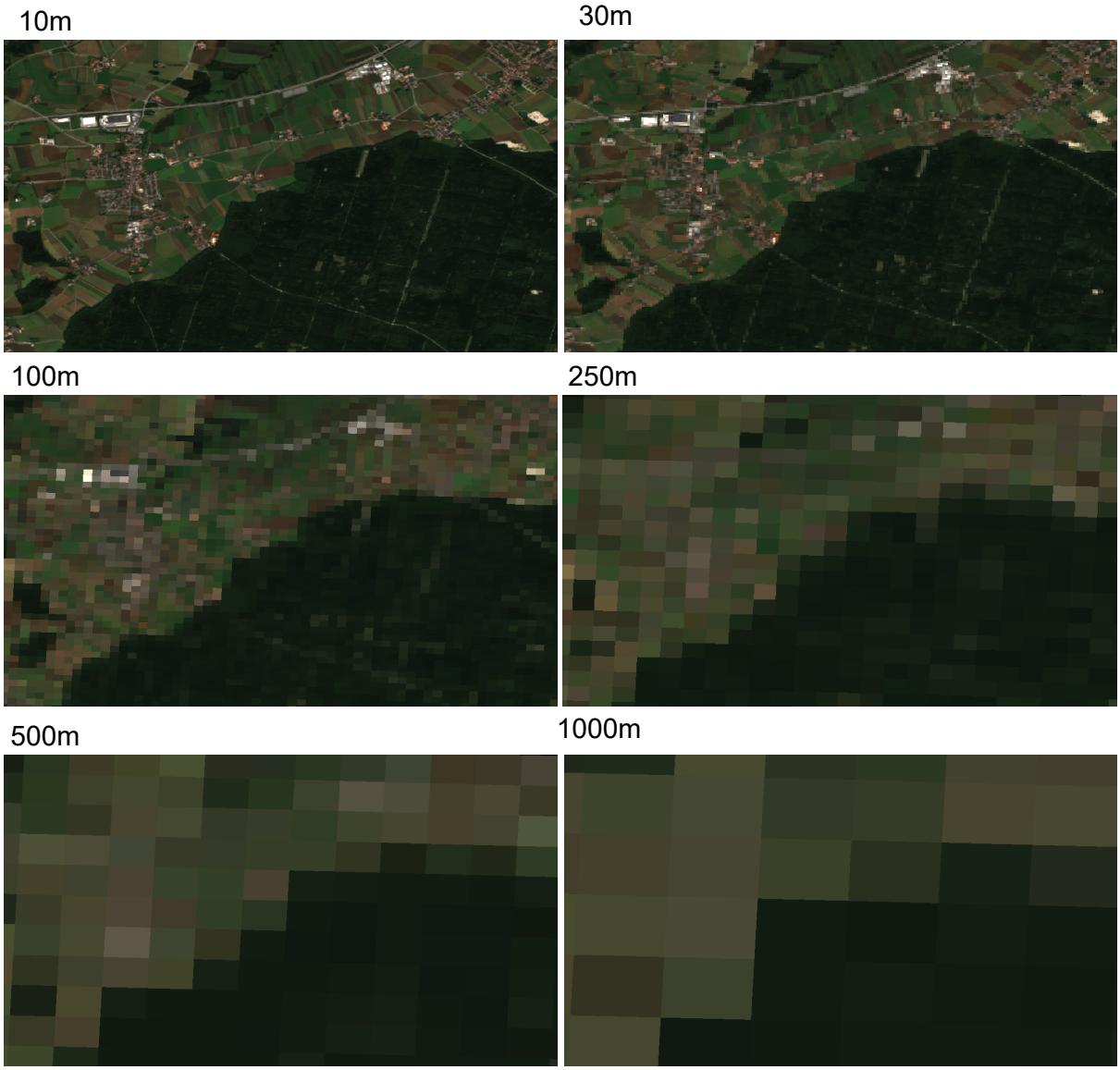


Figure 2.3: How spatial resolution determines the level of observable detail in satellite imagery? This figure shows Sentinel-2 true-color composite over Ebersberger Forest, Germany. The native 10 m resolution image is progressively downsampled to coarser spatial scales: 30 m, 100 m, 250 m, 500 m, and 1000 m. As resolution decreases, fine-scale landscape features such as forest clearings and built-up areas become increasingly indistinguishable.

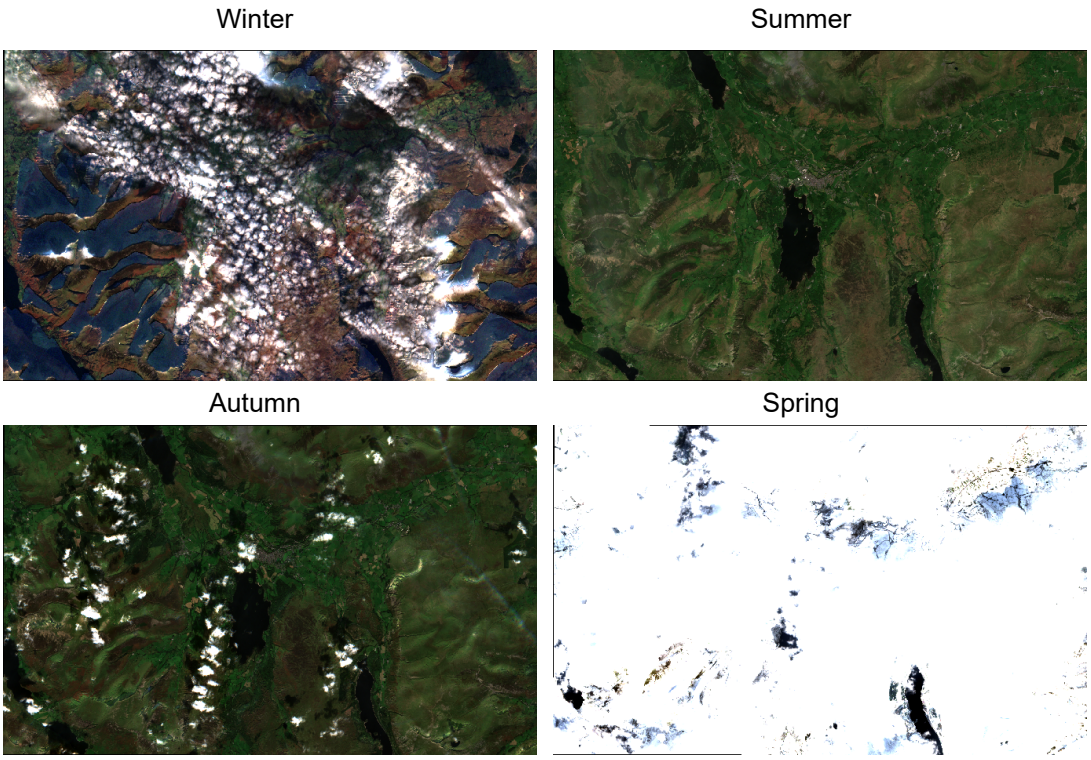


Figure 2.4: This figure illustrates how temporal variability affects the spectral and visual appearance of the same location across seasons. Sentinel-2 true-color composites (Bands 4, 3, 2) over the Lake District National Park, United Kingdom, capture the least cloudy single-date image from each meteorological season in 2023: (a) Winter, (b) Spring, (c) Summer, and (d) Autumn.

neighborhood. This demand for long-range spatial dependency, as captured by Tobler's First Law, "Everything is related to everything else, but near things are more related than distant things", remains a structural limitation for standard Convolutional Neural Networks due to their restricted Effective Receptive Field (ERF). Current geospatial DL approaches do not solve this architectural problem but rather mitigate it by introducing compensatory mechanisms such as self-attention or large-kernel convolutions to implicitly capture broader context. Second, the principle of spatial non-stationarity implies that the relationship between input features and the target variable varies across the map. The same spectral signature may have a different meaning depending on its latitude or regional climate (e.g., a forest's spectral signature in northern Europe differs from that in the tropics owing to variations in phenology, illumination, and atmospheric conditions). Consequently, geospatial DL needs to move beyond the assumption of translation equivariance—a core inductive bias of standard CNNs—by effectively integrating geolocation as an explicit prior to account for this non-stationarity. Incorporating this awareness requires designing methods that can explicitly reason across scales, integrate contextual cues, and produce results that are not only visually discriminative but also geographically consistent with the underlying spatial logic of the geographic domain. Importantly, geospatial deep learning does not seek to replace conventional architectures, but to adapt and extend them to better handle the geographic nature of satellite data. To illustrate these considerations, Figures 2.3–2.5 demonstrate how spatial resolution, temporal variability, and spatial context fundamentally affect the interpretation of remote sensing imagery and, consequently, model design.

Figure 2.3 shows a Sentinel-2 true-color composite progressively resampled to coarser spatial resolutions (10 m to 1000 m). This highlights how spatial resolution defines the level of observable detail — fine-scale patterns such as forest clearings and residential structures become indistinguishable as the pixel size increases. Spatial resolution therefore directly constrains the granularity of features a model can learn and the types of geospatial processes it can represent.

Figure 2.4 demonstrates temporal variability using Sentinel-2 imagery across four seasons in 2023. Each panel represents the least-cloudy, single-date acquisition for that season, emphasizing how seasonality induces strong spectral and visual changes even for the same location. Temporal mosaicking — the process of compositing multiple images to minimize cloud cover and temporal gaps — is often used to mitigate noise and provide a more stable input for models. However, such mosaics can obscure short-term dynamics, whereas single-date images preserve temporal specificity. Understanding this trade-off is crucial for designing representative and robust geospatial learning pipelines.

Finally, Figure 2.5 illustrates the role of spatial context on model perception. Centered on a point within Nairobi National Park, Kenya, Sentinel-2 imagery is shown at increasing input sizes (128×128 to 1024×1024 pixels). Smaller patches capture the

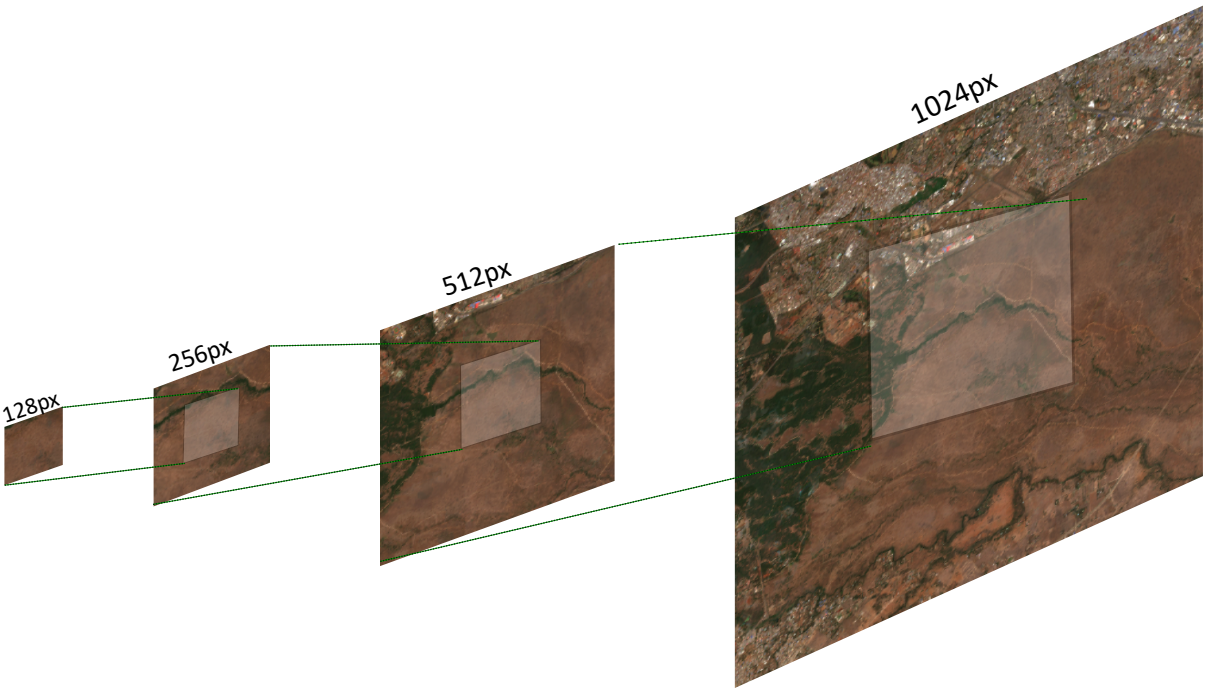


Figure 2.5: This figure shows how increasing the input patch size expands the effective spatial context available for scene understanding. Centered on a point within Nairobi National Park, Kenya, Sentinel-2 surface reflectance composites (June–September 2023) are displayed at progressively larger input scales (128×128 to 1024×1024 pixels). Smaller patches capture natural areas within the park, while larger inputs increasingly include surrounding regions and the urban fabric of Nairobi to the north, illustrating how spatial context enriches scene understanding.

immediate natural environment within the park, largely unaffected by human activity, while larger inputs progressively include surrounding regions and ultimately the dense urban landscape of Nairobi city. Since most CNN-based models operate on fixed-size patches, their receptive field defines the effective spatial context they can learn from. Yet, in geospatial analysis, understanding a location often depends on its surroundings — exemplifying the importance of explicitly modeling spatial autocorrelation and context beyond the cropped input patch.

Overall, these examples underscore that applying DL to EO requires choices that reflect the distinct nature of geospatial data. Robust geospatial models must integrate information across scales (Figure 2.3), account for temporal dynamics (Figure 2.4), and capture contextual relationships (Figure 2.5).

2.3 Human Influence

Domesticating the Earth, a project spanning roughly 12,000 years, has yielded significant advantages for human civilization. Yet, this process has simultaneously exacted a severe toll, manifesting as an altered climate, the erosion of biodiversity, and destabilized ecosystems [22]. The effort to understand human disturbance has deep roots, famously articulated by George Marsh in his seminal 1864 book, *Man and Nature*, with the question, "[...] to what degree are the processes of nature threatened by human activity?" (quoted in [23]). A primary framework for answering this question in the modern era is the concept of the human footprint, which represents the sum total of human ecological impact. It is expressed not as a single number, but as a continuum of human influence stretched across the land surface, revealing the major patterns of human impact on nature [24]. This continuum of influence is constructed from proxies that represent the primary, direct human activities posing the greatest near-term threats to most ecosystems. These include the transformation of land for agriculture or urban development; the direct and indirect effects of roads and the access they afford; and the development of power infrastructure that enables further land transformation. While this is by no means a complete catalog of all conservation challenges, this human footprint continuum provides a foundational basis for understanding conservation efforts [1].

The core methodology involves the heuristic combination of these individual drivers. Technically, the Human Footprint (HF) score for any given pixel is calculated as the additive summation of scores derived from multiple input layers (S_i), where N is the total number of pressure layers:

$$\mathrm{HF} = \sum_{i=1}^{N} S_i$$

This summation is performed under the premise that pressures are not mutually exclusive, and their co-occurrence identifies the greatest human impact, resulting in a maximum

possible score (e.g., 1, 40, 60, or 100 in various indices) [25]. Each pressure layer score, S_i, is first normalized to a common scale, typically 0 to 10, based on its potential relative impact. For access layers like roads or railways, the score is determined by a distance-decay function, often following an exponential or logarithmic relationship, where pressure is maximal near the feature and exponentially decreases as the distance (D) increases. For example, population pressure scores often use a logarithmic scaling function, such as:

$$\text{Pressure Score} = 3.106 \times \log(\text{Population Density} + 1)$$

Conversely, for static, area-based layers such as built environments or croplands, the score is often assigned based on a binary or fractional threshold (e.g., $S_i = 10$ if the urban fraction exceeds 20%, otherwise $S_i = 4$). Traditionally, these input layers encompass seven to eight human pressure variables, typically including built environments, population density, nighttime lights, croplands, pasture lands, roadways, railways, and navigable waterways. These layers are then overlaid and summed according to these pre-assigned weightings and influence distances (as visually summarized in Figure 2.6).

Despite this utility, the interpretation of such measures requires acknowledging several inherent shortcomings. It is crucial to recognize that these indices measure human *influence, not necessarily negative impact*. The drivers themselves, such as population density or access, are not inevitably harmful, and thoughtful practices can allow for coexistence [27]. The measure of anthropogenic pressure, therefore, is better understood as indicating areas where humans have a greater responsibility for biological outcomes, rather than simply quantifying damage. Furthermore, influence is a relative concept; the extremes of the index are theoretical constructs. Even in places where the index is low or equals zero, there remains human influence from climate change or atmospheric deposition. Similarly, even in the most influenced places, such as midtown Manhattan, natural forces are in play: rain falls, winds blow, and plants and animals recruit, grow, and die. Finally, our ability to interpret these patterns is constrained by the sheer complexity of human-nature interactions. As previously noted, the global understanding of human environmental impact is still in its infancy, and all such measures should be considered with caution [28]. However, while knowledge may be incomplete, it is often sufficient to recognize the need for concern and action.

A key critique of early human influence models is their simplistic treatment of the land surface as a "blank slate" upon which human activity is inscribed. This overlooks the crucial fact that the biological outcomes of human influence are not uniform, but are instead shaped by the specific ecosystem types and historical contexts of each region [29]. For example, an absolute human influence score in the temperate broadleaf forests of North America has a vastly different ecological meaning and biological impact than the same score in the tropical rainforests of Africa. This complexity arises because these maps operate on the general premise that the geography of human influence is roughly

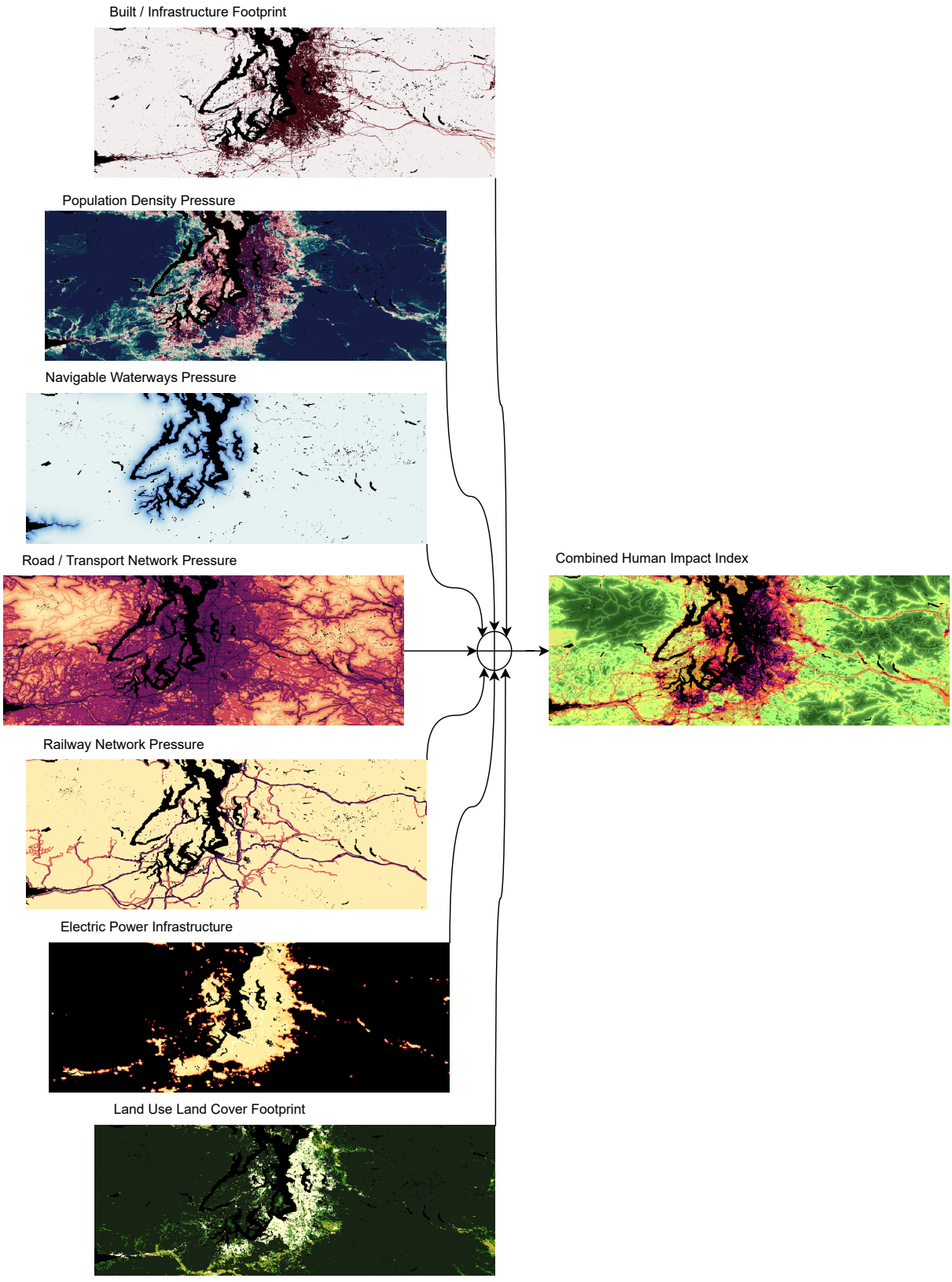


Figure 2.6: Visualization of the individual drivers contributing to the Human Impact Index [26] for the Puget Sound region, United States, and its surroundings. Each panel on the left represents one of the key anthropogenic drivers: *Built and Infrastructure*, *Population Density*, *Navigable Waterways Pressure*, *Road and Transport Network Pressure*, *Railway Network*, *Electric Power Infrastructure*, and *Land Use and Land Cover*. The panel on the right shows the *Combined Human Impact Index*, representing the cumulative human footprint.

the inverse of the geography of natural processes. The expectation is that where human influence values are highest, ecosystems will be most modified, and where values are lower, one can expect more intact and functional natural communities. However, this inverse relationship is not a simple one-to-one mapping. As noted by Redford and Richter (1999) [30], the exact consequences of human influence are complicated and depend on the history of a place, the types of current influence, and the specific aspects of nature being considered. Yet, despite this complexity, it is in these "wildest places" with the lowest measured human footprint that the greatest opportunity to conserve the full range of nature still exists.

Additionally, the heuristic combination of human influence drivers is itself a simplification of reality. This approach assumes that a finite set of proxies can fully represent the complex, multi-faceted nature of anthropogenic influence, which is a concept that may be impossible to capture in its entirety. Furthermore, it often treats each driver's impact as uniform, when in reality, their effects can be vastly different; for example, a road has a direct physical footprint but also creates profound indirect effects by enabling access deep into a landscape.

These limitations, however, do not imply a fixed or predetermined fate. The core premise of sustainability science is that by understanding the effects we are having on the planet, we can learn to mitigate the less desirable ones through better planning, practice, and a commitment to applying that knowledge. For society to successfully shape the human footprint in a way that is consistent with the needs of both current and future generations, it is clear that more must be done to provide for people and nature simultaneously.

Chapter 3

State of the Art

This chapter provides a review of the research relevant to this thesis, organized according to the core research pillars of the thesis. It begins by surveying the field of large scale human influence mapping, examining the evolution from traditional rule-based methods to modern deep learning-derived approaches. Following this, it delves into the domain of interpretable machine learning, focusing on the specific task of feature attribution, which aims to explain a model's behavior. The review then explores adaptations that have been proposed in the literature to make deep learning models more inherently *geospatial*, considering contextual and geographic priors. Subsequently, it covers the topic of distribution shift and out-of-distribution (OOD) detection in the context of Earth Observation. Finally, building upon the surveyed literature, the specific contributions of this thesis are detailed, positioning them within the context of the identified research gaps.

3.1 Paradigms in Human Influence Mapping from Space

3.1.1 Rule-Based Mapping

Quantifying the extent of human impact on the environment has been a long-standing goal, leading to the development of rule-based mapping to monitor anthropogenic pressure. Traditional approaches build on hand-crafted features, involving heuristics about what constitutes pressure on the environment. The goal of these methods has been to objectify the determination of whether an area is "natural" or under strong human influence, a crucial task for conservation planning and policy-making.

Known early efforts include an inventory by McCloskey and Spalding [22]. This study represented a first attempt to map and quantify large blocks of "wild land", seeking to answer how much of the planet's surface still is predominantly influenced by the forces of nature. In the 1990s, owing to rapid advances in remote sensing technology and geographic information systems (GIS), these early inventory surveys were progressively

replaced by digital, data-driven techniques that could integrate multiple geospatial data sources.

A foundational example of this shift is the Human Influence Index (HII), or "human footprint", proposed by Sanderson et al. [1]. To quantify human influence, they integrated and scored four types of geospatial data, or drivers: population density, land transformation, accessibility, and electrical power infrastructure. This approach became a cornerstone, but it was noted for overlooking key pressures such as the destruction of forests and grasslands (i.e., assigning a value of 0 to these areas), which led to potential overestimations of wilderness [1].

The human footprint maps have proven invaluable in monitoring biodiversity [31], assessing habitat fragmentation, and policymaking [32]. Human pressure maps have become an important tool for spatial planning [33] and land use management [34], monitoring the extent of human influence [35], identifying critical remaining intact habitat [32, 36], and predicting how increasing intensity of human–environment interactions affects pathogen transmission, including vector-borne diseases [37].

In the 21st century, research on human influence has undergone significant methodological and data quality advancements. Many subsequent studies followed, continually building toward higher spatial resolution and temporal scales by incorporating updated or novel human pressure drivers (e.g., navigable waterways and night light data) for the first time.

The concept of intactness was further formalized by Allan et al. [2], who produced the first temporally inter-comparable global maps of wilderness for the years 1993 and 2009 using the Human Footprint method. Their work quantified the global erosion of nature, demonstrating that a vast area of wilderness was lost during that period. Such calls for better conservation efforts were repeated by Kennedy et al. (2019) [3], who created a Global Human Modification gradient for 2016 from 13 stressors at a 1 km resolution, finding that over half of the world's ecoregions (52%) are moderately modified.

Theobald et al. (2020) [38] followed with a temporal analysis quantifying the loss of 1.6 million km^2 of natural land between 1990 and 2015, noting that the global rate of loss increased over that period. Similarly, Williams et al. (2020) [36] prepared a global analysis on an ecoregional scale, concluding that 1.9 million km^2 of land relatively free of human disturbance became highly modified between 2000 and 2013, primarily in tropical and subtropical grasslands.

Recent efforts have focused on leveraging higher-quality input data. Mu et al. (2022) [39] generated annual time-series of global Human Footprint maps, addressing the static nature of previous indices. They improved upon previous underestimations by replacing 1 km Nighttime Lights (NTL) data with 30m Landsat-extracted built environment pressure and using higher-quality land cover data. Gassert et al. (2023) [25] further pushed this paradigm by producing global 100 m resolution human pressure map time series.

This trend of increasing resolution and regional specificity was demonstrated by Zhou et al. (2025) [40], who produced annual, 100-meter resolution human footprint maps for the Qingzang Plateau. Their dataset outperformed six existing global and regional maps, highlighting significant discrepancies in spatiotemporal patterns reported by previous work. Zhou et al. attribute these divergences to proxies with errors (e.g., inaccurate land use classification, difficulty detecting weak nighttime lights), the interspersing of static data into dynamic maps, and different cumulative pressure algorithms.

Table 3.1: Comparison of Human Influence Indices: Spatial Resolution and Temporal Coverage, categorized by methodology.

Method	**Human Influence Index**	**Resolution**	**Year-Coverage**
Rule-Based	Sanderson et al. (2002) [1]	1 km	2002
	Allan et al. (2016) [2]	1 km	1993-2009
	Kennedy et al. (2019) [3]	1 km	2019
	Theobald (2020) [38]	300 m	1990-2017
	Williams et al. (2020) [36]	1 km	2000-2013
	Mu et al. (2022) [39]	1 km	2000-2018
	Sanderson et al. (2022) [35]	300 m	2000-2020
	Gassert et al. (2023) [25]	100 m	2015-2020
	Zhou et al. (2025) [40]	100 m	$1990 - 2020$
DL-Derived	Keys et al. (2021) [41]	1 km	2000-2019
	Orihuela-Pinto et al. (2025) [42]	300 m	1999-2024

Despite the foundational understanding these methods provide, rule-based mapping faces several core limitations. First, there is the challenge of the input data itself, as it is impossible to account for all possible forms of anthropogenic pressure; for example, important drivers such as deforestation, mining, and energy extraction were often overlooked or inadequately represented. These methods depend on diverse geo-indicators that are not always complete or accurate [43]. Furthermore, data completeness can vary regionally; for instance, crowdsourced data like OpenStreetMap (OSM) may be more detailed in the Global North, leading to biased assessments [44]. Second, a fundamental weakness is the reliance on subjective, heuristic-based weighting. Human pressure layers are assigned static scores based on a manual assessment of their presumed impact on ecosystems. Such rigid heuristics are problematic as they often fail to generalize, not holding true across diverse spatial contexts or different temporal scales. Finally, these issues of incomplete data and subjective weighting culminate in a significant lack of consensus and stark discrepancies between maps purporting to show the same phenomenon. While maps may show broad agreement globally, significant disagreements emerge at finer scales. A comparative analysis by Riggio et al. (2020) [45] highlights this, where one map estimates 19% of non-Antarctic land as wilderness while another estimates only $\sim 5\%$. This disparity is attributed not just to the choice of input layers, but crucially to the different

subjective methodologies used to combine them. This makes direct, fine-scale comparison between indices suboptimal, hindering their large scale application. Although the spatial resolution of these maps has progressively improved, these fundamental limitations of subjective weighting and incomplete data cannot be solved simply by increasing the pixel count, necessitating a shift toward data-driven methodologies. The progression of these data-driven methodologies, demonstrating a clear movement towards higher spatial and temporal resolution, is summarized in Table 3.1.

3.1.2 Deep Learning-Derived Mapping

Deep Learning models have demonstrated an ability to extract complex, non-linear relationships directly from satellite imagery by learning a mapping, which can potentially capture a more holistic and nuanced representation of the instance and infer this information for any time and place where satellite data is available, removing the dependency on the heuristic combination of static geo-indicators.

The complexity of human influence meant that research focused on refining traditional feature-based approaches by integrating spatial metrics and machine learning (ML) models. For instance, Kim et al. (2025) [46] investigated how human footprint intensity shapes spatial heterogeneity patterns using a case study in South Korea. They used metrics, including Local Indicators of Spatial Association and Gray-Level Co-occurrence Matrix indices, to characterize the spatial structure and texture of the Human Footprint map. Random Forest was used to model the complex, nonlinear relationships between these explicit heterogeneity metrics and ecological responses, substantially improving prediction over using the raw Human Footprint score alone.

This integration of existing human influence data within more sophisticated ML frameworks was also explored in the context of ecological modeling. For example, Cheng et al. (2025) [47] used multiple ML algorithms, including Random Forest, to predict the global distribution of the invasive weed *Conyza sumatrensis*. This study explicitly validated human activity as a primary driver of species distribution by demonstrating the critical role of integrating the Human Footprint variable alongside climatic and edaphic data. By comparing the predictive performance of ten different ML algorithms, the study further highlighted the challenge of selecting the most robust model for complex, data-rich geospatial prediction tasks. While these studies leveraged ML to improve the application and interpretation of human influence data for specific ecological insights, they still relied on the rule-based indices for the input human influence layer itself, leaving the core problem of subjective weighting and incomplete data unsolved.

A critical requirement established by these ML-driven applications is the necessity for high spatial granularity in the underlying human influence maps. The inadequacy of the 1 km standard resolution for conservation studies was confirmed by country-level research:

Presotto et al. (2025) [48] argued that for small-range species like non-human primates in Brazil, 1 km resolution cannot capture fine-scale habitat fragmentation, necessitating mapping at a 10 m resolution. Their resultant Human Impact on the Environment in Brazil (HIBR-10) dataset achieved this 10 m resolution by relying on high-granularity inputs, including Deep Learning-derived building footprints from 30 cm Maxar imagery to replace unreliable Nighttime Lights data. Crucially, the Presotto et al. study is a country-level effort that uses high-resolution source data (including satellite-derived layers like MapBiomas Sentinel-2 and Maxar imagery) which are aggregated using traditional rule-based scoring. This 10 m mapping confirms the feasibility and the critical requirement for a $10,000$ times increase in spatial resolution over current global human influence products, and demonstrated the value of leveraging DL outputs (like building footprints) as key inputs.

While these efforts demonstrate the power of learning complex relationships directly from EO data, the challenge of using DL to map more abstract, composite concepts is considerably more intricate. For such ill-posed tasks, Deep Learning offers a pathway to move beyond subjective, rule-based methodologies, provided the complex spatial structures inherent to human influence are successfully addressed. It is clear that human influence, together with machine learning, is very valuable for deriving ecological insights, and that increasing the resolution is useful for deriving more beneficial information from human influence maps. However, the manual collection and maintenance of all the geospatial layers required for the heuristic calculation of high-resolution human influence maps is not scalable or feasible. This brings in the necessity for a fundamental reformulation of the task to include Deep Learning to learn the mapping directly from satellite imagery, which would break away the dependency on needing to collect, update, check, and heuristically combine multiple pressure layers.

The intersection of human influence maps and Earth Observation, where human influence maps are used as a supervision signal to learn a direct mapping from satellites, remains significantly under-explored. The work by Keys et al. (2021) [41] established the viability of using deep learning to quantify the complex, abstract concept of human influence, marking the first study to formulate this mapping task from a deep learning perspective. Methodologically, they trained a CNN to predict the Human Footprint Index (HFI), using the existing rule-based index by Williams et al. (2020) as the training label (proxy-based supervision). The model was trained using Landsat imagery from the year 2000, specifically using three bands: the red band, and two near-infrared bands. The resulting model was then applied to 2019 Landsat data to generate an updated HFI map. Their resulting map retained the original 1 km spatial resolution of the rule-based index, thus demonstrating the capability of DL to temporally propagate the metric but without increasing its spatial granularity. To manage the global heterogeneity of the Earth's surface, a separate CNN was trained for each continent-scale region, as this technique

improved the accuracy of predictions over training a single global model. Evaluation was performed using Mean Absolute Error (MAE) and through the use of Layer-wise Relevance Propagation (LRP) to confirm that the CNN was focusing on human-made features.

Continuing this work, Orihuela-Pinto et al. (2025) [42] addressed key limitations of the original study by achieving a finer spatial resolution. They used a CNN initialized with a ResNet18 architecture and trained it to predict a more recent and higher-resolution Human Impact Index developed by Sanderson et al. (2022) [35] as the supervision signal. This more recent version, with a 300 m spatial resolution, marked an improvement over the initial 1 km product and relies on Landsat data (green, near-infrared, and shortwave infrared bands) for its predictions.

Beyond improving spatial resolution, predicting abstract and inherently ambiguous concepts like human influence or its inverse—naturalness—necessitates robust interpretability and uncertainty quantification. While Keys et al. [41] briefly utilized explainable AI techniques to verify model focus, a dedicated line of research has emerged to thoroughly decode how deep learning models perceive naturalness in Earth Observation data. Emam et al. (2024) [49] advanced this by leveraging Activation Maximization and Generative Adversarial Networks to explicitly recognize and explain the visual patterns that constitute natural areas in satellite imagery. To quantitatively evaluate the trustworthiness of these visual explanations, they subsequently introduced the Confident Naturalness Explanation (CNE) framework [50]. Recognizing that naturalness mapping is subject to inherent data ambiguities and geographic variations, their recent work has integrated uncertainty awareness directly into the interpretability pipeline. This includes the Confidence-Filtered Relevance (CFR) framework [51], which provides an uncertainty-aware mechanism for naturalness assessment, and NaT-ReX [52], which extends reliable explainability to modern transformer-based architectures. Together, these studies establish that operationalizing deep learning for abstract ecological concepts requires not only accurate spatial predictions, but also a quantified measure of model confidence and transparent reasoning.

The DL-derived human influence mapping approach, which directly maps satellite imagery to the human influence score, remains a promising methodology that warrants further attention to overcome its resolution and generalization limitations. The trend is clear: there is a persistent need for deep learning methodologies capable of producing higher spatial resolution maps using more generalizable approaches that perform robustly across diverse locations.

3.2 Challenges in Large Scale Mapping with Deep Learning

Before detailing the challenges, it is crucial to recognize that the operational creation of large scale map products using deep learning remains highly non-trivial. Notable examples of DL-powered geospatial maps include the mapping of physically grounded quantities like anthropogenic structures (e.g., buildings [53]), and the estimation of crucial ecological variables (e.g., canopy height [54, 55], above-ground biomass [56]). While these efforts validate the capacity of DL to produce large scale products, they underscore that every step, from model training to final product stitching, is fraught with unique computational and generalization risks.

While both rule-based and deep learning paradigms aim to produce comprehensive map products, a significant set of practical challenges arises when moving from local models to large scale, operational deployment. For rule-based methods, the challenge at scale is primarily one of data engineering and harmonization, as the limitations of varying data completeness and quality are magnified when creating a consistent global product from disparate sources.

For deep learning models, the challenges are different and more profound. The first challenge concerns robustness and reliability. A deep learning model's ability to fit complex relationships comes with the critical caveat that it can only be reliably applied to new data that is statistically similar to its training data. This is particularly acute for remote sensing map products, as they are the outputs of complex and often black-box machine learning algorithms whose errors can be systematic and difficult to characterize [57, 58]. The high variability and heterogeneity inherent in geospatial data—derived from diverse sensors measuring across varying physical settings, and subject to differences in spatial and temporal resolution, acquisition conditions, and atmospheric disturbances—mean that inference data often present distinct statistical distributions compared to the training data. This disparity significantly increases the chance that a model encounters an environment it has not been reliably trained to predict. When these maps are used in downstream scientific analyses, these uncharacterized errors can lead to significantly biased population-level estimators, a factor that has been cited as a barrier to their adoption by stakeholders such as the World Bank and the United Nations [59, 60]. This presents a fundamental problem for large scale spatial mapping, where the very goal is to make predictions for new geographic spaces.

The deep learning models lack the built-in capability to quantify the reliability of their predictions. This absence of reliability-awareness, which is essential for climate-sensitive and policy-relevant applications, prevents end-users from knowing how much trust to place in a specific model output. This critical need for reliability estimation has led to the development of the Area of Applicability (AoA) concept, as introduced

by Meyer et al. (2020) [61]. The AoA is defined as the domain where a model's learned relationships are valid and its estimated performance is expected to hold. Meyer et al. (2020) argue that since training data points are usually few compared to the large application area, models often make predictions for new geographic locations that differ considerably in their environmental properties from the training data, leading to highly uncertain results. To address this, they proposed a methodology that delineates the AoA by calculating a Dissimilarity Index, which is based on the weighted minimum distance in the multidimensional predictor space to the training data. By applying a threshold to the DI, the AoA is defined as the area where the prediction error is, on average, comparable to the model's cross-validation error, ensuring that predictions are restricted to spaces where the model has learned the underlying relationships [62]. A key challenge for large scale DL mapping is therefore not just producing predictions, but also understanding their reliability.

Addressing this, recent work has begun to focus on integrating multiple uncertainty signals directly into model prediction frameworks. For instance, Cohrs et al. (2025) [63] introduced a framework combining three complementary reliability signals: Out-of-Distribution (OOD) detection in the raw input space, OOD detection in the learned embedding space, and task-specific predictive uncertainty. By integrating these signals, SHRUG-FM aims to provide a comprehensive reliability assessment, which can be further linked to geospatial context to diagnose interpretable gaps in the foundation model's pretraining and guide deployment decisions in real-world environments. The development of such reliability-aware frameworks is critical for increasing the trustworthiness, robustness, and utility of large scale DL-powered geospatial maps.

Furthermore, the very task of reliably assessing the performance of large scale geospatial ML predictions presents unique difficulties [64, 65]. Standard machine learning evaluation practices (e.g., random train-test splits) are often invalidated by the inherent spatial autocorrelation present in geospatial data, which can lead to overly optimistic performance estimates. Compounding this, the frequent presence of covariate shifts between training and evaluation regions, along with issues of data availability, quality, and representivity, makes robust performance assessment challenging [62]. Rolf (2023) [21] argues that model performance extends beyond simple map accuracy metrics and emphasizes the need for evaluation frameworks tailored to the specific intended use case, considering factors like spatial generalization and the inherent structures within geospatial data.

In response, the field has begun to develop adaptation strategies to improve model transferability and operational readiness. Research has focused on areas such as geographic adaptation, where methods like label transfer can anchor a model to a new region with minimal local data for calibration [66]. Other work addresses pipeline adaptation, for instance by aligning a model to shifts in data processing, such as moving from Sentinel-2 L1C to L2A products [67]. These efforts underscore a growing consensus that

operational models must be accompanied by robust strategies for adapting to real-world scenarios.

Beyond the challenges of reliability and performance assessment, there are significant operational hurdles tied to the sheer volume of data involved. Large scale mapping requires processing massive remote sensing inputs which necessitates splitting the input scene into smaller tiles, running inference on each tile independently, and then stitching the resulting prediction outputs back together. This tiling process is computationally necessary but can introduce undesirable discontinuities or artifacts at the boundaries where tile predictions are merged [68]. The issue of tiling artifacts is a recognized drawback in global mapping efforts. For example, Neumann et al. (2025) [69] created the Natural Forests of the World map, which used a single multi-modal, multi-temporal transformer model trained on Sentinel-2 and auxiliary inputs to generate a global 10 m resolution probability map. The authors noted that subtle tiling artifacts could still appear in the final probability map near tile corners, despite employing an overlapping inference approach designed to minimize these discontinuities. Furthermore, the development of specialized tiling strategies like Flip-n-Slide, designed to maximize spatial context and eliminate data redundancy during training for semantic segmentation, underscores that existing tiling methods present an open problem for preserving contextual information [70]. This demonstrates that the operational challenge of creating seamless, large scale map products remains a critical concern for deployment.

3.3 Feature Attribution in Geospatial Deep Learning

The recent push towards Data-Centric Machine Learning emphasizes the critical need to understand and optimize the data used throughout the ML cycle [71]. Interpretable machine learning directly supports this data-centric view by providing methods to move beyond treating models as opaque black boxes and gain insight into their decision-making processes [72]. A primary task within this field is feature attribution, which aims to estimate the influence or importance of individual input features on a model's predictions.

Historically, feature importance has been assessed using three main categories: filter, wrapper, and embedded methods [73]. Filter methods act as a pre-processing step, using statistical tests to rank features independently of any model. Wrapper methods evaluate feature subsets by repeatedly training a model (e.g., Recursive Feature Elimination), while embedded methods incorporate feature selection directly into model training (e.g., L1 regularization or decision tree importance). While foundational, these traditional methods often struggle with the high dimensionality and complex, hierarchical feature representations learned by modern deep neural networks. Consequently, specific feature attribution techniques have been developed for deep learning. One major family operates in a post-hoc manner on trained models, often model-agnostically. Notable examples

include SHapley Additive exPlanations (SHAP), which uses game theory concepts to assign contributions [74], and Local Interpretable Model-agnostic Explanations (LIME), which approximates model behavior locally through perturbations [75]. Another category involves building attribution capabilities intrinsically into the model's architecture (i.e., *ad-hoc*). Attention mechanisms are a key example, learning to assign importance weights to different input parts [76, 77].

Within EO, these modern methods are increasingly applied. Post-hoc techniques like SHAP have been used to clarify decision-making in multi-label classification [78] and identify important variables in Random Forest models for tasks like biomass estimation [79]. Najjar et al. (2024) used Shapley Value Sampling combined with an iterative deletion framework (inspired by ROAR [80]) to identify necessary and sufficient feature sets, demonstrating substantial potential for data reduction in EO tasks [81]. Intrinsic methods like attention have also shown utility, improving performance and interpretability in segmentation tasks [82] by highlighting relevant spectral channels [76] or spatial regions. The value of identifying influential modalities is further underscored by recent work showing that thoughtful multi-modal fusion can significantly improve model data-efficiency and out-of-distribution generalization [83]. This push for informed feature selection is further supported by work that uses an evolutionary algorithm, the Univariate Marginal Distribution Algorithm, to select optimal spectral band compositions for deforestation detection, demonstrating that a carefully selected subset of bands can outperform the full set [84].

Occlusion sensitivity analysis represents another important family of attribution techniques. Occlusion sensitivity maps, which assess feature importance by masking parts of an input image and observing the output change, have been used in EO for tasks such as exploring wilderness characteristics [85]. In a related sense, occlusion has also been performed at the modality level (often termed "sensor dropout"), primarily as a tool to evaluate or enhance model robustness against missing sensor data during inference. This line of research investigates strategies like zero-padding or learned imputation [86].

While these diverse analytical approaches exist in Earth Observation, a significant gap persists in translating feature attribution techniques into scalable, operational strategies. Current methods primarily focus on post-hoc analysis rather than integrating attribution for proactive model control. This limits the ability to leverage interpretability for computational and data-efficiency gains, such as identifying and discarding non-influential input modalities to reduce storage and inference costs at scale. Addressing this gap has the potential to identify influential data sources and reduce reliance on shortcut or spurious correlations.

3.4 Making Deep Learning Geospatial

As established in the theoretical background (Section 2.2.2), standard deep learning methods have inherent limitations when applied directly to geospatial data. While powerful, their architectural biases are not always aligned with the properties of Earth Observation imagery. To bridge this gap, the field has explored various specialized adaptations to make deep learning models more inherently geospatial. This section reviews two categories of such methods explored in the literature: the modelling of spatial correlation and location encoding.

3.4.1 Modelling Spatial Correlation

This first approach focuses on architectural innovations that allow a model to see beyond the pixels of an immediate, isolated patch. It seeks to model the relative spatial relationships and multi-scale patterns necessary to understand an object or scene. However, this design contains a fundamental limitation for geospatial tasks: a severely limited receptive field. This limitation has driven a series of architectural innovations designed to capture multi-scale and global context. A key research goal in this domain is centered around introducing new inductive biases that force models to consider patterns spanning far beyond the immediate pixel neighbourhood.

There has been a lot of research dedicated to optimizing architectural capacity for multi-scale and global context capture. The Context-Aware Convolutional Neural Network [87] tackles the challenge of object detection in Very-High-Resolution remote sensing imagery by integrating a Context-Regions-of-Interests (Context-RoIs) mining layer to enrich feature representation. This layer automatically extracts context features by mining adaptive Context-RoIs surrounding the foreground proposals, which are then fused with the original Region-of-Interest features, leading to significantly improved detection performance. Multi-Scale Representation Concatenation [88] was introduced to transform arbitrary single-scale algorithms into multi-scale ones, balancing the trade-off between capturing fine-grained individual heterogeneity and broader contextual information. For the task of semantic segmentation, a method used a multi-kernel convolutional layer for fast aggregation of predictions at multiple scales, improving pixel-wise scene labeling accuracy [89]. Further addressing the joint complexity of scale and data source heterogeneity, the M3Fusion framework was proposed to simultaneously leverage the temporal knowledge contained in multitemporal High Spatial Resolution data and the fine spatial information available in Very High Spatial Resolution images [90]. On the foundation model front, TerraFM enriches spatial and semantic coverage by leveraging large spatial tiles and a local-global contrastive learning strategy to unify multisensor inputs [91].

While the emergence of geospatial foundation models suggests a move toward universal, pre-trained backbones, their application to high-resolution, continental-scale mapping

presents distinct operational challenges. These models typically rely on heavy vision transformer architectures with massive parameter counts, resulting in significant computational memory and latency costs during inference. For mapping tasks that require processing terabytes of imagery, this overhead is often prohibitive. Furthermore, generic foundation models are designed to maximize transferability across diverse tasks rather than optimizing for specific geospatial inductive biases [67]. Consequently, there remains a need for specialized, lightweight CNN architectures that prioritize inference efficiency and explicit geospatial modeling over the generic expressivity of heavy foundation models.

The drive to overcome limited receptive fields continues to inspire architectural exploration. For example, while vision transformers have become dominant in self-supervised learning, recent work demonstrated that Convolutional Neural Networks can achieve comparable or better performance by simply scaling up convolutional kernel sizes and making other minor adjustments, showing that increasing the effective receptive field through architectural tweaks is key to closing the performance gap [92]. The xT framework proposes a two-stage architecture for vision transformers to model extremely large images by combining local high-resolution details with global context [93]. This streaming approach uses nested tokenization and a lightweight long-sequence context encoder (like Mamba or Transformer-XL) to process massive images end-to-end without incurring quadratic memory growth. Moreover, research continues to reinforce the importance of integrating context, such as the BRIDGE-LC framework for continental land cover mapping, which showed significant performance gains by explicitly integrating both fine-grained (latitude/longitude) and coarse-grained (biogeographical region) geospatial information. This integration helps disentangle region-specific from region-agnostic features [94].

These methods show strong benchmark performance, but they are neither designed for nor tested in large scale, real-life applications that demand robustness, generalization, and computational lightness. The fundamental gap remains in developing computationally lightweight and efficient architectural solutions that can reliably transfer and operate at a continental scale. This leaves a critical operational need to rethink how these architectures are practically deployed and evaluated.

3.4.2 Location Encoding

This approach focuses on infusing models with an understanding of their absolute position on Earth. It seeks to model the unique, heterogeneous characteristics of a specific geographic coordinate, leveraging location as a powerful prior.

There has been much work on this topic, starting with early methods that adopted straightforward techniques. Early work in 2015 demonstrated an approach to incorporate location context into image classification by extracting handcrafted geographic features

from GPS coordinates and introducing them into a Convolutional Neural Network as auxiliary information [95]. For instance, studies have leveraged geo-location in the design of pre-text tasks within self-supervised learning frameworks to exploit the spatio-temporal structure of remote sensing data, thereby closing the gap between self-supervised and supervised learning [96]. A common technique to prepare these coordinates for a neural network is the cyclic encoding of longitude, which addresses the challenge that the Earth's longitude is continuous and wraps around at ± 180 degrees [54]. This is achieved by transforming the single longitude value, θ, into a two-dimensional vector $(\sin(\theta), \cos(\theta))$. The use of both sine and cosine preserves the cyclical nature of the data, ensuring that $\pm 180°$ are numerically the same point, while also providing a unique, non-singular encoding; the cosine function recovers information where the sine function is zeroed out (at $\pm 0°$ and $\pm 180°$).

Recently the literature has since moved towards specialized deep learning methods. For instance, the Geolocation-Aware Deep Coding approach sought to stabilize global model training on heterogeneous data by using location as a discrete switch to select expert regional subnetworks rather than a continuous vector [97]. Other major efforts focused on learning rich, context-aware embeddings of location via self-supervision: the CSPCSP framework used a dual-encoder with contrastive learning to bootstrap a location embedding from unlabeled images, which was then used as auxiliary information to improve geo-aware image classification [98]; SatCLIP pursued general-purpose, global location embeddings by employing Contrastive Alignment to distill objective, global ground conditions from Sentinel-2 imagery into a generalized location vector [99]; and building upon this, the RANGE approach aimed to create multi-resolution, high-information geo-embeddings by introducing a Retrieval-Augmented Neural Field to correct information loss inherent in contrastive models [100].

Another recent advancement focused on fine-grained visual classification and ecological mapping by using a metadata-aware self-supervised learning framework, BirdSAT, that unifies contrastive learning and masked image modelling, enriching the embedding space with location and time metadata from ground-level imagery and corresponding satellite data [101]. A parallel stream of research treated the Earth's surface as a continuous function: the GeoCLIP model, designed for worldwide geo-localization, used Random Fourier Features to learn high-frequency features across the globe [102]; and the GAIR framework extended this continuous modeling into a multi-modal setting by utilizing Factorized Encoders and an Implicit Neural Representation to create a continuous feature map from remote sensing data, enabling precise spatial alignment of patch and point data for fusion [103]. Finally, another recent effort focused on planet-scale image geolocalization by combining semantic geocell creation, multi-task contrastive pre-training, and a novel loss function, resulting in the PIGEOTTO model which achieved state-of-the-art performance and generalized effectively to unseen places [104].

Many studies propose location encoding methods, yet almost none are designed for or applied in large scale, real-life deployments where robustness, generalization, and lightness are essential. Specifically, complex location encoders often introduce computational overhead, presenting a barrier to large scale deployment in low-resource settings. This highlights an operational need to rethink the practical deployment and assessment of these location-aware techniques, echoing the unaddressed operational demands also observed for spatial correlation methods in the previous section.

3.5 Distribution Shift in Earth Observation

A core assumption in supervised learning is that the inference data will be independent and identically distributed (*i.i.d.*) to the training data. However, in real-world applications, this closed-world assumption is frequently violated by test-time distribution shifts that can significantly degrade model performance and harm generalization [105, 106]. Deep learning models often exhibit overconfidence in their predictions, even when confronted with data that diverges from their training distribution [107, 108, 109]. This poses a critical challenge, especially in high-stakes applications where incorrect predictions can have severe consequences. To address this, it is critical for predictive models to be able to identify when new observations fall outside the training distribution, a task known as out-of-distribution detection.

The literature contains several families of OOD detection methods, each with different underlying principles and trade-offs. One prominent category relies on scoring functions that assign a numerical score based on the model's output, such as Maximum Softmax Probability (MSP) [108] and its variants like ODIN [110] and Energy Score [111], or methods using distance metrics in feature space, such as Mahalanobis distance [112]. Another key category comprises activation-based methods that analyze and manipulate the model's internal activations to enhance separation, including techniques like ReAct [113], DICE [114], and Activation Shaping (ASH) [115]. A third line of research leverages clustering-based methods in the model's learned feature space to isolate structural patterns corresponding to ID versus OOD samples [116, 117]. While these general methods establish various approaches for OOD detection, their trade-offs—ranging from degrading primary task performance to reliance on hyperparameter tuning with hypothetical OOD samples—often limit their real-world adaptability.

In the domain of EO, distribution shifts are not an exception but the norm. They commonly occur in the form of covariate shift (changes in the input distribution with fixed labels) and semantic shift (changes in the label space or meaning of inputs). A specific and common case of covariate shift is subpopulation shift, which arises when certain seasons, regions, or acquisition conditions are underrepresented during training but appear more prominently at test time [118]. The magnitude of these shifts exists

along a spectrum: near-distribution shifts may occur within the same satellite source due to temporal, geographic, or environmental variation (e.g., seasonal changes, sensor degradation) [119]; whereas far-distribution shifts involve more substantial differences, such as changes in satellite platforms, sensor types, or entirely different imagery domains.

The critical importance of robustness in EO is highlighted by the observation that geographic distribution shift—a specific form of domain shift where performance degrades significantly in regions unseen during training—is a primary cause of performance failure [120, 121]. Despite this necessity, the application of OOD detection in EO remains relatively underexplored. Existing studies often develop specialized models, such as Dirichlet Prior Networks, framing OOD detection as an anomaly or open-set task [122, 123]. However, a significant research gap persists: many existing OOD methods, both general and EO-specific, often presume access to test-time distributions, adhere to closed-set assumptions, or degrade model performance on the primary task [124, 125, 126]. Furthermore, dedicated benchmarks reveal that standard domain generalization techniques offer little improvement over powerful contrastive pre-trained baselines when facing spatial shifts [127].

This motivates a critical research direction: moving beyond simple OOD flagging to actively modeling and anticipating the effects of geographic distribution shifts to build more reliable systems. Recent work has begun to tackle this issue not just as a post-hoc detection problem, but as a core modeling challenge. For instance, some studies demonstrate that integrating geographic coordinates directly into the model via location encoders can significantly improve robustness to spatial shifts [128], rather than treating locations as discrete, isolated domains. Complementing this, other work focuses on assessing model trustworthiness before deployment, developing methods to predict performance drops in out-of-domain regions even without access to target labels [129]. This underscores the need for post-hoc, model-agnostic OOD detection methods that can be deployed at scale without compromising primary task performance, thereby enabling the development of truly trustworthy, global-scale geospatial models.

3.6 Contribution of the Thesis

The preceding review of the state of the art has identified several key research gaps at the intersection of human influence mapping, interpretable machine learning, and geospatial deep learning. This thesis addresses these gaps through the following primary contributions:

- **A Framework for the Novel Task of Naturalness Mapping:** Building on the legacy of human influence mapping, this thesis formally establishes the task of *Naturalness Mapping*. To support this task, it introduces a rule-based *Naturalness*

Index as an annotation source and an accompanying globally-sampled, multi-modal dataset named *MapInWild-NI*.

- **A Method for Influential Modality Selection:** To address the challenge of multi-modal data fusion, this thesis contributes a feature attribution method designed to assess and identify the most influential input data modalities for the naturalness mapping task, addressing the research gap in multi-modal interpretability.
- **A Geospatially-Aware Model Architecture:** Following the selection of the most informative data source, this work develops a model architecture explicitly tailored to the challenges of satellite imagery. It advances the practice of making deep learning truly *geospatial* by integrating both spatial-contextual priors (to model spatial correlation) and location encoding (as a geolocation prior), making the final model both context- and location-aware.
- **A Deployment Pipeline for Large Scale Mapping:** To address the challenges of generalization, robustness, and distribution shift in real-world EO applications, this thesis proposes a post-hoc, model-agnostic out-of-distribution (OOD) detection method that operates without compromising the model's primary task performance. The method is designed to perform for near-distribution shifts and for scenarios where domain divergence is attributed to semantic and covariate shifts. This produces a companion *Reliability Map* for the final continental-scale *Naturalness* map product, with a specific focus on feasibility in low-resource settings.

Chapter 4

Dataset Curation for Naturalness Mapping

This chapter details the data curation process that underpins the deep learning methodology developed in this thesis. The resulting dataset, named MapInWild-NI, was first presented in [130, 131].

Training a model to perform supervised, pixel-wise naturalness mapping requires a large scale, globally representative dataset comprising both satellite imagery and corresponding annotations. The curation process involved two major stages. The first stage focuses on the development of the annotation source: it introduces the rule-based *Naturalness Index* (NI), explains the conceptual framework and data sources used, details the computation methodology, and provides a validation of the index's suitability as a proxy for naturalness. The second stage describes the construction of the final MapInWild-NI dataset: it details the sampling strategy used to curate a globally diverse dataset, the acquisition and quality assurance procedures for the multi-modal satellite data, and the final assembly and characterization of the dataset, pairing the imagery with the NI annotations. Together, these steps establish the data foundation upon which the subsequent deep learning models are built and evaluated.

4.1 From Human Influence to a Naturalness Proxy

The work in this thesis is a continuation of existing research on mapping human influence. Understanding the extent of human impact is crucial for conservation planning and environmental monitoring. Geospatial methods have long sought to quantify this impact by engineering composite geo-indices from drivers like population density and infrastructure. These indices serve as proxies for human influence, providing a structured way to measure disturbance.

This thesis updates the foundational methods of the classic human footprint with contemporary, higher resolution geospatial human influence drivers. By doing so, it creates

a more accurate and detailed proxy for anthropogenic pressure, one that is suitable to serve as an annotation source for training models. This modernized index, viewed from an inverse perspective, forms the basis for what this thesis terms naturalness mapping—the task of measuring how free a landscape is from modern, large scale human influence, as visible from space. This reframing highlights remaining undisturbed spaces, helping to identify ecologically intact regions.

4.2 Conceptual Framework and Data Sources

A rule-based naturalness index is developed to quantify the absence of human influence. The approach is an extension of the seminal Human Influence Index (HII) framework proposed by Sanderson et al. [1], which integrates four key proxies for anthropogenic pressure: population density, land transformation, accessibility, and electrical power infrastructure.

This work modernizes the HII by utilizing more recent and significantly higher-resolution global datasets. For example, where the original HII used 1 km land cover data from the 1990s, this work leverages 10 m ESA WorldCover data from 2020. Similarly, updated sources are used for population density and nighttime lights. A detailed comparison of the data sources is provided in Table 4.1. While each source provides valuable information, their individual ability to characterize naturalness is limited. Therefore, a rule-based data fusion approach is applied to integrate them into a single, comprehensive index. The workflow is shown in Figure 4.1.

Table 4.1: Comparison of geospatial data sources for HII and NI

Data Proxy	HII [1]	NI
Population Density	GPW v2 (1995, 2.5')	GPW v4.11 (2020, 30")
Land Transformation	GLULC v2 (1992-93, 1 km)	ESA WorldCover (2020, 10 m)
Accessibility	Vector Maps (1960-90s)	OpenStreetMap (2020)
Electrical Power	DMSP Stable Lights (1994-95, 30")	VIIRS Nighttime Lights (2014+)

4.3 Creating the Naturalness Index

Each geospatial proxy undergoes a standardized workflow of pre-processing, scoring, and normalization before being aggregated into the final NI.

Population Density: As a primary indicator of human impact, population density is normalized using a threshold of 1000 persons/km^2, above which the maximum influence

score is assigned. A logarithmic transformation and min-max scaling are then applied to map the values to an influence score range of [0, 10].

Land Transformation: High-resolution land cover data (ESA WorldCover V100) is used to classify land into natural and human-influenced categories. Weight factors are assigned to different land cover classes based on a rule-based scheme reflecting their degree of modification.

Accessibility: The presence of roads, railways, and navigable rivers from OpenStreetMap determines accessibility. The vector data is rasterized to a 10 m resolution and used to generate a cost surface based on travel time, accounting for terrain slope. Tobler's hiking function is used to model off-path travel costs, as illustrated by the workflow in Figure 4.2. The final accessibility score is normalized to a range of [0, 10], with a maximum walking time threshold of eight hours.

Electrical Power Infrastructure: Nighttime illumination data from VIIRS serves as a proxy for human presence and energy infrastructure. The raw radiance values are logarithmically transformed and normalized to an influence score, with an upper limit of 30 nW/cm^2/sr.

Aggregation: Once processed, the influence scores from all four components are summed to obtain a modernized version of the HII. The final NI is then produced by inverting this aggregated score and scaling it to the range [0, 100], where higher values indicate a higher degree of naturalness. Finally, large water bodies are masked using the Global Forest Change water mask [38].

4.4 Validation as an Annotation Source

To assess the validity and suitability of the NI as an annotation source for the subsequent deep learning task, a detailed qualitative and quantitative validation was performed. The following analysis compares the NI to existing human influence maps and evaluates its alignment with established patterns of land use and protection.

A visual comparison of the NI across three distinct European regions—Bavaria (Germany), Lapland (Finland), and Scotland (UK)—is presented in Figure 4.3. The differences are immediately apparent: Bavaria exhibits NI values mostly in the medium range, with higher values in mountainous regions. Lapland shows the highest overall naturalness, while Scotland is placed between the two, with large natural areas but a clearly identifiable, human-influenced Central Belt. A direct comparison with the Human Footprint (HF) and Human Modification (HM) maps for the same regions reveals a general agreement in broad-scale patterns, though the NI consistently provides more granular detail (Figure 4.4). This is further confirmed by the histograms of NI values for the study areas, shown in Figure 4.5, where the distributions align with the qualitative assessment of each region's overall naturalness.

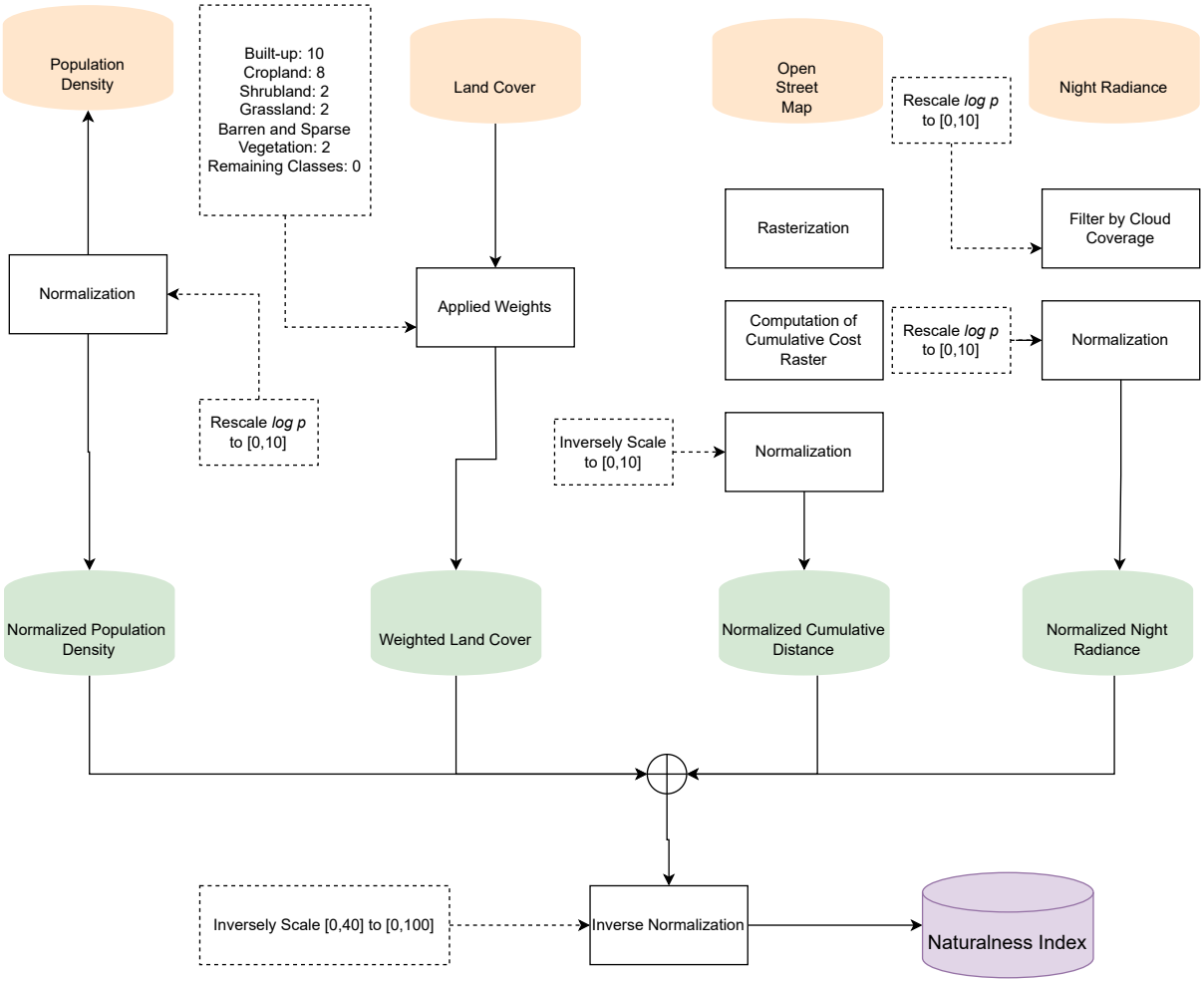


Figure 4.1: Integrated workflow for the computation of the final Naturalness Index (NI). The process involves normalizing and aggregating three main components: Population Density, Weighted Land Cover (based on 11 discrete classes), and Normalized Cumulative Distance (derived from OpenStreetMap, incorporating Tobler's hiking function for travel cost). All stages are adapted and implemented within the Google Earth Engine (GEE) platform.

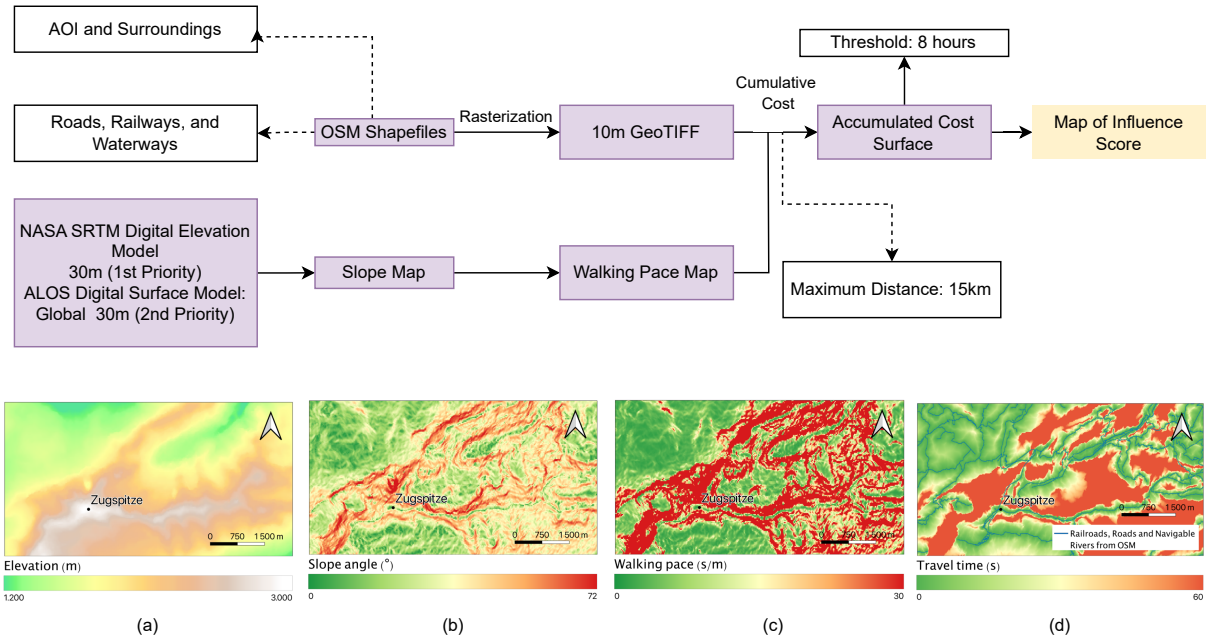


Figure 4.2: (Top) Workflow to deduce accessibility. (Bottom) An example image of (a) a digital elevation map (source: SRTM), (b) a slope map calculated from elevation, (c) a walking pace map based on terrain slope calculated using Tobler's hiking function for off-path travel, (d) and an accumulated cost surface calculated over 15 km^2 maximum distance.

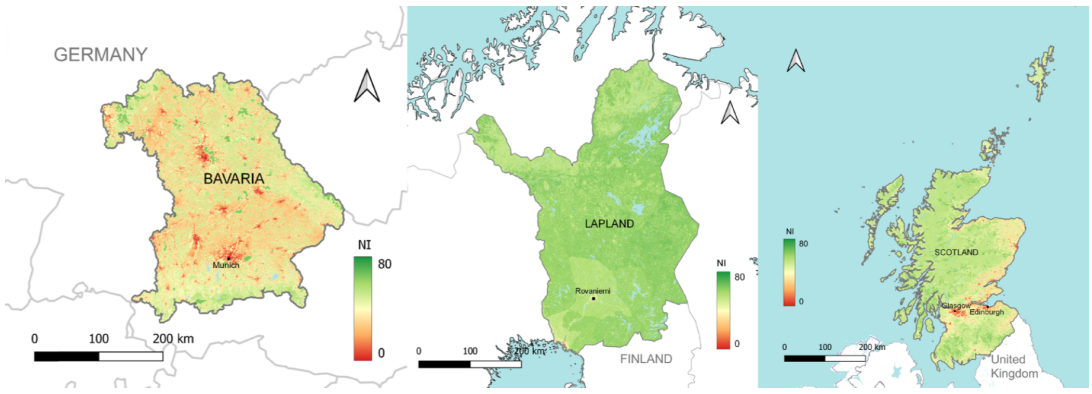


Figure 4.3: The resulting Naturalness Index (NI) maps for the three distinct case study regions: Bavaria (Germany), Lapland (Finland), and Scotland (United Kingdom). The values range from red (low naturalness) to green (high naturalness), demonstrating a clear gradient from heavily modified metropolitan areas to low-pressure landscapes.

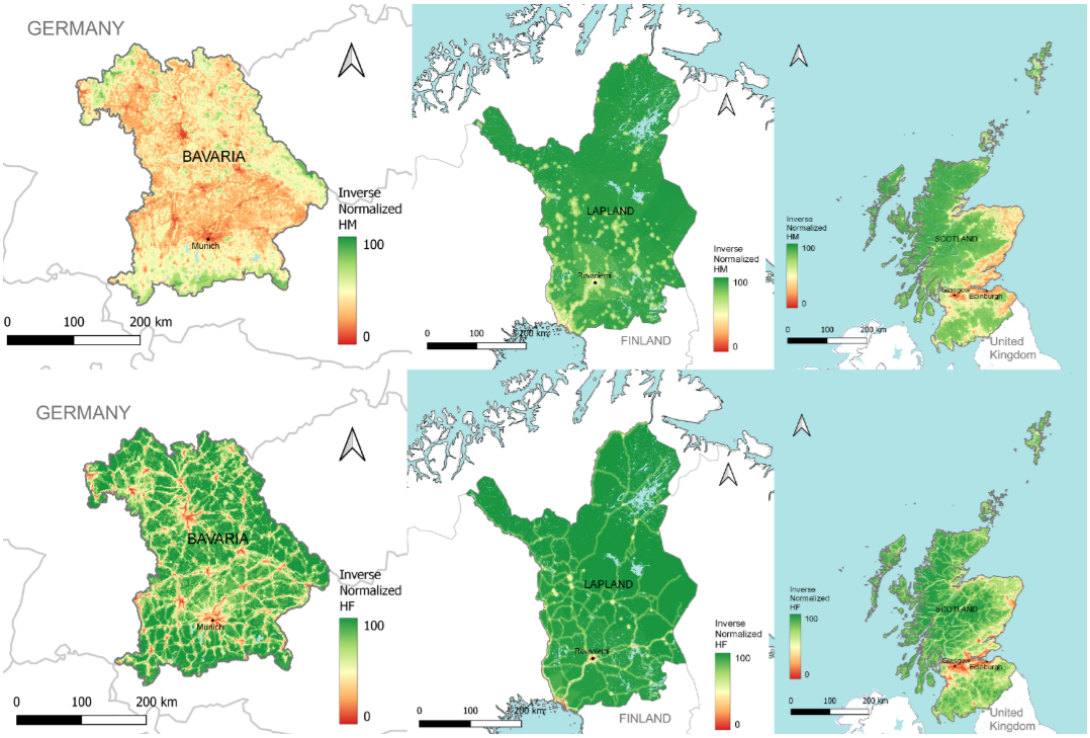


Figure 4.4: Inverted Human Modification (HM, top row) and Human Footprint (HF, bottom row) maps for Bavaria, Lapland, and Scotland. A visual comparison to the regional NI maps reveals a general agreement, although the HF seems to put a much stronger focus on transportation networks, while both HM and HF appear to exaggerate the naturalness of the land occasionally.

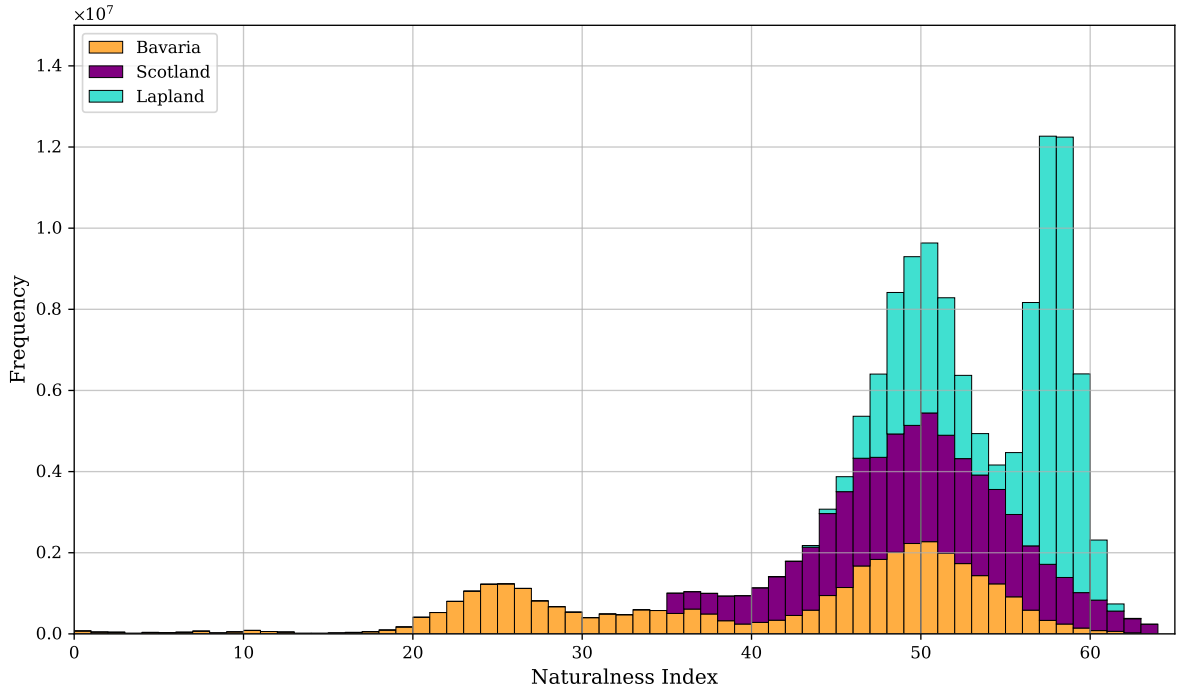


Figure 4.5: Histogram of the NI values for the study areas. While the NI values for Bavaria are relatively spread out, the NI values for both Scotland and Lapland peak at higher levels.

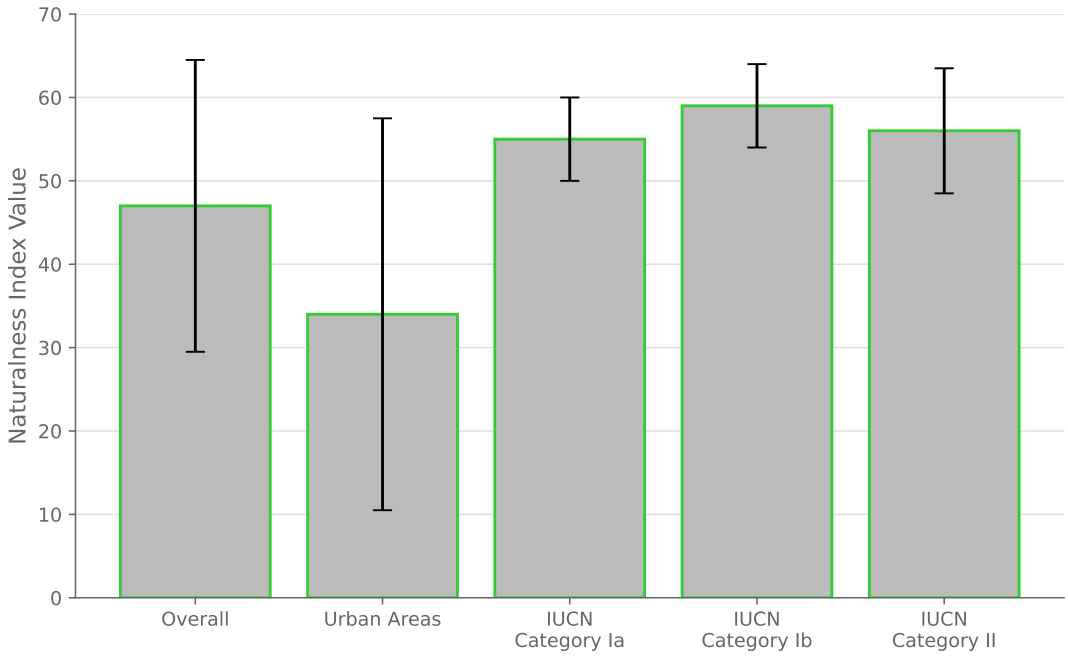


Figure 4.6: Comparison of the average NI values for different area categories. The observation that protected conservation areas exhibit higher NI values than, e.g., urban areas, is clearly visible.

To further validate the NI quantitatively, Figure 4.6 compares average NI values across wide-ranging categories and substantially different anthropogenic pressure levels. The area categories labeled Ia, Ib, and II represent the most stringent protected regions sourced from the World Database of Protected Areas (WDPA) [132], as detailed in Section 4.5. The results confirm that protected areas exhibit significantly higher NI values than urban and other human-influenced areas. This supports the reliability of the NI as a measure for distinguishing more natural regions from less natural ones.

A close-up examination of individual protected areas, shown in Figure 4.7, highlights the advantage of the NI's higher resolution. Focusing on the Clyde Valley Woodlands National Nature Reserve and Cairngorms National Park, the NI successfully identifies and delineates the fine-scale features of the Clyde Valley reserve, whereas the coarser HF and HM maps fail to do so. In the case of Cairngorms National Park, the NI map reveals clear variations in naturalness across the region, aligning with the patterns shown in existing HF and HM classifications. This detailed analysis supports the robustness of the NI as a refined measure of human impact and confirms its suitability as a high-resolution annotation source. In the subsequent chapters of this thesis, this index serves as the annotation source for the primary contribution: a framework designed to learn and generalize these patterns directly from satellite imagery.

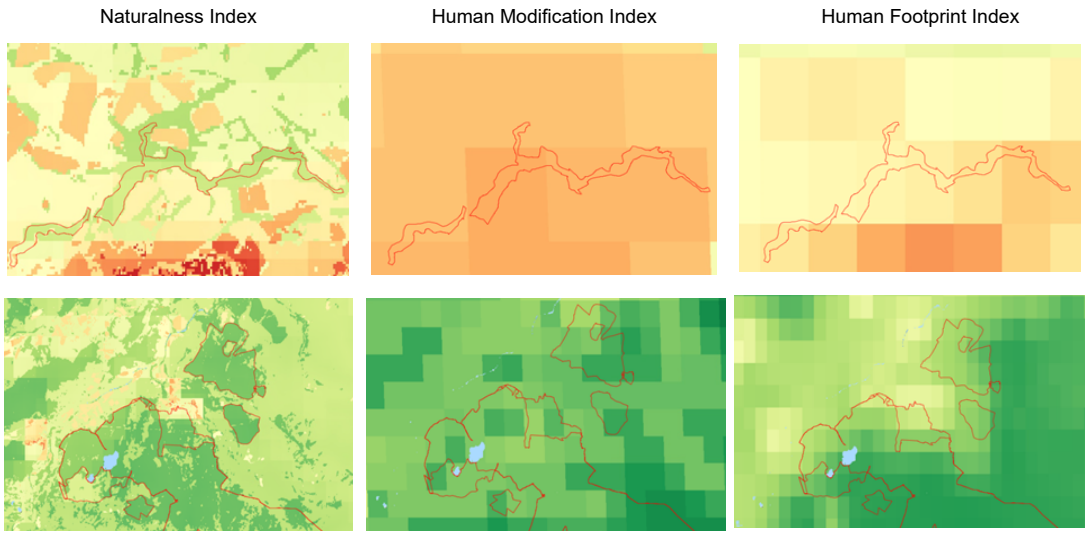


Figure 4.7: Close-look, from left to right: NI, HM, and HF values. Red polygons indicate IUCN designated areas. First row: Clyde Valley Woodlands National Nature Reserve, Scotland. Second row: Cairngorms National Park, Scotland.

4.5 Construction of the *MapInWild-NI* Dataset

A representative dataset is crucial for training a model to perform supervised, pixel-wise naturalness mapping. This section introduces the *MapInWild-NI* dataset, which was tailored for the task at hand. The core design principle was to create a dataset that represents the full continuum of human disturbance, from minimally impacted (natural) to highly modified (anthropogenic) landscapes. The following subsections detail the three main stages of the curation process: the sampling strategy for selecting Areas of Interest (AOIs), the acquisition and quality assurance of the raw geospatial data, and the final assembly and characterization of the dataset.

4.5.1 Design Principles and Sampling Strategy

The sampling strategy was designed to capture a globally representative set of landscapes by anchoring the two ends of the human disturbance spectrum. This was achieved by sourcing highly natural examples from the WDPA and manually augmenting the dataset with highly modified examples from regions with high human pressure.

A core methodological choice in this design is the use of the WDPA as the primary source for identifying highly natural landscapes. This is a deliberate, and indeed opinionated, decision that carries a key assumption: that areas legally designated for strict conservation (specifically IUCN Categories Ia, Ib, and II) serve as a reliable proxy for the natural end of the human influence spectrum. This approach is not without limitations; it does not imply that all protected areas are pristine, nor that all unprotected areas lack naturalness. However, it provides a globally consistent and systematically defined

Table 4.2: IUCN protected area categories

WDPA Class	Description
Ia: Strict nature reserve	Strictly protected areas where human presence is strictly limited and controlled.
Ib: Wilderness area	Slightly modified areas with little human presence in the form of indigenous and local communities.
II: National park	Functioning ecosystems that are subject to tourism through zoning.
III: Natural Monument or Feature	Areas where preserving a particular feature hosting a cultural value is the dominant goal.
IV: Habitat/species management area	Areas where flora species, fauna species, or habitats are preserved and/or restored through informed interventions.
V: Protected landscape/seascape	Areas where a distinct value is created by human presence over time.
VI: Protected area with sustainable use of natural resources	Areas where the conservation of natural ecosystems and ecological processes takes place.

starting point for creating a supervisory signal. The IUCN protected area categories, along with their official descriptions, are detailed in Table 4.2.

To implement this strategy, a multi-step process was designed. The highly natural samples were sourced from the specified IUCN categories within the WDPA. Given that the raw WDPA polygons are unevenly distributed and vary in size, a guided, semi-automated stratified sampling approach was developed to ensure a representative selection (Figure 4.8). This process involved applying class-specific minimum area thresholds ($5\,\mathrm{km}^2$ for categories Ia/Ib, and $100\,\mathrm{km}^2$ for category II), followed by a stratified sampling operation using the Köppen-Geiger climate and ESA WorldCover maps as auxiliary guides. By inversely weighting the sampling based on the frequency of land cover types within each climate zone, underrepresented polygon types were oversampled. To reduce spatial clustering, polygons within defined proximity limits were also excluded.

Following this automated sampling of highly natural examples, 108 highly modified AOIs were manually sampled from densely populated regions worldwide to ensure the model is exposed to various forms of human disturbance. Given that the dataset comprises 1,017 total scenes, these 108 samples constitute over 10% of the data.

4.5.2 Data Acquisition and Quality Assurance

For each sampled Area of Interest (AOI), a suite of geospatial data was acquired using Google Earth Engine (GEE) [133]. This strategy was driven by the foundational assumption that accurately mapping naturalness requires leveraging the complementary information offered by diverse, globally available sensor types. The curated modalities include Sentinel-1, four-season Sentinel-2, ESA WorldCover, and Visible Infrared Imaging Radiometer Suite (VIIRS) data. These sources provide measurable proxies for human influence: Sentinel-2 data provides valuable spectral information on what is on the ground;

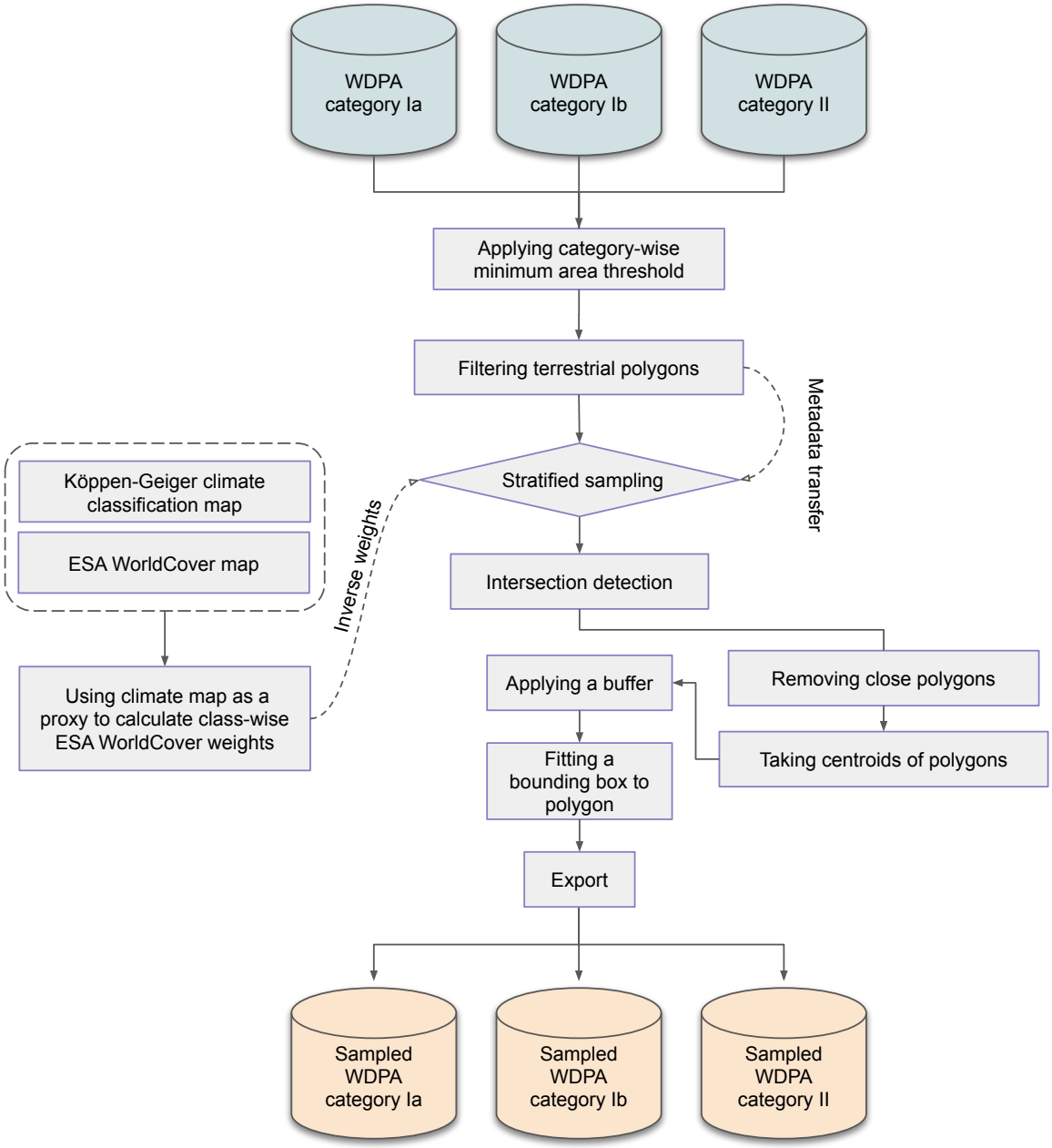


Figure 4.8: Sampling workflow: The guided stratified sampling of the WDPA polygons. The WDPA polygons that are varying in size and shape are filtered through a semi-automated sampled process that is guided by the climate zones and the land cover types.

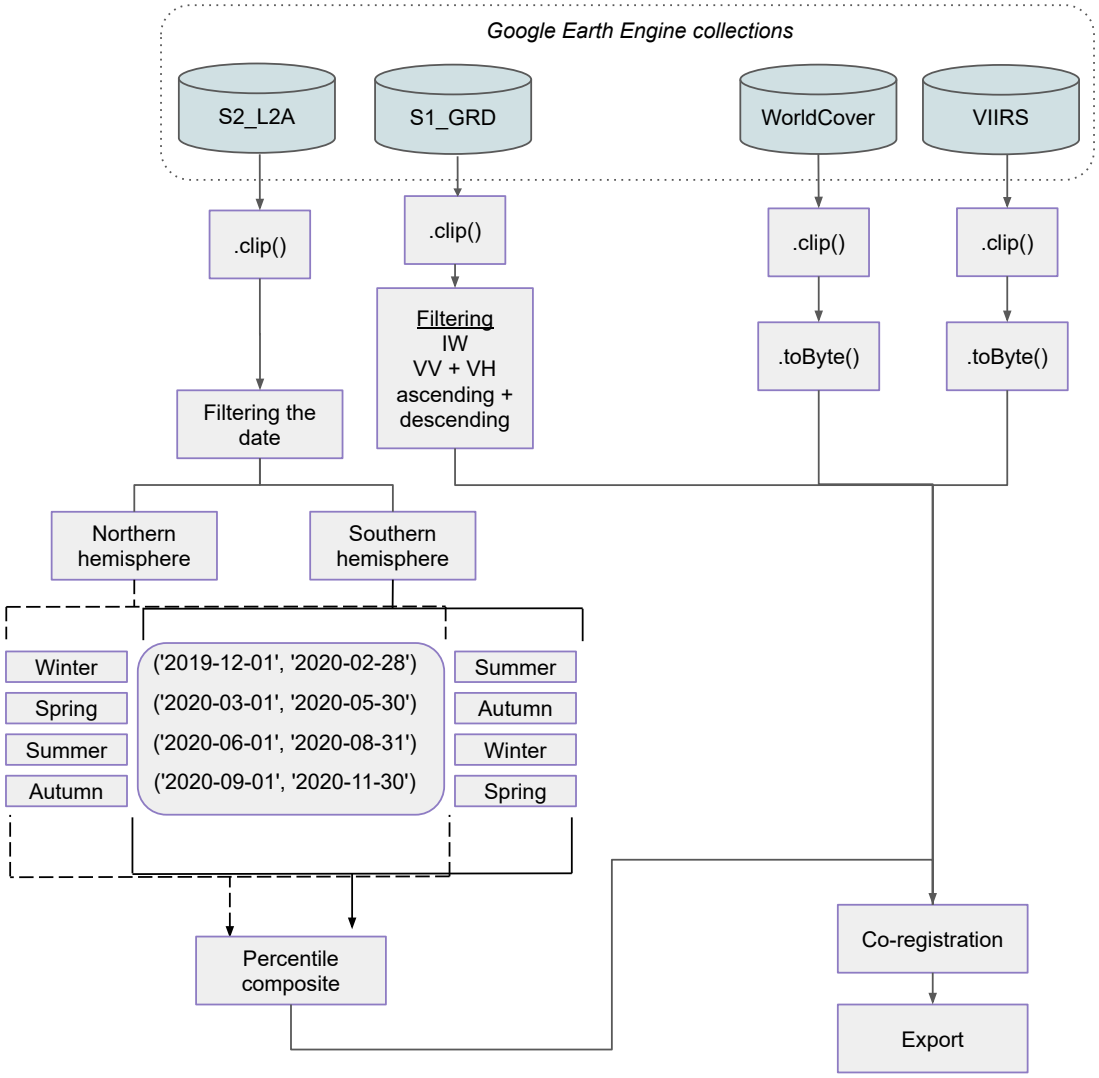


Figure 4.9: Curation workflow: Querying and filtering the geodata sources in MapInWild-NI dataset. L2A: Level-2A, GRD: Ground Range Distance, VIIRS: Visible Infrared Imaging Radiometer Suite.

Table 4.3: Seasonal distribution of the dataset across train, validation, and test sets.

Single Temporal Subset	Train	Validation	Test
Autumn	46	2	5
Spring	141	15	10
Summer	508	72	70
Winter	117	10	21

Sentinel-1 backscatter reveals physical structure; ESA WorldCover delineates how land is used; and VIIRS data provides a surrogate for human activity and infrastructure density by quantifying nighttime radiance.

Sentinel-1 imagery was acquired as a ground-range-detected (GRD) product with dual (VV and VH) polarization. For Sentinel-2, Level-2A imagery containing 10 spectral bands was acquired for all four seasons of 2020 for each AOI. A 25th percentile compositing method was used to create cloud-reduced mosaics for each season, a method which empirical analysis showed produced more informative images than median compositing. The curation workflow is illustrated in Figure 4.9.

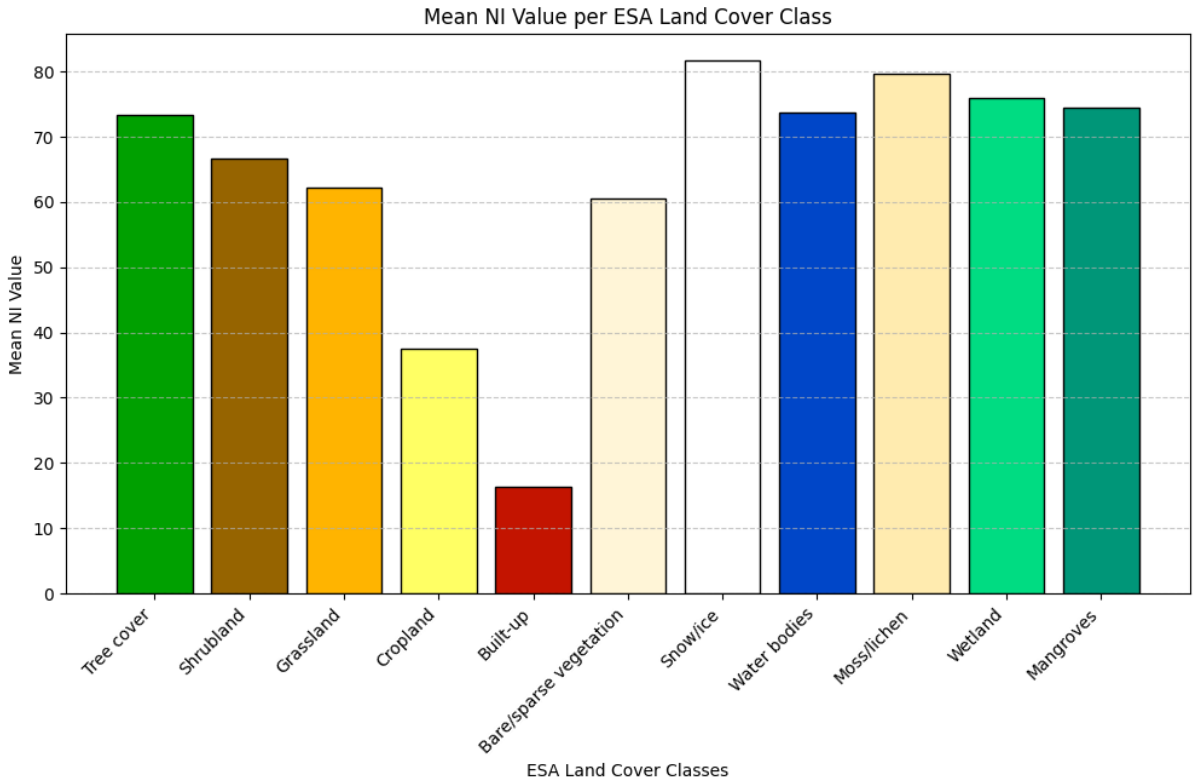


Figure 4.10: Mean NI values for each ESA WorldCover land cover class. This figure highlights the average NI for different land cover types, indicating how human influence varies across landscapes. Higher NI values correspond to more natural or less disturbed areas, whereas lower values indicate stronger human impact.

An important step in the curation process was a manual quality assurance check performed by remote sensing experts via a dedicated evaluation platform. Each seasonal

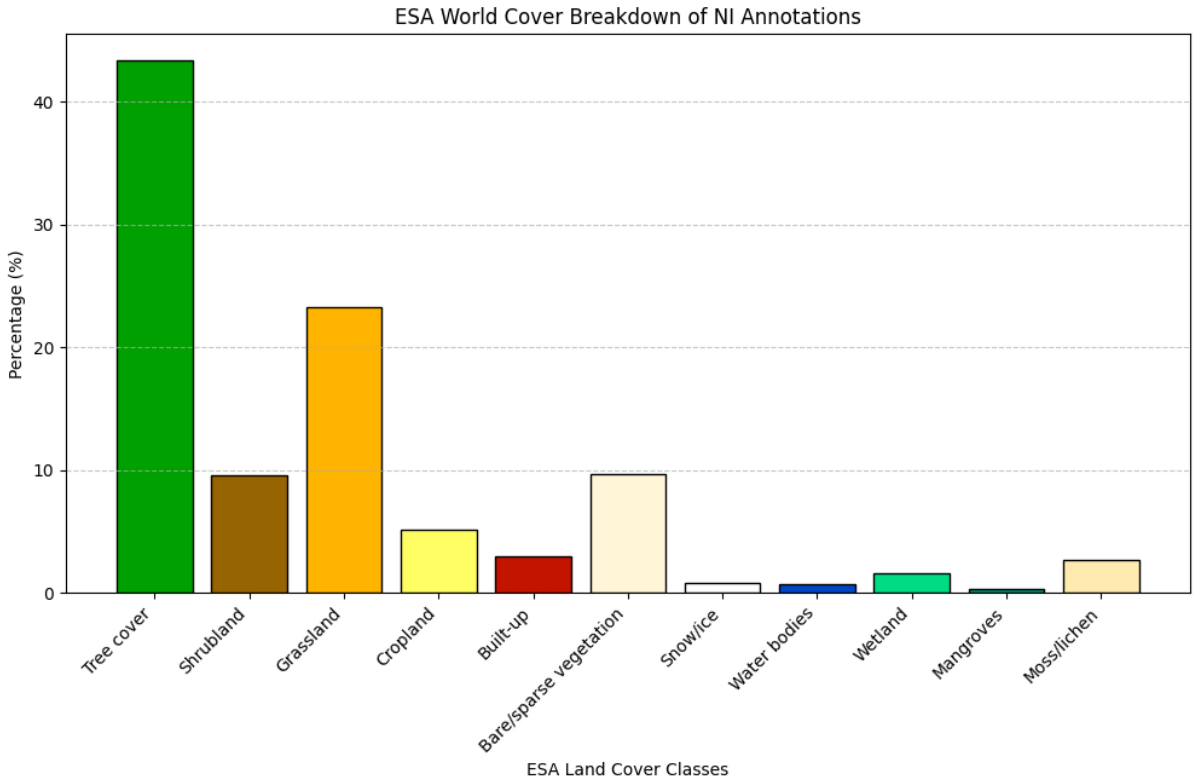


Figure 4.11: Percentage breakdown of ESA WorldCover land cover classes in the annotation. This figure shows the relative proportion of each land cover type present within the analyzed areas, helping to contextualize how land cover is represented in the NI dataset.

Sentinel-2 composite for every AOI was evaluated for informativeness, considering factors like cloud cover and processing artifacts. Based on these quality scores, a single, high-quality temporal subset (prioritizing summer in cases of identical scores) was selected for each AOI. The final seasonal distribution of the selected images is shown in Table 4.3.

4.5.3 Final Dataset

The final step was to assemble the MapInWild-NI dataset by pairing the curated geospatial data with the corresponding NI annotations, which were generated following the methodology described in Section 4.3. The resulting dataset, named MapInWild-NI [134], contains multi-modal imagery and pixel-wise NI annotations.

MapInWild-NI consists of data from diverse modalities, including Sentinel-2, Sentinel-1, VIIRS nighttime lights, and ESA WorldCover. The dataset comprises 1,017 images, each with a size of 1920 × 1920 pixels, amounting to approximately 350 GB in total.

To ensure that the model is trained on a representative sample of the full naturalness spectrum, a stratified 80-10-10 split was applied based on the distribution of NI values. The dataset was first binned into ten intervals; if a bin contained fewer than three samples, they were all assigned to the training set, otherwise, the 80-10-10 split was applied. This strategy mitigates class imbalance by ensuring that rare, extreme NI values are adequately

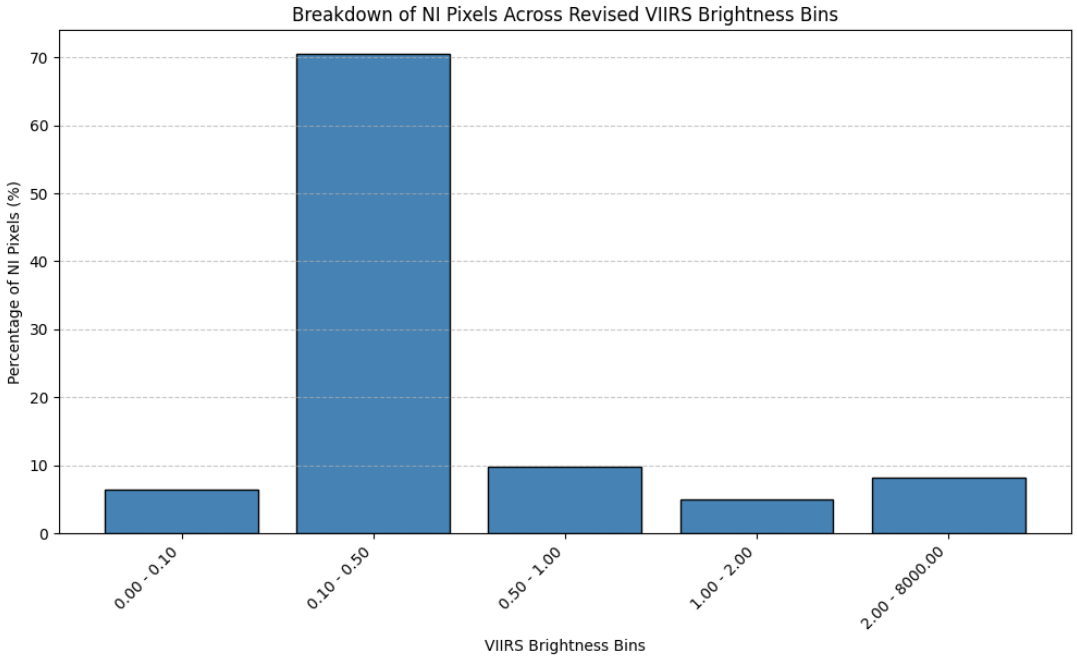


Figure 4.12: Mean NI values across VIIRS brightness bins. The results indicate a clear negative relationship, where higher nighttime light intensity, representing increased human activity and urbanization, corresponds to lower NI values, suggesting greater human impact on the landscape.

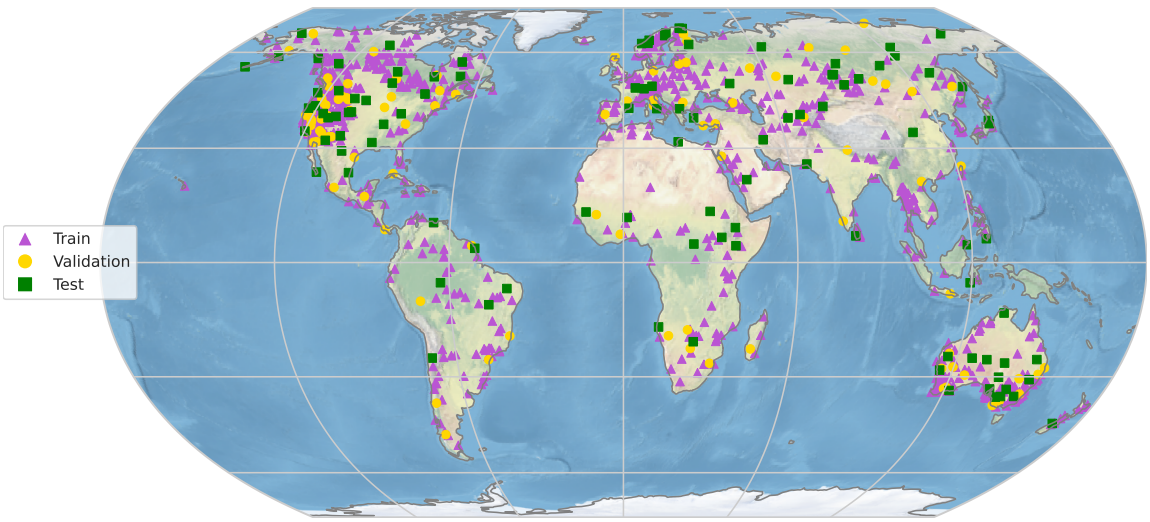


Figure 4.13: Distribution of the AOIs and training, validation, and test sets across continents. The points represent the center of each study area and are enlarged for better visualization.

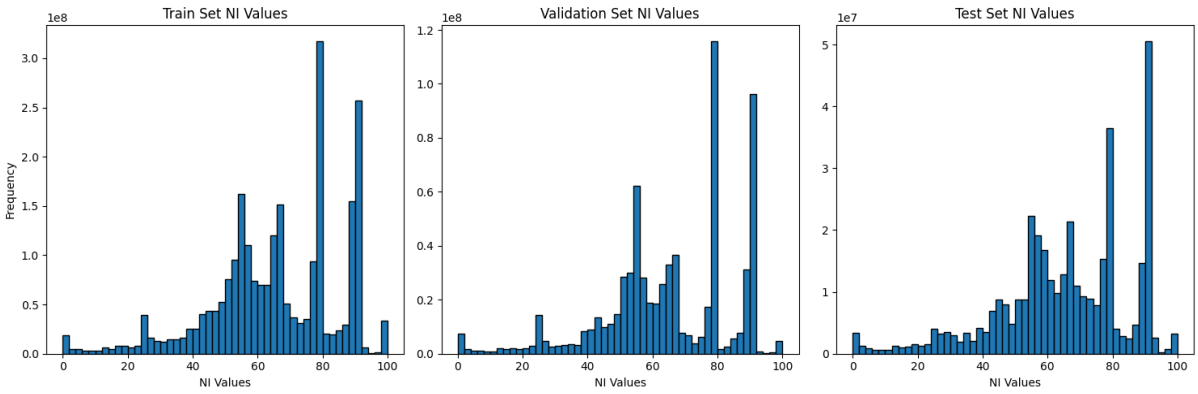


Figure 4.14: Distribution of annotated Naturalness Index (NI) values in the training, validation, and test sets. Each histogram shows the frequency of NI values, indicating the data balance and presence of dominant NI ranges across the train, validation, test splits.

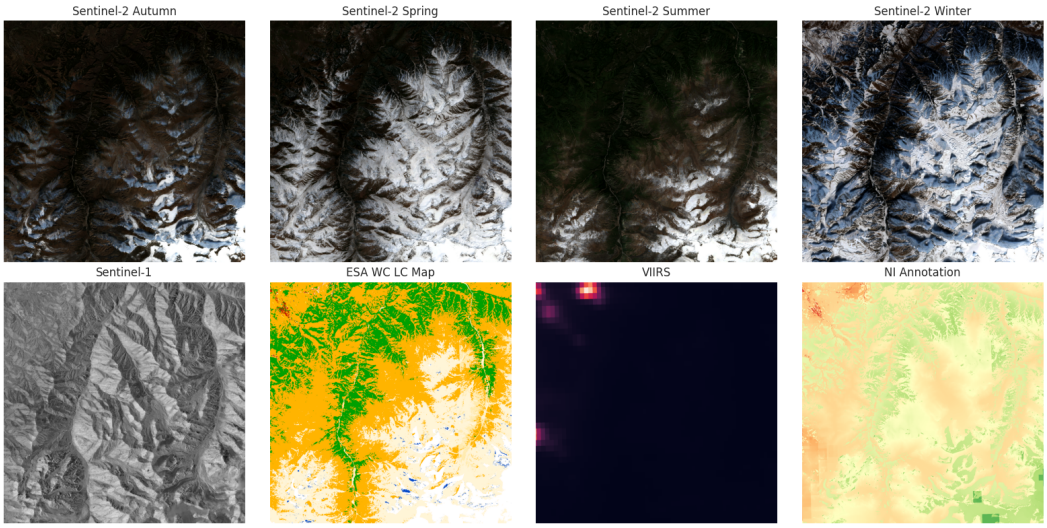


Figure 4.15: A wilderness sample from MapInWild-NI dataset (Ile-Alatau State National Nature Park, Kazakhstan. Dataset ID: 100018). The first row shows the Sentinel-2 images of four seasons, from left to right: spring, summer, autumn, and winter. Second row, same order: Sentinel-1 image, ESA WorldCover map, VIIRS Nighttime Day/Night band, and WDPA annotation. For visualization purposes, the VV and VH bands of the Sentinel-1 image are treated as Red and Green bands and only visible bands (Red: B4, Green: B3, Blue: B2) in the Sentinel-2 image are used.

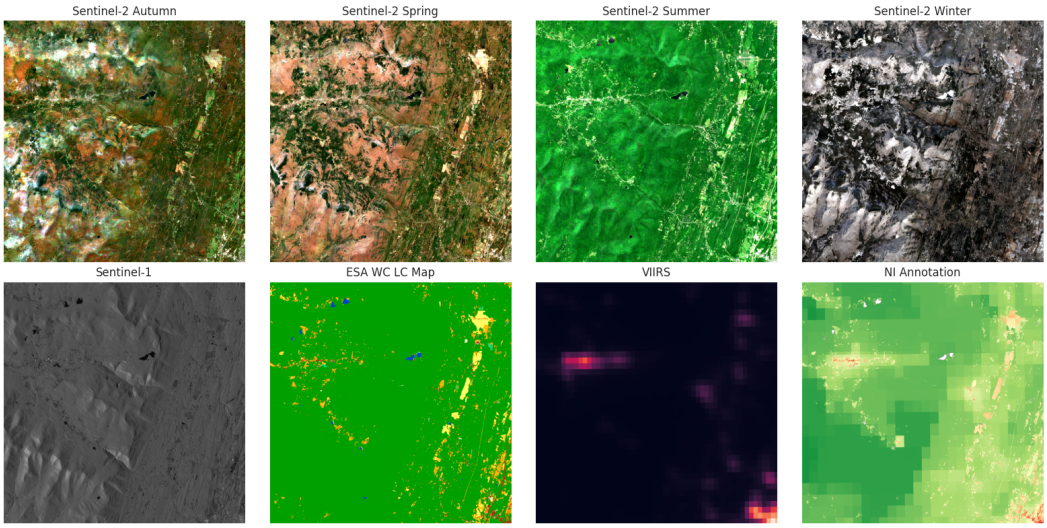


Figure 4.16: A wilderness sample from MapInWild-NI dataset (Kaaterskill, USA. Dataset ID: 555582). The first row shows the Sentinel-2 images of four seasons, from left to right: spring, summer, autumn, and winter. Second row, same order: Sentinel-1 image, ESA WorldCover map, VIIRS Nighttime Day/Night band, and WDPA annotation. For visualization purposes, the VV and VH bands of the Sentinel-1 image are treated as Red and Green bands and only visible bands (Red: B4, Green: B3, Blue: B2) in the Sentinel-2 image are used.

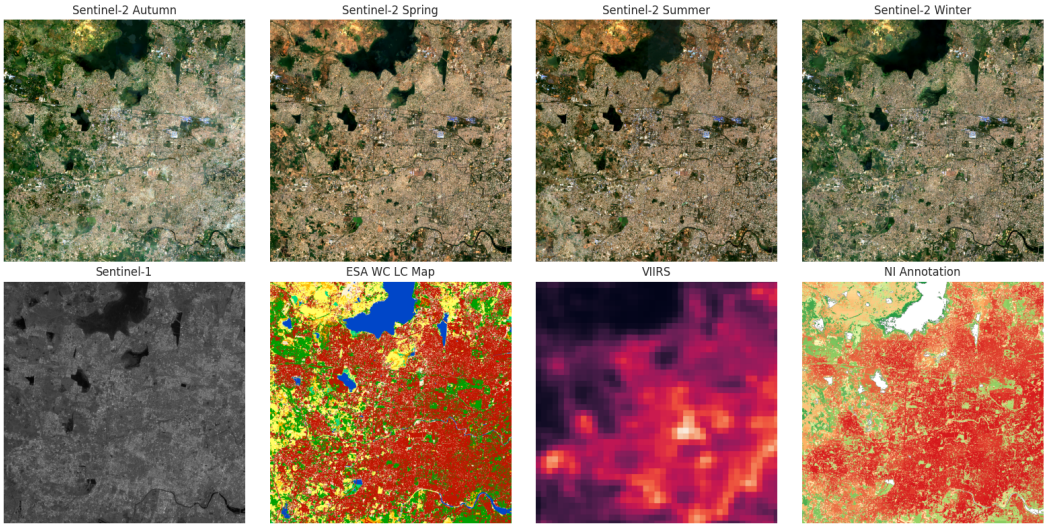


Figure 4.17: A sample representing an area of intense anthropogenic pressure from MapInWild-NI dataset (Chennai, India. Dataset ID: 900000010). The first row shows the Sentinel-2 images of four seasons, from left to right: spring, summer, autumn, and winter. Second row, same order: Sentinel-1 image, ESA WorldCover map, VIIRS Nighttime Day/Night band, and WDPA annotation. For visualization purposes, the VV and VH bands of the Sentinel-1 image are treated as Red and Green bands and only visible bands (Red: B4, Green: B3, Blue: B2) in the Sentinel-2 image are used.

represented in the training set. This resulted in a final split of 812, 99, and 106 samples for the training, validation, and test sets, respectively. Each scene measures 1920×1920 pixels. During training, these scenes are cropped on-the-fly into smaller patches, which is the input size the model operates on. This strategy significantly increases the effective number of samples in the dataset. Given the 10 m spatial resolution, each 1920×1920 scene represents an area of 368.64 km^2. Consequently, the 1017 scenes in the dataset collectively cover a vast total analyzed area of approximately 375 000 km^2.

The resulting distribution of NI annotations across these splits is illustrated in Figure 4.14. The figure clearly demonstrates that the NI annotation distribution is multi-modal, exhibiting significant peaks in the mid-to-high range (approximately 50 to 80) and a strong concentration of the highest naturalness values (near 90 to 100). Conversely, the frequency of samples is very low at the extremes, particularly for the lowest NI values (near 0 to 10). Crucially, the shape of the histogram is highly consistent across the Training, Validation, and Test sets. This uniformity confirms that the stratified splitting process was successful, ensuring that the test set is a representative sample of the overall NI data distribution, which is essential for a robust and generalizable evaluation.

An analysis of the final dataset confirms its quality, diversity, and geospatial coverage. The spatial distribution of the AOIs is global, providing broad geographic representation across the train, validation, and test sets (Figure 4.13). A characterization of the dataset's land cover composition reveals a wide range of environments, with a logical correlation between land cover type and mean NI value (Figure 4.10 and Figure 4.11). For example, tree cover and wetland classes have high average NI scores, while built-up areas have very low scores. Similarly, a clear negative correlation is observed between NI values and VIIRS nighttime light intensity, further validating the dataset's internal consistency (Figure 4.12). This comprehensive curation and characterization process yields a robust dataset that serves as a test-bed for experiments detailed in the remainder of this thesis. Examples of the final curated data, showcasing both highly natural and highly modified samples, are provided in Figure 4.15, Figure 4.16, and Figure 4.17.

Chapter 5

Deep Learning Methodology for Naturalness Mapping

Having established the foundational *MapInWild-NI* dataset in the previous chapter, this chapter details the core deep learning methodology developed in this thesis. The objective is to leverage this dataset to train a model capable of inferring naturalness directly from satellite imagery, thereby overcoming the limitations of static, rule-based approaches. This chapter presents the complete analytical framework, focusing on three key methodological innovations designed to create an efficient, geospatially-aware, and robust model suitable for large scale deployment.

The chapter begins by introducing the interpretable machine learning method used for influential input modality selection, which allows for the identification of the most salient data sources. It subsequently details the specific architectural adaptations—incorporating spatial correlation modeling and location encoding—that make the model inherently *geospatial*. Finally, the chapter addresses the critical challenge of model robustness and generalizability in real-world applications by presenting the methodology developed for out-of-distribution detection.

5.1 Influential Input Modality Selection

The underlying approach of this section is first presented in [135].

The abundance of multi-modal EO data presents a challenge in determining the true influence of each individual data stream (or modality) on a deep learning model's final prediction. Feature attribution methods are essential for resolving this opacity, as they quantify the contribution of each input variable to the model's output. In the context of multi-modal EO, identifying the most influential modalities is critical for simplifying model complexity, optimizing computational resources, and enhancing model trustworthiness.

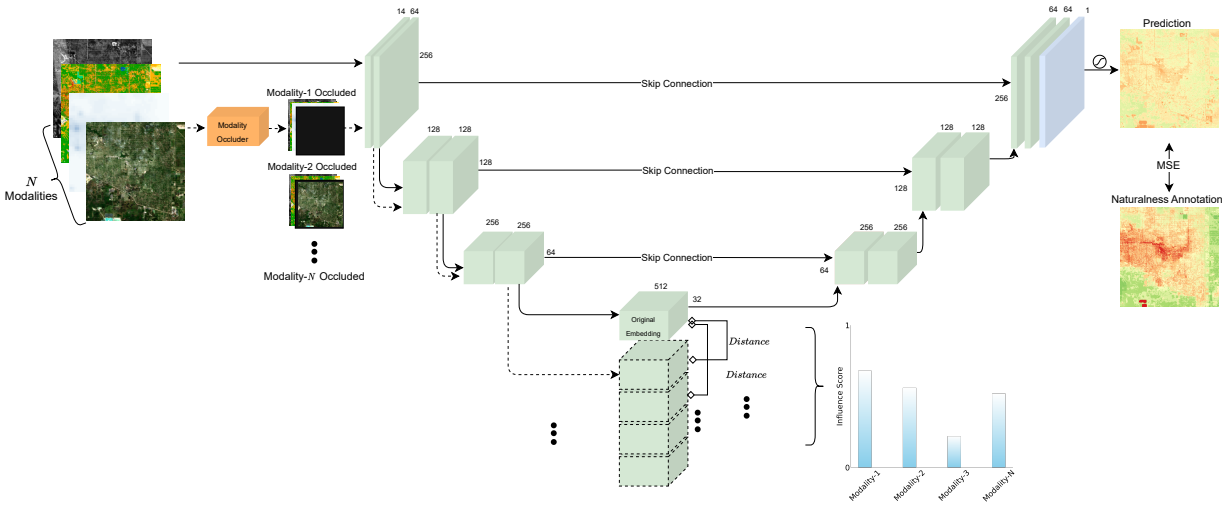


Figure 5.1: The framework for assessing land naturalness using multi-modal Earth Observation data. The iterative occlusion process, shown by dashed lines, is repeated N times for the N input modalities. In each iteration, one modality is zeroed out to create an occluded embedding. Comparing these occluded embeddings with the baseline from the non-occluded input reveals the impact of each modality on the model's output.

To achieve this, a Deep Occlusion Framework is proposed to enhance interpretability through feature attribution on the embedding level. This approach systematically analyzes the impact of different modalities by leveraging an occlusion strategy within the model's latent representation space, rather than solely at the output layer.

As illustrated in Figure 5.1, the method involves generating two sets of embeddings from the model's encoder: one from the original, non-occluded multi-modal input (e_{original}) and another from the input where a specific modality is set to a zero-state (i.e., channel values set to zero) (e_{occluded}). The purpose is to discern how the absence of an input modality affects the model's embedding space and, by extension, its prediction capability.

To quantitatively assess the influence (or attribution score, I_i) of each occluded modality i, the Euclidean distance (L_2 norm) between the two embeddings is computed:

$$I_i = D_i = ||e_{\text{original}} - e_{\text{occluded}}||_2$$

A larger distance D_i indicates a significant alteration in the embedding due to the occlusion, suggesting that the occluded modality plays a crucial role in the model's decision-making process. Conversely, smaller distances imply minimal change, indicating the lesser importance of that modality.

It is important to note that the distances provide relative rather than absolute information about modality importance. The comparison of distance metrics across different modalities offers insights into their relative contributions to the model's predictions. This approach not only aids in understanding the model's sensitivity to specific input features but also assists in feature selection and enhancements to model interpretability.

Beyond its utility as a post-hoc analysis tool, the framework incorporates an intrinsic regularization mechanism during training. The modality influence scores are utilized to formulate an embedding regularization loss (L_{reg}), which is the average of the distance scores across all N occluded modalities in a batch:

$$L_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} D_i$$

This regularization loss is then combined with the primary task loss (L_{task}) to create the total loss function:

$$L_{\text{total}} = L_{\text{task}} + \lambda L_{\text{reg}}$$

where λ is a hyperparameter balancing the influence of the regularization term. By minimizing L_{total}, the model is encouraged to achieve its main objective while simultaneously minimizing the differences in internal representations caused by occluding input features. This methodology ensures that the model's performance is not overly dependent on any single input modality, which enhances both its generalizability and interpretability. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Feature Attribution with Re-purposed Occlusion Sensitivity Maps at Embedding Level: The framework utilizes the embedding distance (L_2 norm) as a regularization loss during training, calculated through sequential channel occlusion.

Input: Image batch of shape $B \times C \times H \times W$, Annotation of shape $B \times 1 \times H \times W$
Output: Regression output, Combined Loss (Regression Loss and Embedding Distance Loss)
▷ Training Scheme
Initialize I_{distance} to zero
for each channel c in C **do**
▷ Step 1: Occlude channel
    $x_{\text{occluded}} \leftarrow \text{ChannelOccluder}(x, c)$
▷ Step 2: Obtain embeddings
    $x_{\text{encoded}} \leftarrow \text{Encoder}(x)$
    $x_{\text{encoded\_occluded}} \leftarrow \text{Encoder}(x_{\text{occluded}})$
▷ Step 3: Calculate Embedding Distance
    $I_c \leftarrow \text{Distance}(x_{\text{encoded}}, x_{\text{encoded\_occluded}})$
    $I_{\text{distance}} \leftarrow I_{\text{distance}} + I_c$
end for
▷ Compute Combined Loss
$L_{\text{regression}} \leftarrow \text{Regression}(x_{\text{encoded}}, \text{Annotation})$
$L_{\text{combined}} \leftarrow L_{\text{regression}} + \lambda \times \frac{I_{\text{distance}}}{N}$
return L_{combined}
▷ Testing Scheme
▷ Repeat steps 1-3 for post-hoc explanation without loss calculation

5.2 Incorporating Geospatial Priors in Deep Learning

The underlying approach of this section is first presented in [136].

As established in Chapter 2.2.2, standard CNNs have several inherent limitations for geospatial analysis. This section details the proposed model architecture, which is explicitly designed to address these shortcomings by including geospatial priors. The methodology aims at patching these classic intrinsic properties of CNNs:

- A small effective receptive field (ERF): Standard convolutional layers have a limited view of the input. To counteract this, the proposed method injects a tile embedding from a wider context tile. This provides a global-contextual signal by fusing the latent space L_T from a much wider input field, granting the final prediction a larger virtual receptive field to compensate for the local constraints of the UNet_{enc}'s native processing scope.
- A weak positional signal: Translation-invariant filters ignore absolute location and large scale layout. The proposed architecture addresses this by encoding the absolute geographic position (L_C) via cyclic coordinates to mitigate Spatial Non-Stationarity, and by fusing the relative macro-geometry extracted by L_T, providing a robust, contextual position in the global domain.

The framework (Figure 5.2), utilizing a multi-encoder architecture, integrates these priors. For a given small patch P, the objective is to predict a naturalness map M. This is achieved by combining three distinct latent representations: one from the patch itself (L_P), one from its broader context tile T (L_T), and one from its geographic coordinates C (L_C). The following subsections detail each component of this geospatially-aware model.

For the task of pixel-wise regression, the vanilla UNet [137] architecture is utilized as the backbone. This architecture consists of an encoder, UNet_{enc}, and a decoder, UNet_{dec}. The encoder processes the patch P (of shape $256 \times 256 \times 3$) through a series of convolutional operations, batch normalization, ReLU activation, and max pooling to create an embedding $L_P = \text{UNet}_{\text{enc}}(P)$. The decoder, UNet_{dec}, mirrors the encoder's structure and uses transposed convolutional layers. Subsequently, the decoder generates the predicted naturalness map M through an upsampling process, where $M = \text{UNet}_{\text{dec}}(L_P)$. The UNet architecture features consecutive blocks with channel sizes of $64, 128, 256$, and 512 across its depth. The final layer outputs raw logits as the prediction output.

Several studies have previously used encoding techniques for geographic coordinates [96, 99, 54]. The rationale for the cyclic encoding of longitude coordinates is derived from the continuous nature of the Earth's longitude, which wraps around at $+180°$ and $-180°$.

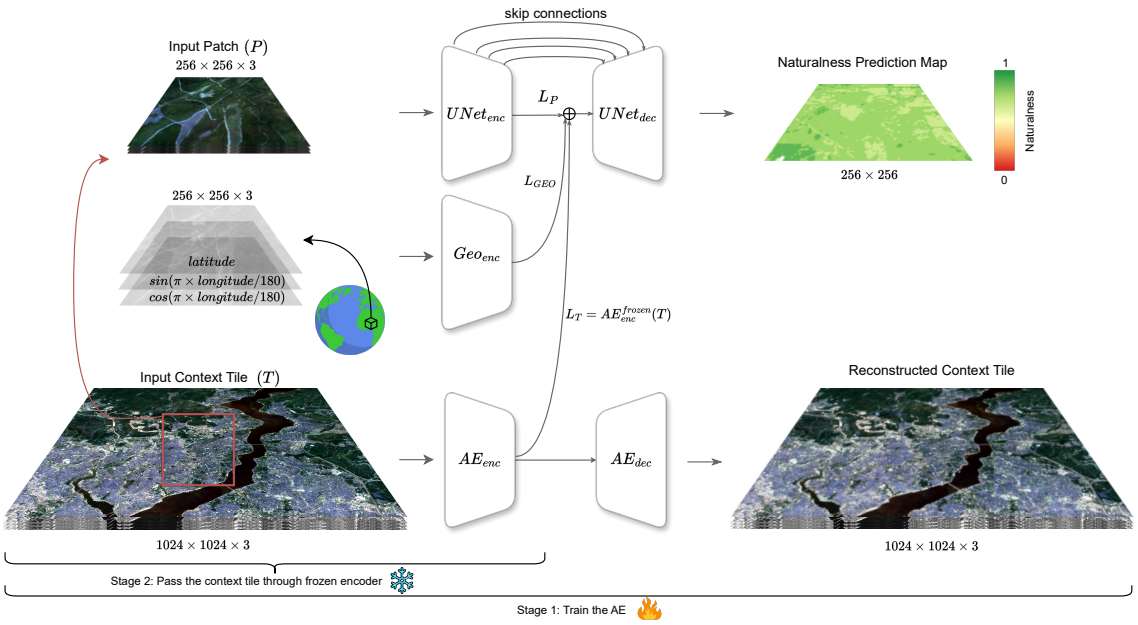


Figure 5.2: The proposed framework works as follows: Initially, an Autoencoder (AE) is trained to reconstruct the input context tiles, encoding the contextual information into a latent space. The broader context tiles pass through the now frozen AE encoder AE_{enc}, while the smaller patches cropped from these tiles are fed to the UNet encoder UNet_{enc} and the encoded coordinates are fed to the Geo_{enc}. The three high-dimensional latent representations obtained from all three encoders are then channel-wise concatenated and input into UNet_{dec} to produce the naturalness prediction map.

Traditional scalar representations fail to capture this continuity; for instance, $+179°$ and $-179°$ are geographically adjacent but numerically distant. The longitude coordinate C_{lon} is transformed into a cyclic format using sine and cosine functions:

$$C_{\text{cyclic}} = \left[\cos\left(\frac{2\pi \times C_{\text{lon}}}{180}\right), \sin\left(\frac{2\pi \times C_{\text{lon}}}{180}\right)\right]$$

The use of both sine and cosine is motivated by the $\pi/2$ phase difference between them, which prevents information loss at zero-crossings. The cyclically encoded geographic coordinates, shaped as $256 \times 256 \times 3$ (including the latitude), are fed into a dedicated encoder Geo_{enc} that structurally mimics the UNet_{enc}.

The incorporation of additional information, such as scale or context, is achieved via an Autoencoder (AE), which is designed for image reconstruction on larger context tiles T. The AE in this study encodes broader spatial information into a latent space, $L_T = \text{AE}_{\text{enc}}(T)$, and reconstructs the high-dimensional embedding with its decoder, AE_{dec}. The encoder segment AE_{enc} is architecturally similar to the UNet_{enc}, featuring convolutional layers with increasing channel sizes $(64, 128, 256, 512)$, followed by pooling layers to reduce spatial dimensions. Importantly, the broader context tiles T are shaped as $1024 \times 1024 \times 3$ (using Sentinel-2 bands B4, B3, B2) for the initial training of the AE.

The contextual information is leveraged by passing the larger context tile T through the frozen AE_{enc} to produce a latent representation L_T, which encapsulates the broader spatial context ($L_T = \text{AE}_{\text{frozen-enc}}(T)$). Concurrently, the UNet_{enc} maps the patch P into a latent space representation ($L_P = \text{UNet}_{\text{enc}}(P)$), and the dedicated encoder Geo_{enc} transforms the coordinates into a high-dimensional feature vector ($L_C = \text{Geo}_{\text{enc}}(C)$). These three latent representations are then concatenated channel-wise at the bottleneck to create a combined feature map:

$$L_{\text{combined}} = L_P \oplus L_C \oplus L_T$$

This combined feature map is then passed through the UNet_{dec} to generate the final naturalness map ($M = \text{UNet}_{\text{dec}}(L_{\text{combined}})$). The decision to inject the contextual and geographic embeddings at the bottleneck represents the point of maximum semantic compression, forcing the model to create a unified representation that integrates local patterns, broad spatial context, and absolute geographic location before beginning the pixel-wise reconstruction process in the decoder. The spatial correspondence between the context embedding L_T and the patch embedding L_P is learned implicitly by the UNet_{dec} during the regression task.

In summary, auxiliary supervision is integrated at the embedding level, introducing broader contextual information from the larger tile and spatial-geographic relations from the encoded coordinates. The latent space learned by the autoencoder captures high-level abstractions, which are considered to be useful for understanding spatial relationships from compressed data representations.

5.3 Distribution Shift for Large Scale Deployment

The underlying approach of this section is first presented in [138].

Let $\mathcal{X}$ represent the set of all possible data the model may encounter, and $\mathcal{Y}$ the set of class labels. The dataset on which a model is trained is defined as in-distribution (ID), denoted $D_{\text{ID}} \subset \mathcal{X}_{\text{ID}} \times \mathcal{Y}$. During inference, however, the model may encounter data from an unknown distribution, referred to as the WILD dataset $D_{\text{WILD}} \subset \mathcal{X}_{\text{WILD}}$, which may contain both ID and OOD samples.

A pre-trained neural network $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ is assumed, which extracts features $z \in \mathbb{R}^F$ from a specified internal layer. The objective is to distinguish between ID and OOD samples within D_{WILD}. This is achieved by fitting a binary classifier $g : \mathbb{R}^F \to \{0, 1\}$, parametrized by θ, to operate on these features z for a given sample in $\mathcal{X}_{\text{WILD}}$. It is defined that $g(z; \theta) = 0$ if $x \sim D_{\text{ID}}$, and $g(z; \theta) = 1$ if $x \sim D_{\text{OOD}}$.

The proposed method, TARDIS (illustrated in Figure 5.3), addresses the challenge that training a classifier g typically requires labeled OOD data, which is unavailable in real-world settings. TARDIS is a post-hoc method that operates with a pre-trained model f, known ID samples, and WILD samples. Its core mechanism involves generating surrogate ID/OOD labels for the WILD samples using their proximity to the known ID set in the feature space. This labeled set is then used to train the final binary distribution classifier g.

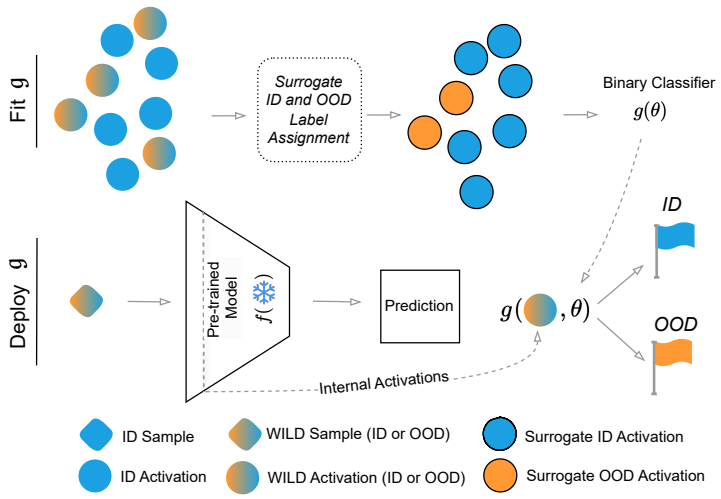


Figure 5.3: Overview of the proposed OOD detection method. Given a pre-trained model, a set of ID samples, and WILD samples (from unknown distributions), TARDIS assigns surrogate ID or OOD labels to WILD samples and fits a binary classifier g (top row). During deployment, g uses internal activations of an unseen sample to predict whether it is ID or OOD (bottom row).

TARDIS operates in a four-step workflow (Figure 5.4). Given M ID and N WILD samples, both are passed through the frozen model f, and internal activations z are

extracted from a specified layer. These activations are spatially downsampled (pooled) to obtain one-dimensional feature vectors.

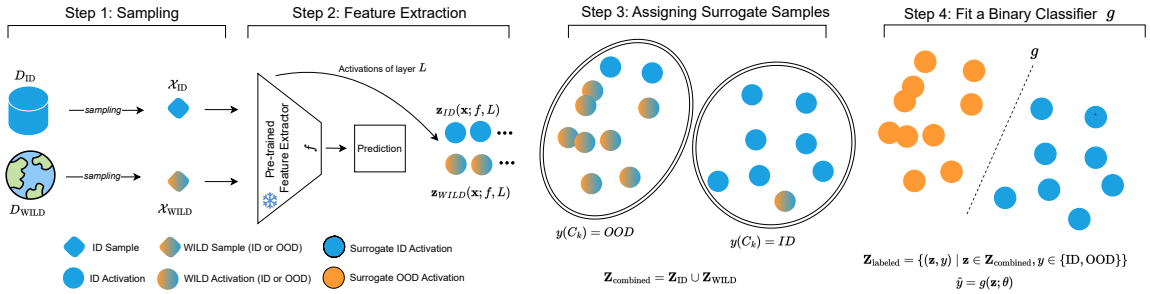


Figure 5.4: The proposed framework consists of four key steps: (1) Sampling in-distribution (ID) and WILD samples; (2) Extracting internal activations from a pre-trained model f for both ID and WILD samples; (3) Clustering the combined feature space and labeling WILD samples as surrogate-ID or surrogate-OOD; (4) Fitting a binary classifier g on the labeled feature representations to distinguish between ID and OOD samples. The classifier g, during deployment, flags out-of-distribution inputs.

The feature space, combined from ID and WILD samples, is clustered into k groups (e.g., using k-means clustering). The distance to the known ID samples then serves as a proxy for assigning surrogate labels to the WILD data. For each cluster, if the proportion of ID samples is greater than or equal to a threshold T, all samples within that cluster are assigned the surrogate label 0 (ID); otherwise, they receive the surrogate label 1 (OOD). This results in a fully labeled training set (original ID samples + surrogate-labeled WILD samples).

In Step 4, a binary classifier $g(z; \theta)$ is fitted on this fully labeled set to map feature representations z to their respective distribution labels. During deployment, activations are extracted from an unseen sample x, and its distribution label is predicted as $\hat{y} = g\left(\text{Downsample}\left(f(x)\right); \theta\right)$. The output can then be interpreted as a probability of domain shift or thresholded for binary OOD classification.

Optimal Hyperparameter Selection. To select the optimal clustering parameters (k and T) for surrogate labeling, a composite objective is minimized:

$$(k^*, T^*) = \arg\min_{k,T} \quad H(S) + P_{\text{mis-ID}} - P_{\text{corr-ID}}$$

where $H(S)$ is the average cluster entropy (encouraging cluster homogeneity), $P_{\text{mis-ID}}$ is the proportion of known ID samples assigned the surrogate OOD label, and $P_{\text{corr-ID}}$ is the proportion of known ID samples correctly assigned the surrogate ID label. This optimization balances three goals: achieving coherent clusters, minimizing ID data loss, and maximizing accurate ID classification, ensuring high surrogate label quality.

Chapter 6

Experiments and Results

This chapter details the experimental validation of the methodologies described in Chapter 5 using the MapInWild-NI dataset introduced in Chapter 4. The Naturalness Index serves as the supervision signal for the model. The resulting prediction output from the model trained on the MapInWild-NI dataset is referred to as DL-NI throughout the remainder of the thesis.

The initial phase establishes the evaluation metrics used throughout the work. A foundational analysis is performed to investigate the necessity of deep learning for the naturalness mapping task, comparing its performance against traditional machine learning approaches. This is followed by an investigation into a data-centric machine learning strategy, specifically for core-set selection. Subsequently, an experiment on influential modality selection is conducted to identify which specific data sources contribute most significantly to the task. Following the identification of the most influential data sources, methods are explored to make the model inherently more geospatially aware by incorporating geospatial priors. Finally, a distribution shift detector, designed for large scale deployment, is introduced. The culmination of these experiments is the final model deployed for continental-scale inference in Chapter 7.

6.1 Setting the Stage: Experimental Setup and Foundational Analysis

6.1.1 Evaluation Metrics

The model's predictive performance is evaluated using the mean absolute error (MAE, equation 6.1), the root mean squared error (RMSE, equation 6.2), and the mean error (ME, equation 6.3). The MAE is used to measure the average magnitude of prediction errors without considering their direction, indicating how far the predictions deviate from the annotation on average. The RMSE is computed as the square root of the mean squared error and is used to penalize larger errors more heavily, thereby increasing

sensitivity to outliers. The ME is used to quantify systematic bias, with negative values indicating underestimation (i.e., predictions systematically lower than the reference values), and positive values indicating overestimation.

In addition to these global metrics, balanced versions are computed—denoted as bMAE, bRMSE, and bME—to better reflect predictive performance across the full range of naturalness values. In the balanced setting, the respective error is calculated independently within ten equal-width (evenly spaced) bins of annotation naturalness values (e.g., $[0.0–0.1), [0.1–0.2), \ldots, [0.9–1.0]$), and the resulting per-bin errors are averaged. This approach ensures that each target interval contributes equally to the final score, regardless of its frequency in the dataset, and enables a more robust assessment of how predictive performance varies across different levels of naturalness.

While the standard MAE and RMSE provide an overall view of model accuracy and are suitable for model training and comparison, they are inherently influenced by class imbalance—errors in frequently occurring mid-range values dominate the result, potentially masking poor performance in underrepresented value ranges. In contrast, the balanced metrics (bMAE, bRMSE) offer a fairer and more interpretable perspective by equalizing the weight given to each target range. These are particularly useful for understanding whether predictive performance is consistent across the full spectrum of naturalness levels or biased toward more common ranges.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{6.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{6.2}$$

$$\text{ME} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) \tag{6.3}$$

where y_i and $\hat{y}_i$ denote the annotation and predicted naturalness values at pixel i, respectively, and N refers to the total number of valid pixels.

6.1.2 Assessing the Need for Deep Learning

This experiment is conducted to establish a foundational performance baseline and to technically justify the necessity of complex models. The motivation is to determine if the task's mapping—from 10-channel spectral input $X \in \mathbb{R}^{B\times 10\times H\times W}$ to a single-channel score map $Y \in \mathbb{R}^{B\times 1\times H\times W}$—is sufficiently simple to be solved by traditional regressors operating on a pixel-by-pixel basis. Here, the input data is all ten available channels of Sentinel-2. Out of all modalities in the MapInWild-NI dataset, Sentinel-2 is chosen for this foundational analysis because it provides the highest number of descriptive data

Table 6.1: Determining the appropriate learning formulation for the task: an empirical comparison of linear and non-linear learners.

Model	**MAE** (↓)	**RMSE** (↓)	**ME** (↓)
Linear Regressor	0.2856	0.3590	-0.158
MLP Regressor	0.2791	0.3448	0.0340
Random Forest Regressor	0.2798	0.3421	0.0500
U-Net	0.1995	0.2495	0.0020

channels (10 bands), thereby offering the richest possible feature set for the non-spatial solvers. This approach tests the hypotheses of linear separability and the sufficiency of non-linear spectral interactions alone, absent any spatial information.

To isolate spectral information, the input images are transformed from a spatial layout to a flat matrix, $X_{\text{flat}} \in \mathbb{R}^{(B \cdot H \cdot W) \times 10}$, where each row represents the 10-channel spectral vector of a single pixel. Four models are benchmarked:

1. Linear Regressor: A single nn.Linear layer mapping $\mathbb{R}^{10} \to \mathbb{R}^1$ with a *Sigmoid* output. This model tests the hypothesis of a linear relationship between the spectral bands and the target value.

2. MLP Regressor: A shallow neural network with one hidden layer of 64 units and $ReLU$ activation ($\mathbb{R}^{10} \to \mathbb{R}^{64} \to \mathbb{R}^1$). This model assesses the benefit of capturing simple non-linear spectral interactions.

3. Random Forest (RF) Regressor: A non-parametric ensemble (100 estimators, max_depth = 20). This serves as a strong non-linear, non-spatial benchmark capable of modeling complex partitions of the spectral feature space.

4. U-Net: A fully convolutional encoder–decoder architecture with skip connections [137], trained end-to-end on the original spatial tensors $X \in \mathbb{R}^{B \times 10 \times H \times W}$. The encoder captures hierarchical spatial–spectral representations via successive convolution and pooling, while the decoder reconstructs a dense prediction map through upsampling and concatenation with encoder features. This model uses both spectral and spatial context, representing the deep learning baseline.

The performances of the four models are summarized in Table 6.1. The results yield three clear observations. First, the modest performance improvement from the linear model to the MLP and RF models confirms that the relationship between spectral bands and the target is non-linear. Second, the high error of the Random Forest model demonstrates that spectral information from a single pixel is fundamentally insufficient for accurate prediction. Finally, the noticeable error reduction achieved by the U-Net underscores that incorporating spatial context drastically improves prediction performance. This provides empirical evidence justifying the use of deep learning models for this task.

6.1.3 Data Centric Machine Learning: Core-set Selection

Deep Learning models are often over-parameterized, which can lead them to learn strong biases and spurious correlations within the training data. When the dataset contains contradictory labels (noise) or when covariate shifts are present, the overparameterized model's high capacity causes it to memorize these spurious, often contradictory associations instead of learning the true, generalizable patterns. Core-set selection is a method used for identifying a small, highly representative subset of a larger training dataset [71]. For instance, Santos et al. [139] applied core-set selection logic to satellite time series data. They specifically used clustering to identify samples that were mislabeled or exhibited low prediction accuracy. By removing these noisy instances from the training set, their goal is to prevent a decrease in model performance caused by memorizing detrimental associations.

For instance, desert or barren land labeled as "low naturalness" in most samples yet labeled as "high naturalness" in a few noisy samples can force the model to learn long-range artifacts that degrade generalization. This cleaning procedure follows the rationale that samples consistently contributing disproportionately high error are likely conflicting with the dataset's main patterns or are fundamentally mislabeled, and thus should be discarded to improve the overall signal-to-noise ratio.

To implement this idea, the RMSE is measured per sample, and samples repeatedly exhibiting high error are removed. RMSE is chosen over other metrics because it disproportionately penalizes large errors, making it better at identifying inherently difficult or mislabeled tiles. In contrast, the MAE treats all errors uniformly and may miss problematic samples, while the ME reflects overall bias rather than providing insight into fundamental annotation issues.

Per-sample RMSE scores from the last three epochs of training are collected, and a 90th percentile threshold is used to flag tiles that exceed this threshold in at least two of the three epochs. A greedy iterative scheme is then applied: newly flagged tiles are compared to those identified in the previous iteration; consistently high-error samples remain flagged, and any that no longer exceed the threshold are unflagged. After each cleaning step, the model is retrained on the pruned dataset. This procedure is repeated until no new samples trigger removal. In total, 7 iterations are performed, leading to the exclusion of 86 out of the 1017 samples.

6.2 Influential Modality Selection

The experimental validation of the Deep Occlusion Framework (introduced in Chapter 5.1) consists of three main steps: first, the collection of sample-level and test set-level modality influence values for post-hoc analysis; second, a complementary ablation study

to validate influence scores against predictive performance drops; and third, an investigation into the robustness and versatility of the Modality Occluder by testing alternative occlusion strategies (e.g., random noise versus zeros). The proposed approach performs the embedding-level distance calculation and subsequent evaluations during the test phase.

6.2.1 Model and Training Details

The model uses a multi-modal input formed by concatenating all MapInWild sources into a 14-channel data cube ($C = 14$): 10 channels from Sentinel-2, 2 from Sentinel-1, a single ESA WorldCover band, and a single VIIRS DNB band. Sentinel-2 bands are normalized by dividing by 10,000; this domain-standard physical scaling converts the raw pixel values back to their native Bottom-of-Atmosphere (BOA) surface reflectance range of roughly [0, 1]. In contrast, because the remaining sources possess unbounded or fundamentally different physical scales, they are standardized using their respective mean (μ) and standard deviation (σ) values calculated from the training set. Data augmentation performs on-the-fly random cropping to a 256×256 pixel shape, along with random horizontal and vertical flips. To mitigate class imbalance within the NI annotations, the inverse mean value of each patch is utilized as sample weights during loss calculation. Training is optimized using the Adam optimizer with an initial learning rate $\eta = 10^{-4}$ and a batch size of 16. The loss criterion is the Mean Squared Error. To manage convergence and prevent overfitting, an early stopping mechanism with a patience of 40 epochs based on validation error is implemented. A learning rate scheduler with warm restarts is used, where the initial cycle length is set to three epochs. NoData values are explicitly excluded from all loss and error calculations. Model training and evaluation are performed using the TorchGeo geospatial deep learning library [140].

6.2.2 Post-Hoc Analysis of Sample-Level Modality Influence

The study calculates influence values for each test set sample to enable fine-grained, sample-level interpretations. These results provide insight into the attribution values assigned to individual input modalities during inference. For instance, Figure 6.1 demonstrates that within the context of residential areas, the nighttime lights modality exhibits a locally strong influence on the model's predictions. This is shown by the embedding distances measured when this modality is occluded. The sparsity of nighttime lights is hypothesized to provide easily discernible cues when present, serving as a distinct signal of human activity in built-up regions. Furthermore, Figure 6.2 illustrates a scenario where Sentinel-2 has the greatest influence on the prediction, followed by Sentinel-1 and Land Cover. While this focuses on modality attribution values for single samples or extreme cases, this analysis is extended to the full test set by averaging the attribution values across all samples, as described in the subsequent section (Figure 6.3).

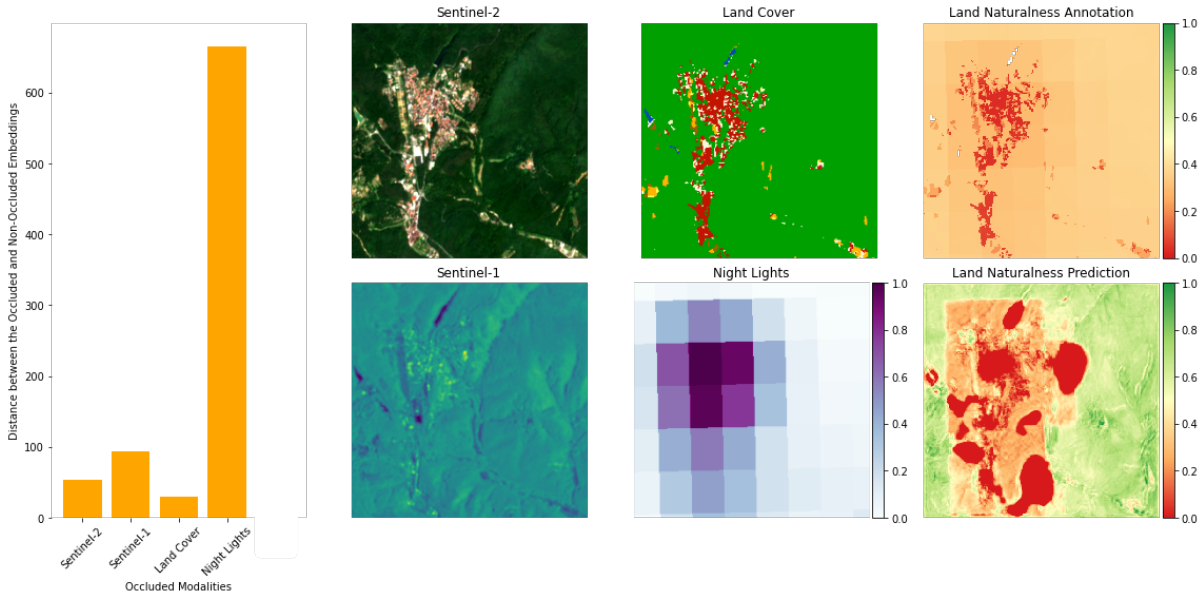


Figure 6.1: Sample-level influence scores, highlighting the impact of occluding the nighttime lights modality on the model's output. The dominant influence of nighttime lights is reflected in the significantly larger distance on the embedding level.

6.2.3 Validation via Ablation Study and Aggregated Influence

To validate the influence scores derived from the embedding distance, an ablation study is conducted. First, the test performance is benchmarked when each modality is individually input to the model (Table 6.2). Sentinel-1 and Sentinel-2 achieve the lowest MAE and RMSE among all single-modality inputs, followed by nighttime lights and land cover.

The table also presents results from experiments where the model is trained from scratch using the remaining modalities after one is occluded (X Occluded). The drop in predictive performance observed in these occlusion scenarios confirms that the occluded modality contributes unique, useful information. It is hypothesized that the individual performance metrics in this table serve as a quantitative proxy for the underlying importance of each modality, aligning with the qualitative findings of the deep occlusion method.

The study compares the MAE values from the single-modality scenarios against the baseline scenario (*All Modalities*). The resulting absolute differences, termed relative performance deviations, quantify the performance impact of relying solely on an individual modality. The modality influence values are then calculated over the test set using the proposed occlusion-based embedding distance method. Finally, a comparison is conducted between the aggregated influence scores and the relative performance deviations (Figure 6.3). The high degree of agreement in the distributions suggests that inferring relative influence information from the latent space is a valid method for generating input feature attribution values. This analysis confirms that in the test set, land naturalness prediction heavily depends on the *Sentinel-2* bands, followed by *Sentinel-1*. Conversely,

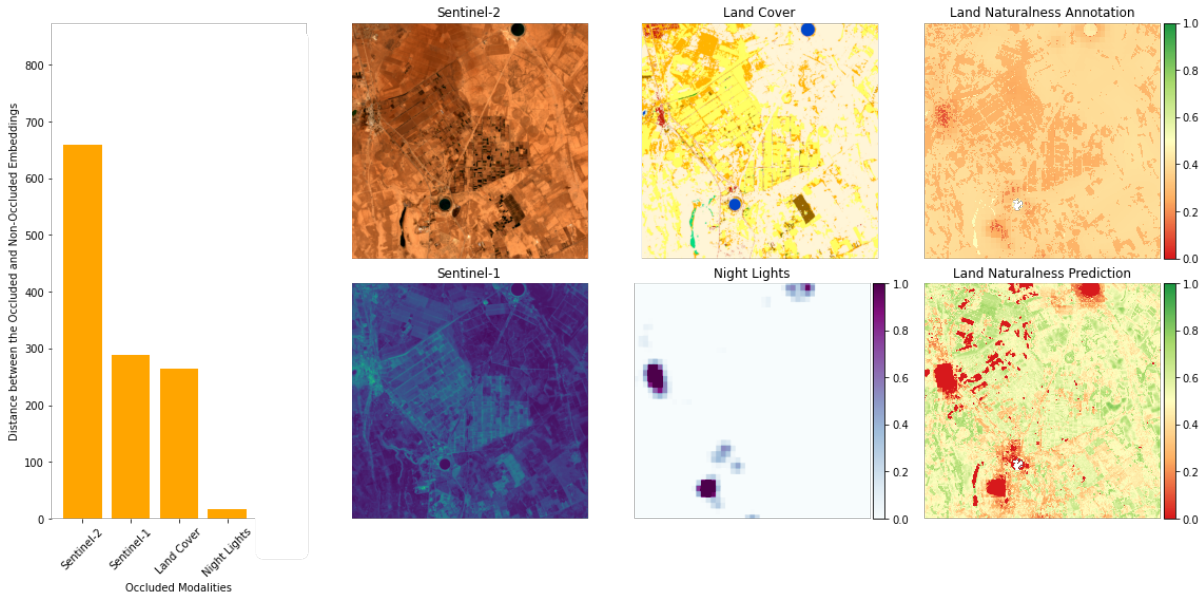


Figure 6.2: Sample-level influence scores, highlighting the impact of occluding the Sentinel-2 modality on the model's output, with other modalities influencing the prediction to certain extents. The relatively higher impact of Sentinel-2 is reflected in the larger distance on the embedding level.

Modality	**MAE** (↓)	**RMSE** (↓)
All Modalities	0.109	0.136
Sentinel-2	0.125	0.159
Sentinel-1	0.123	0.157
Land Cover	0.128	0.163
Night Lights	0.126	0.157
Sentinel-2 Occluded	0.170	0.194
Sentinel-1 Occluded	0.145	0.166
Land Cover Occluded	0.130	0.161
Night Lights Occluded	0.132	0.173

Table 6.2: Ablation Study: Influence metrics per modality

the other modalities exhibit relatively weak influence on average. This global finding stands in contrast to the sample-level observation in Figure 6.1, where nighttime lights demonstrate locally strong influence; the aggregated result suggests that while NL is crucial in specific built-up contexts, its overall importance is diminished when averaged across the entire test set.

The model demonstrates a high reliance on the Sentinel-2 bands, which primarily capture the visible and near-infrared portions of the electromagnetic spectrum. Contrastingly, Sentinel-1 bands (VV and VH) provide less predictive power for this specific task. While SAR is sensitive to geometric structure (e.g., buildings), the concept of 'naturalness' in this context relies heavily on biophysical indicators (vegetation health, bare soil indices) which are best captured by optical reflectance.

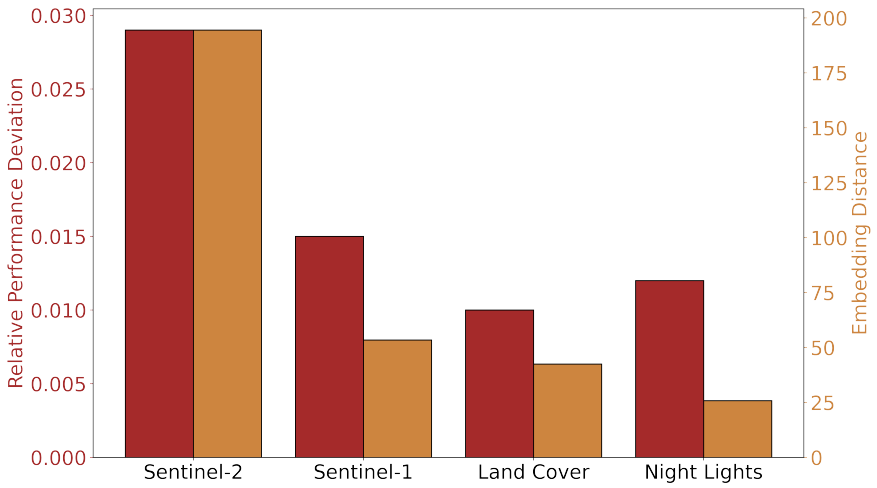


Figure 6.3: Comparison of relative performance deviations and embedding distances for each modality, aggregated over the entire test set. Red bars represent performance deviations from an all-modality scenario, while orange bars show the embedding distances between occluded and non-occluded embeddings, measuring their influence.

The ESA WorldCover land cover data does not significantly contribute to the model's decision-making process. This is indicative of a certain level of redundancy between the spectral data and the land cover classes, suggesting that the Sentinel-2 bands may already encapsulate much of the information required to assess naturalness, particularly since ESA WorldCover is itself derived from Sentinel-1 and Sentinel-2 data.

The VIIRS DNB (Nighttime Lights) data, on the other hand, makes a contribution to the model's predictive capacity. Its ability to capture low-light emissions, both natural and anthropogenic, is argued to provide a unique perspective on human activity, serving as a direct or inverse proxy for landscape naturalness depending on the context.

6.2.4 Robustness Check via Occlusion Strategy Investigation

In the final experiment, four occlusion strategies—filling with zeros, ones, random noise, and Gaussian noise—are applied to the input modalities to observe their impact on the model's representations. Each occlusion method alters the input modality in a unique way, simulating the absence of information or introducing variability to gauge the model's dependency on the given modality.

The resulting pattern (Figure 6.4) suggests a consistent trend across all occlusion types regarding the relative influence of each modality. Sentinel-2 and Sentinel-1 modalities appear to have a more significant impact on the embedding space, as indicated by larger distances, suggesting these modalities are highly influential in the model's prediction process. The coherence of these findings across different occlusion strategies reinforces the reliability and robustness of the occlusion sensitivity analysis as a method for interpreting model behavior through its manipulated embedding space. The agreement between the

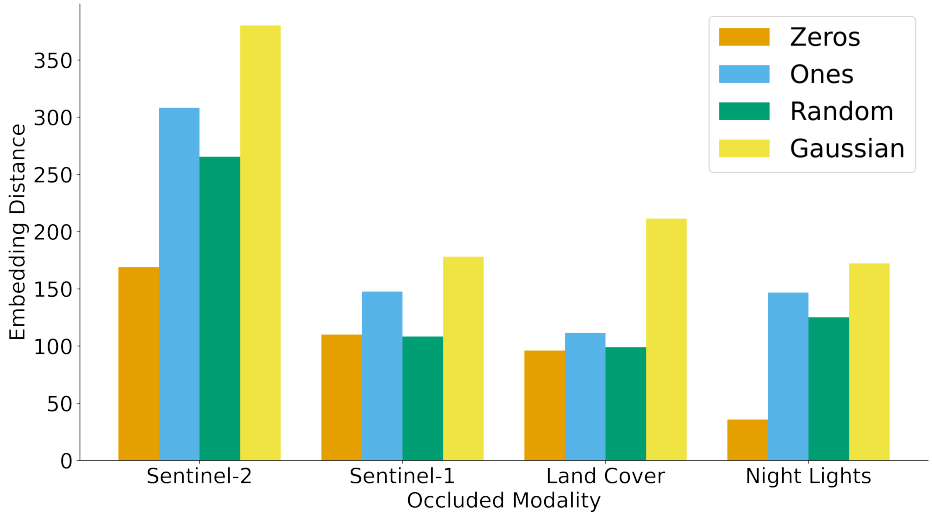


Figure 6.4: Comparison of embedding distances for different occlusion strategies across modalities. Each bar represents the embedding space perturbation when a specific modality is occluded with zeros, ones, random noise, or Gaussian noise.

occlusion types supports the conclusion that despite the difference in how information is withheld or distorted, the model consistently identifies the same modalities as more or less influential, which is a promising result for the robustness of the proposed method.

6.3 Incorporating Geospatial Priors

The Influential Modality Selection experiment (Section 6.2) demonstrates that the Sentinel-2 bands provide the most informative signals for the task of naturalness mapping. For this reason, the remainder of the thesis uses Sentinel-2 data exclusively as the core input for all subsequent experiments and analysis. While the full 10-band Sentinel-2 input may offer marginal predictive utility, the final model's input stack is constrained to the three visible bands (Red, Green, and Blue) to ensure operational feasibility during continental-scale deployment and to leverage pretrained weights. This decision mitigates the computational and storage burden associated with large scale multi-band inference and context tile acquisition. Specifically, the model uses the pretrained weights Sentinel2_Resnet50_SI_RGB [141]. This model was trained on 856K Sentinel-2 tiles at zoom level 13 and has demonstrated superior performance across multiple downstream tasks, justifying its selection as a robust feature extractor.

In this section, two experimental setups are formed to compare the baseline model (a plain UNet) against the proposed, geospatially-aware model (Figure 5.2). The proposed model uses Sentinel-2 as input, complemented by its geographic coordinates and a wider context tile (T). Quantitative results are tabulated in Table 6.3, and qualitative results are provided in Figure 6.5.

Features	MAE ↓	MSE ↓	MSSIM ↑
Baseline	0.24	0.29	0.78
+ Coordinate and Context	0.14	0.25	0.89

Table 6.3: The comparison study between the baseline (plain UNet) and the proposed model (UNet augmented with Coordinate and Context embeddings).

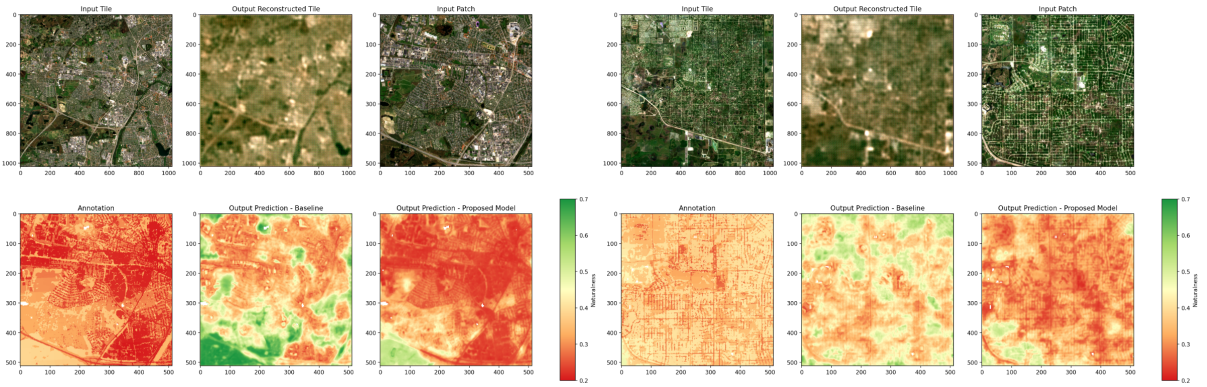


Figure 6.5: Inference results of the patches with dataset IDs 900000061 and 900000068. The areas cover parts of Cape Coral, Florida, USA (left), and Copenhagen, Denmark (right). For each prediction, the first row shows the context tile (T), its reconstruction, and the main input patch (P). The second row shows the naturalness annotation and the prediction results of the baseline and proposed models.

This comparison strongly validates the need for incorporating geospatial priors. As shown in Table 6.3, the inclusion of geographic coordinates and broader context information leads to a substantial improvement across all metrics. The MAE drops by 42% (from 0.24 to 0.14), and the structural similarity (MSSIM) increases significantly from 0.78 to 0.89. This quantitative shift suggests that the model is no longer relying solely on purely local pixel patterns but is leveraging the auxiliary embeddings drawn from spatial relationships and broader contextual information, leading to more robust and accurate predictions of naturalness. The patches are cropped from the center of context images, and the reconstructed tiles are presented to demonstrate the performance of the AE. Details on the dataset and implementation can be found in Appendix A.

The visual results in Figure 6.5 further confirm this, particularly for heterogeneous landscapes. The proposed model consistently shows enhanced predictive performance for areas with complex spatial features, such as sharp transitions between urban structures (Cape Coral) and natural landscapes (Copenhagen). This suggests that providing the model with explicit geographic context helps it to disentangle spurious local correlations and adhere more closely to real-world geographic principles.

6.4 Final Model

The final model, using RGB bands of Sentinel-2 as input, achieves a competitive level of predictive performance on the test set. Global test set evaluation yields a MAE of 0.137 and a RMSE of 0.173. The ME, which detects systematic bias, is -0.038. This value indicates a minor residual overall underestimation by the model.

To investigate performance across the full range of naturalness values, balanced metrics are computed by averaging errors over ten equal-width bins (Figure 6.6). Per-bin analysis confirms that the prediction error remains non-uniform, with MAE and RMSE being consistently highest at the extreme ends of the annotation scale, particularly in the $[0.8, 1.0]$ bin, where the error measures nearly MAE $= 0.23$ and RMSE $= 0.25$.

These findings suggest that the model's core limitation is a persistent central prediction tendency. While the overall bias is near zero, this conservative behavior leads to the overestimation of low NI values and the underestimation of high NI values.

Analysis of Performance Per Annotation Range Figure 6.6 presents the MAE, RMSE, ME calculated per bin. The per-bin analysis clearly shows the systematic divergence in prediction bias. In the lower and moderate bins (approximately [0.0–0.6]), positive mean errors are observed, indicating the model consistently overestimates the naturalness score. Conversely, negative mean errors dominate the higher bins (approximately [0.6–1.0]), confirming that the model systematically underestimates the highest naturalness scores (which is validated by the train set annotation distribution illustrated in Figure 4.14). This S-curve pattern in the Mean Error plot is definitive evidence of a central prediction tendency (or regression toward the mean), where systematic errors with opposing signs at the extremes effectively cancel out, leading to a deceptively small global ME.

Distribution Comparison Figure 6.7 visualizes the disparity between annotation and predicted values. The annotations exhibit a broad, multi-modal distribution spanning $[0, 1]$ (with strong peaks near 0.8 and 0.9), whereas the model's predictions show a significant shift toward the mid-range. The predicted distribution remains wide but becomes predominantly unimodal, over-concentrating its values and spiking sharply around 0.6, effectively minimizing the distinct high-value modes seen in the annotation (near 0.8 and 0.9). This concentration confirms the model's central tendency, explaining the high errors at the spectrum's extremes. This conservative behavior suggests the limitation stems from training data characteristics, specifically data scarcity at the NI scale extremes, or potentially inconsistent/misleading NI annotations near the 0.0 and 1.0 boundaries, which prevents the model from confidently generating accurate outlier predictions.

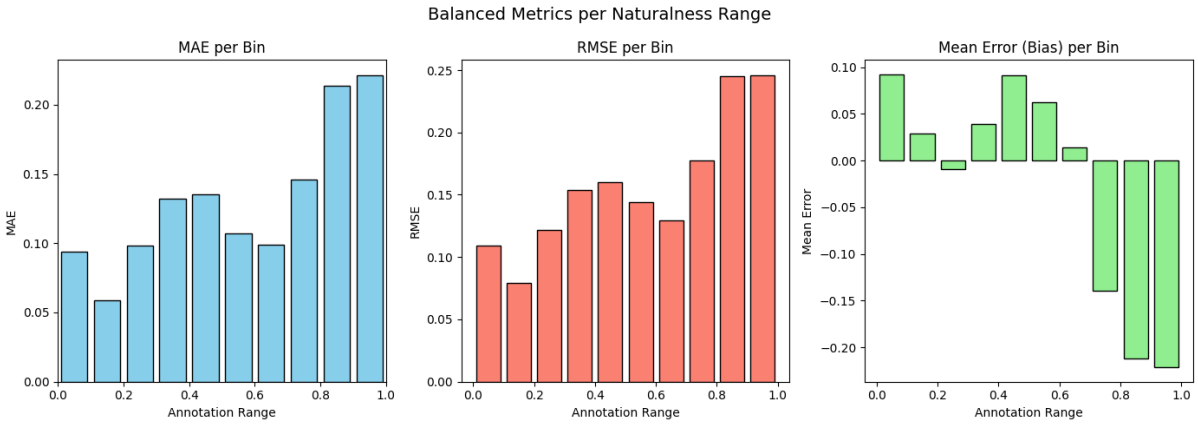


Figure 6.6: Performance of the final model across five equal-width bins of the NI annotations. The bar plots illustrate the MAE, RMSE, and ME per bin, revealing a clear trend of performance variation and systematic bias depending on the annotation range.

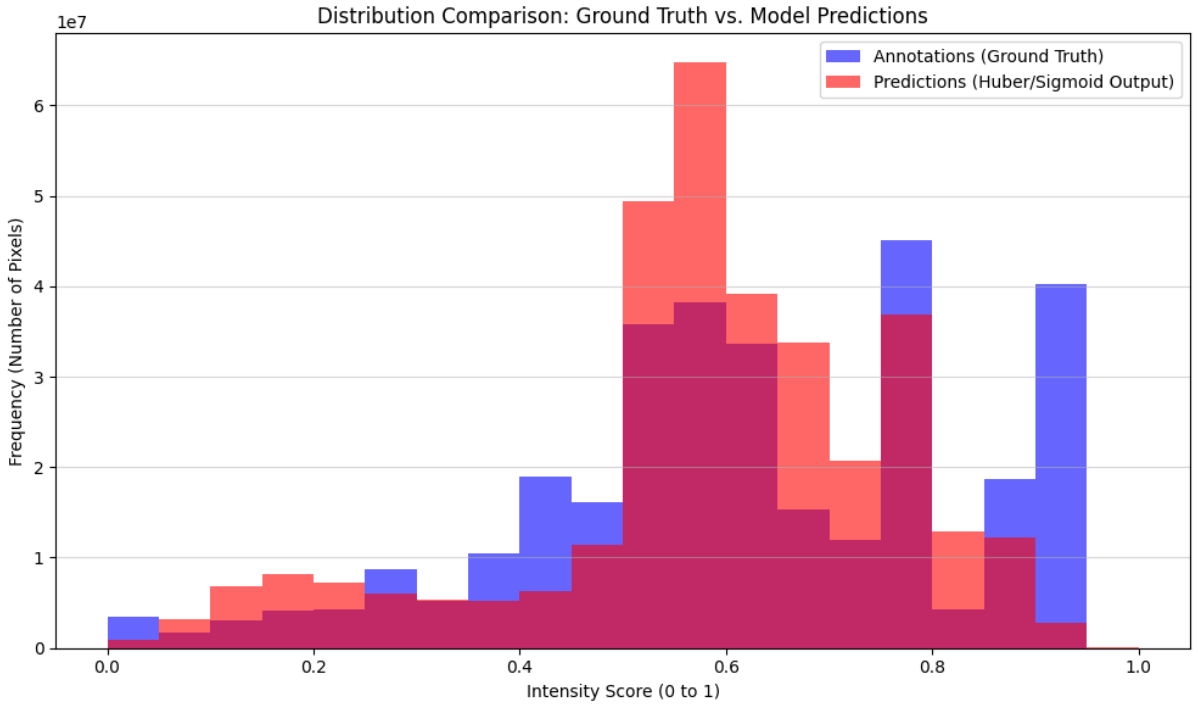


Figure 6.7: Distribution comparison between the NI annotation and the final model's output. The sharp contrast between the annotation's broad, multi-modal distribution and the model's tightly clustered predictions around the mean highlights a regression toward the mean phenomenon.

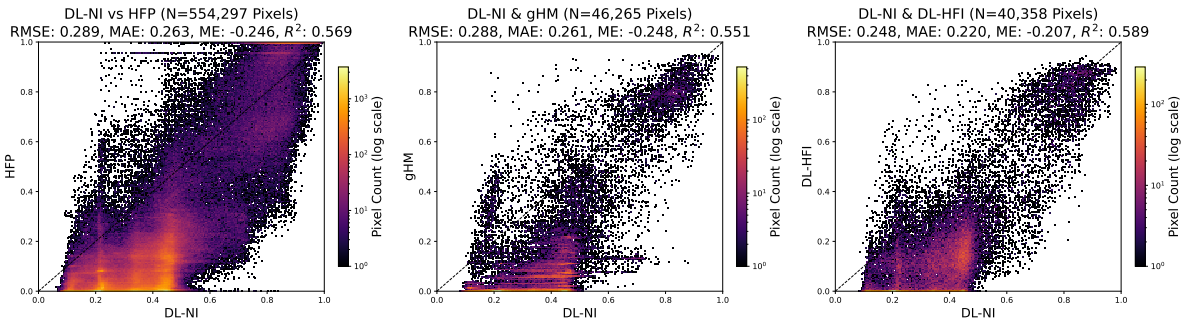


Figure 6.8: Density scatter plots illustrating the correlation between the 10 m resolution DL-NI map and the human influence indices: HFP, gHM, and DL-HFI.

Having evaluated the models' predictive performance, the following evaluations compare the DL-NI with existing human influence indices: the second-generation Human Footprint Index (HFP) [35], the Global Human Modification map (gHM) [3], and the Deep Learning-based Human Footprint Index (DL-HFI) [41]. The selection of these three indices is motivated by their distinct methodologies and foundational roles in the field of human influence mapping. The HFP (2020, 300 m) and gHM (2016, 1.1 km) are established foundational rule-based indices that provide similar yet distinct representations of human pressure. Specifically, the HFP approaches human influence by aggregating 9 different stressor variables, while the gHM aggregates 13 different stressor variables. Additionally, the underlying geospatial layers that contribute to the calculation of the resulting maps are of different resolution and source. The DL-HFI (2019, 1.1 km) serves as the direct deep learning baseline because it is the only existing study, at the time of writing, that uses satellite imagery alone to map human influence using deep learning and has the resulting data publicly available for evaluation. All three indices are products of multi-temporal studies, compiled over several years. To minimize temporal discrepancy, the latest available map iteration is selected for each index; however, it is important to note that a temporal gap still exists between the reference maps (2016–2020) and the proposed DL-NI (2024), which provides a much finer spatial resolution of 10 m, representing a significant spatial improvement of 110-fold compared to the only existing deep learning-based study (the 1.1 km DL-HFI). The same set of indices is also used for comparison against the continental DL-NI map in Chapter 7.

Inter-Map Correlation Analysis Figure 6.8 displays density scatter plots comparing the DL-NI with the three indices for all test samples in MapInWild-NI dataset. The analysis confirms a fairly strong statistical correlation (R^2 values ranging from 0.551 to 0.589), indicating that the DL-NI captures the same broad patterns of human influence as the established indices. This empirical agreement is essential for validating the DL-NI as a reliable successor to these coarser maps.

Despite the overall correlation, the plots appear to reveal systemic differences inherent to the change in spatial resolution. All three scatter plots show a high concentration of pixels along the diagonal identity line, especially in the mid-to-high naturalness range (0.6 to 1.0). However, for the DL-NI vs. HFP and DL-NI vs. gHM plots, there is a substantial mass of pixels (low pixel count on log scale) clustered near the bottom-left corner (0.0 on both axes). This cluster represents areas of very low naturalness where the DL-NI assigns a low score, yet the original 1 km or 300 m indices show considerable disagreement, likely due to the spatial blurring effect of their coarse resolution.

Qualitative Analysis of Prediction Maps Figure 6.9 shows several test samples from the dataset, displayed alongside their corresponding NI annotations and DL-NI predictions, which are then compared to the existing benchmark metrics introduced in the literature. As discussed in Section 3.1.1, HFP appears to be overfitting to the pattern of certain accessibility layers, such as roads, a geo-proxy used in the formation of the composite metric. This observation highlights the downside of having to manually and heuristically combine individual pressure layers. Another observation is that gHM and DL-HFI tend to over-assign human influence values to patches where visual evidence indicates low anthropogenic presence (e.g., sample ID: 6138). The patch (ID: 5955) demonstrates the capabilities of the DL-NI model benefiting from contextual information, where the low human influence value of an inner-city park appears to be correctly modulated by the high anthropogenic pressure surrounding it. Overall, while the DL-NI predictions exhibit structural patterns with sharper features (e.g., roads, clear boundaries), the model reasonably captures the expected transition of human influence intensity not as a sharp binary shift, but as a more gradual gradient (e.g., sample ID: 90000004). This gradient-like transition is ecologically sensible as human influence is expected to change gradually across a landscape. Further, in most instances, the benefit of the higher resolution provided by the DL-NI predictions is evident, allowing for superior delineation of human influence at a fine scale.

6.5 Out-of-Distribution Detection

This section first describes the preparation of the Distribution Shift Detector (introduced in Section 5.3) and subsequently evaluates its performance on benchmark datasets.

6.5.1 Preparing the Distribution Shift Detector

The application of the distribution shift detection method involves three main stages: feature extraction, surrogate label assignment, and binary classifier training. Once prepared, the trained classifier is ready to be integrated with the model for deployment.

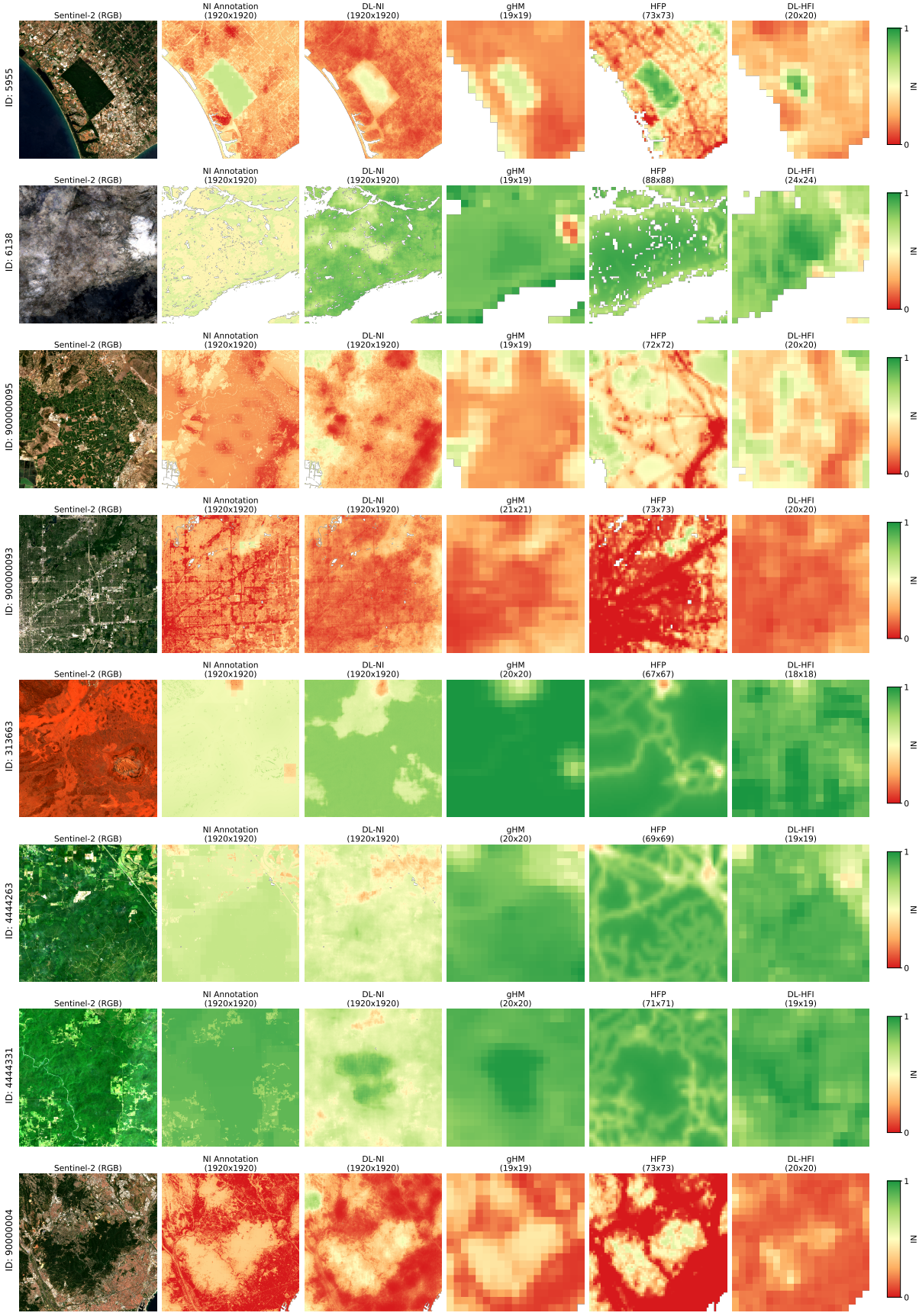


Figure 6.9: Qualitative comparison of globally-distributed test regions. Each row shows the Input Sentinel-2, NI annotation, the final model's prediction (DL-NI, trained on NI), and the three benchmark indices: gHM and HFP (rule-based), and DL-HFI (DL-derived).

First, both In-Distribution (ID) and WILD samples are processed by the pre-trained deep learning model (f) to extract intermediate features. The feature vector z is obtained by registering a forward hook to the `model.encoder.layer4` module. This layer, situated in the encoder's bottleneck, is selected because bottleneck features are known to provide compressed yet informative representations, which is beneficial for transfer and detection tasks. The resulting high-dimensional feature map ($\mathbb{R}^{C\times H\times W}$) is subsequently reduced to a single feature vector ($\mathbb{R}^{C}$) using Adaptive Max Pooling for spatial downsampling.

The extracted intermediate features populate the activation space, which is then partitioned using K-Means clustering to assign surrogate labels. The number of clusters (K) is set as 10% of the total number of extracted features. A maximum ID proportion threshold is chosen such that any cluster where the proportion of ID samples is greater than $T = 0.1$ is flagged as Out-of-Distribution (1); otherwise, it is flagged as In-Distribution (0). This process effectively assigns pseudo-labels to the WILD samples. The resulting combined ID and pseudo-labeled OOD dataset is split into training and testing sets using an 80/20 ratio.

The binary classifier (g) is a Logistic Regression model. To ensure robustness against potential class imbalance between the assigned ID and OOD samples, the model's features are pre-processed using *StandardScaler* (removing the mean and scaling to unit variance), and the classifier is configured with the `class_weight="balanced"` setting.

Upon training, the final classifier is integrated as a wrapper to the model. This augmented model operates dually during inference: it provides the primary task prediction (e.g., naturalness map) while simultaneously outputting the distribution shift probability (ranging from 0 to 1) for the given image. This multi-task framework allows for the quantification of prediction reliability during large scale deployment.

6.5.2 Detecting Controlled Shifts

While the previous sections evaluated the model on the specific task of Naturalness mapping, validating a Distribution Shift Detector requires controlled environments with known ground-truth shifts (e.g., specific disaster types or land cover changes). Therefore, TARDIS is first benchmarked on benchmark community datasets (EuroSAT, xBD) to establish its theoretical validity before being applied to the Naturalness Map in the deployment phase.

To this end, the experimental setups used to evaluate under controlled distribution shifts are described. The EuroSAT [142] dataset is used for patch-level classification and the xBD [143] dataset is used for semantic segmentation. EuroSAT contains 27,000 labeled images across ten land use/land cover classes. xBD includes pre/post-disaster imagery labeled for building damage assessment, covering hurricanes, floods, wildfires, and more. For the purposes of this study, xBD is formulated as a building footprint

Table 6.4: Overview of experimental setups for evaluating under controlled distribution shifts. The table details the covariate and semantic shift scenarios, including the dataset used, split method, and the composition of the ID and out-of-distribution OOD sets. All conditions are post-event except where specified.

Type	Dataset	Split Method	ID (Training Set)	OOD (Test Set)
Covariate Shift	EuroSAT	Spatial Split (Longitude)	West Europe	East Europe
	xBD	Similar Disasters - Distant Locations	Nepal Flooding	Midwest Flooding
		Similar Disasters - Nearby Locations	Santa Rosa Wildfire	Woolsey Fire
		Different Disasters - Distant Locations	Hurricane Matthew	Nepal Flooding
		Different Disasters - Nearby Locations	Hurricane Matthew	Mexico Earthquake
		Temporal	Portugal Wildfire (Pre-Disaster)	Portugal Wildfire
Semantic Shift	EuroSAT	Unseen Class (Repeated for 10 Classes)	9 out of 10 classes (e.g., Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River)	1 holdout class (e.g., Sea/Lake)

segmentation task with two classes: building footprint and background. These datasets offer diverse evaluation conditions in terms of volume, task, sensor, spatial/temporal coverage, and geolocation, making them suitable test-beds for the OOD method.

To simulate controlled distribution shifts, the train, validation, and test splits in both datasets are adjusted so that the model, trained on the modified training set, encounters realistic shifts during testing. Two shift types are evaluated: covariate shift, where input distribution changes between train and test, and semantic shift, where unseen classes appear at test time. Covariate shifts include spatial (different locations), temporal (different time periods), and environmental (different landcover/disaster types) variance. For semantic shift, the model is trained on nine of ten classes and is tested on the held-out class, rotating through all classes. Table 6.4 summarizes these setups.

These setups assume access to clear ID (train) and OOD (test) labels—an unrealistic assumption in real-world deployments. To address this, the method is evaluated in two ways. First, known ID and OOD labels are used to directly train and evaluate the binary classifier g, referred to as g_{oracle}. Second, test-time OOD labels are treated as unknown

(i.e., WILD) and the surrogate label assignment is applied to generate surrogate ID and OOD labels. This version, trained and evaluated using surrogate labels, is denoted as g^*. The version g_{oracle} is considered as an upper bound on performance, since it uses ground-truth labels, while g^* relies on inferred surrogates.

Table 6.5 shows results for g^* and g_{oracle} across 17 setups. Overall, g^* closely approaches g_{oracle}, which serves as the upper bound. In 13 setups for AUROC and all 17 for FPR95, the performance gap is not statistically significant. In the few remaining cases, the gap is small. These results show that TARDIS can reliably assign surrogate labels without access to test-time labels, thereby validating its utility across diverse EO distribution shift types.

Benchmark studies on the g_{oracle} classifier are conducted to explore the factors influencing the detector's ability to detect test-time distribution shifts. This ability depends significantly on the parameters k (the number of clusters) and T (the maximum ID proportion threshold). That is why a comparison of g^* to g_{oracle} as the parameters k and T are tuned is required for surrogate label assignment. To determine k and T, a Tree-structured Parzen Estimator is used for sampling. Specifically, the search boundaries are set for k between 2 and $0.3 \times M$, where M is the total number of samples, and for T (the fraction of in-distribution samples within a cluster, below which the cluster is labeled as surrogate OOD) between 0.01 and 0.2. These values are empirically determined based on preliminary tuning for optimal performance. 20 independent experiments are then run, and the best-performing (k, T) pair is selected. Fixing T based on the hyperparameter search results, a recurring pattern is observed that enables fixing k to $0.3 \times M$ across all setups. This choice is based on the observation that the classifier g^* reaches performance levels close to g_{oracle} when k is set to 30% of the training samples. This trend holds for both datasets, with the xBD disaster scenario (semantic shift) and the EuroSAT Pasture experiment (covariate shift) showing similar behavior. In both cases, performance improves as more clusters are added, approaching that of g_{oracle}. In Figure 6.10, the process used to determine the optimal (k, T) pair that achieves high surrogate label assignment performance across all setups is shown. The same pair generalizes to real-world deployment, removing the need for hyperparameter tuning.

Table 6.5: Performance metrics comparing the oracle classifier g_{oracle} and the surrogate classifier g^* for EuroSAT and xBD datasets across various experimental setups. The g_{oracle} classifier acts as an upper bound for the g^*, which is the proposed method. The † notation indicates that over 10 measurements, the difference is not found to be statistically significant ($p < 0.05$). All conditions are post-event except where specified. T: Distribution Shift Type, D: Dataset.

T	**D**	**OOD (Test Set)**	**ID (Training Set)**	**AUROC↑**		**FPR95↓**	
				g_{oracle}	g^*	g_{oracle}	g^*
Covariate Shift	EuroSAT	Eastern Europe	Western Europe	0.91 ± 0.04†	0.89 ± 0.06†	0.32 ± 0.11†	0.37 ± 0.10†
	xBD	Midwest Flooding	Nepal Flooding	1.00 ± 0.00†	1.00 ± 0.00†	0.01 ± 0.02†	0.01 ± 0.02†
		Woolsey Fire	Santa Rosa Wildfire	0.94 ± 0.02†	0.93 ± 0.02†	0.23 ± 0.08†	0.27 ± 0.09†
		Nepal Flooding	Hurricane Matthew	0.99 ± 0.01	0.98 ± 0.01	0.06 ± 0.04†	0.07 ± 0.02†
		Mexico Earthquake	Hurricane Matthew	0.96 ± 0.02†	0.94 ± 0.03†	0.11 ± 0.05†	0.19 ± 0.08†
		Portugal Wildfire (Pre)	Portugal Wildfire	0.99 ± 0.00	0.95 ± 0.01	0.06 ± 0.04	0.23 ± 0.20
		Santa Rosa Wildfire	Woolsey Fire	0.94 ± 0.02†	0.93 ± 0.02†	0.23 ± 0.08†	0.27 ± 0.09†
Semantic Shift	EuroSAT	Forest	Remaining 9 classes	0.99 ± 0.01†	0.97 ± 0.02†	0.08 ± 0.04†	0.08 ± 0.05†
		Herbaceous Vegetation		0.93 ± 0.03†	0.88 ± 0.07†	0.26 ± 0.11†	0.29 ± 0.14†
		Highway		0.63 ± 0.08†	0.56 ± 0.06†	0.78 ± 0.13†	0.88 ± 0.10†
		Industrial		0.97 ± 0.02†	0.96 ± 0.02†	0.13 ± 0.14†	0.22 ± 0.07†
		Pasture		0.98 ± 0.01	0.95 ± 0.01	0.09 ± 0.06†	0.28 ± 0.19†
		Permanent Crop		0.92 ± 0.03†	0.89 ± 0.03†	0.26 ± 0.08†	0.33 ± 0.11†
		Residential		0.87 ± 0.03†	0.82 ± 0.06†	0.46 ± 0.21†	0.43 ± 0.15†
		River		0.87 ± 0.04	0.81 ± 0.06	0.44 ± 0.14†	0.57 ± 0.10†
		Sea-Lake		1.00 ± 0.00†	1.00 ± 0.00†	0.00 ± 0.00†	0.00 ± 0.00†
		Annual Crop		0.92 ± 0.03†	0.91 ± 0.04†	0.25 ± 0.10†	0.32 ± 0.16†

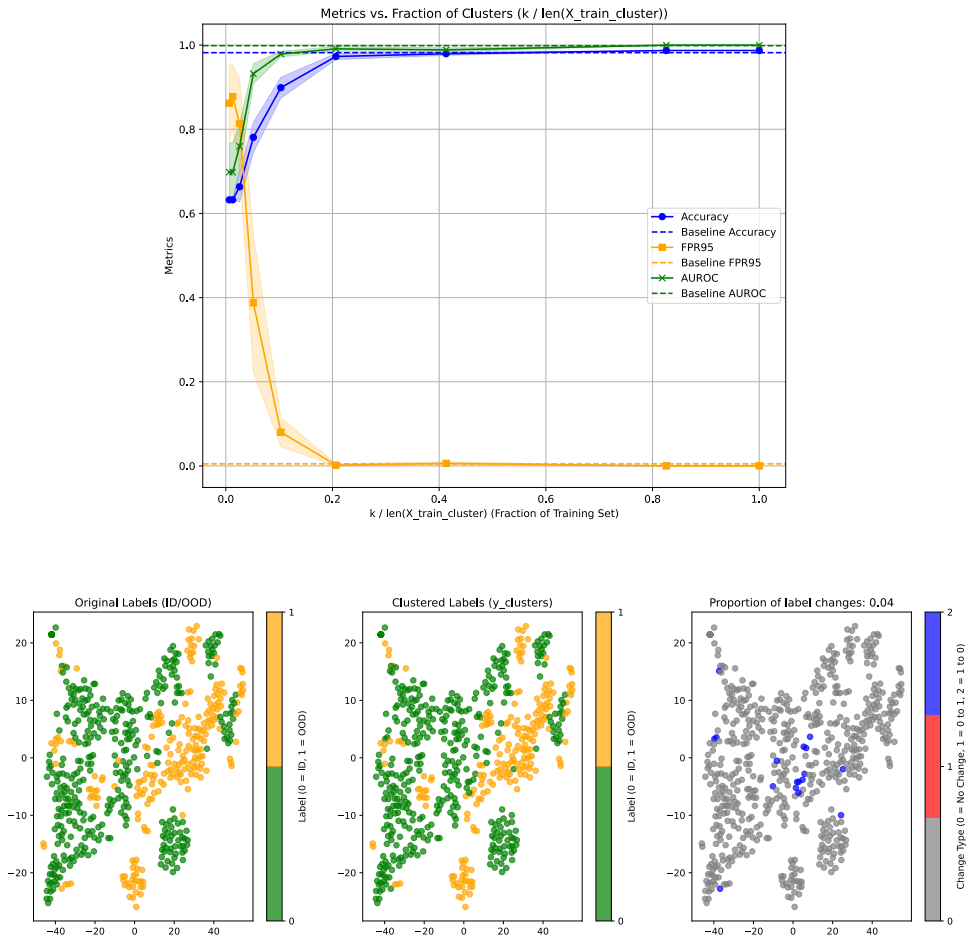


Figure 6.10: xBD *Nepal Flooding-Midwest Flooding* disaster experiment on surrogate label assignment. The upper plot shows the performance metrics (Accuracy, FPR95, AUROC) for the oracle classifier g_{oracle} and the surrogate classifier g^* as the ratio of clusters to training samples $k/\text{len}(X_{\text{train}})$ increases. As k grows, g^* gradually improves and approaches the performance of g_{oracle}. The lower plot visualizes the feature space before and after clustering, showing how original ID and OOD labels are reassigned to surrogate ID and OOD labels based on the clustering logic.

6.6 Discussion

6.6.1 Influential Modality Selection

The experimental results confirm the significance of each input modality for the naturalness mapping task. The core finding—that the task is heavily reliant on a subset of the available data—justifies the strategic decision to simplify the input stack for the final model. Although the initial impulse might be to utilize all data for maximum predictive performance, this approach is often computationally sub-optimal, especially when the end goal is deployment at a continental scale; it is often noted that greater data volume can also introduce greater noise. Focusing on the most influential modality significantly reduces the computational overhead and storage requirements, thereby making large scale inference feasible.

Regarding the deep occlusion method used for feature attribution, the current implementation relies on the Euclidean distance between latent space embeddings to quantify influence. While this simple distance metric effectively serves its purpose of identifying key modalities, exploring more sophisticated distance or divergence measures (e.g., Kullback-Leibler divergence or specialized manifold distances) could potentially provide a more nuanced and accurate quantification of how feature occlusion perturbs the model's internal representation. However, for the specific objective of rank-ordering input influence, the current distance-based approach proves sufficient and robust.

6.6.2 Incorporating Geospatial Priors

The experiments on incorporating geospatial priors demonstrate a notable boost in performance compared to the baseline model, directly validating the core assumption that geographic context is crucial for accurate naturalness mapping. The primary focus of this experiment is not to find the single optimal method for geospatial conditioning, but to provide a fundamental and deployable framework for integrating such prior knowledge. The use of simple coordinates and a lightweight autoencoder strikes a necessary balance: it leverages contextual information without relying on resource-intensive, recent location encoding methods. This strategy is essential, as the final deployed model must remain computationally lightweight for efficient continental-scale inference.

6.6.3 Analysis of Final Model Performance and Trade-offs

As described, the final model's architecture—utilizing the most influential modality and incorporating geospatial conditioning—is intentionally designed for efficiency and informed prediction.

One key diagnostic observation from the final model evaluation is the smoother appearance of the predictions compared to the annotations. While this might initially seem

like a limitation, it is a favorable characteristic for a dense-prediction task at 10-meter resolution. Annotations often exhibit high-frequency noise and sharp, sometimes inconsistent, boundaries. The model's smoothing behavior acts as a regularizer, assigning a more generalized and spatially coherent score, which is more desirable for the task at hand.

The model demonstrably struggles to accurately predict values at the extreme ranges, particularly in the underrepresented annotation value ranges. This is a common challenge in regression tasks with imbalanced data and can stem from two main sources: data scarcity at the extremes, or inconsistent/ambiguous annotations in those rare regions, which prevent the model from forming a confident, non-conservative mapping function. Despite these limitations, a qualitative comparison to existing traditional human influence maps suggests that the NI predictions are in decent agreement.

6.6.4 Out-of-Distribution Detection

From a technical standpoint, the method relies on clustering extracted features to assign surrogate labels. A necessary operational consideration is that the clustering would typically require periodic updates when the system is deployed in an environment with rapidly evolving data streams. This periodic recalibration ensures the ID boundary remains relevant to the evolving data characteristics. However, for the scope of this thesis, since the target inference imagery is acquired once and its distribution is static, the clustering boundaries are not required to be updated post-deployment.

A potential technical limitation arises in edge-case scenarios where the incoming WILD data is almost entirely in-distribution. Unlike density-based clustering methods that naturally handle outliers, the current distance-based K-Means approach may not reliably detect very subtle shifts or assign meaningful OOD labels if the feature space is dense with ID samples.

The fundamental purpose of the OOD tool is to quantify covariate shift and, in turn, provide an inference-time reliability check. The tool operates under the assumption that if the input imagery is of a similar domain to the training set, one major source of prediction error—distribution shift—can be flagged as low risk. The tool provides a quantification of this similarity: "The inference imagery is of a similar domain". Conversely, if the tool raises a probability indicating a significantly different distribution, the implication is a crucial warning: "Given the significant distribution shift, the prediction's reliability may be compromised, as the input is far from what the model has been trained on". In essence, the tool transforms a hidden risk (distribution shift) into an explicit, quantifiable metric, serving as a critical check before trusting model predictions, especially in real-world, large scale mapping efforts.

Chapter 7

Naturalness Map of Europe

This chapter details the continental-scale deployment of the final model, as configured in Chapter 6. The model is encapsulated within the Distribution Shift Detector (Section 6.5), enabling the system to concurrently yield both the primary naturalness prediction and associated distribution shift scores. These scores facilitate the generation of the Reliability Map, which is presented alongside the primary output. Furthermore, the large scale prediction is evaluated by benchmarking the results against established traditional human influence maps.

7.1 Large Scale Deployment: Naturalness Map

For large scale inference, the model is deployed on a Sentinel-2 mosaic of continental Europe and its surroundings, produced for the year 2024. The spatial bounds of the mosaic span 47.8°W to 81.6°E longitude and 45.0°S to 84.4°N latitude. This extent encompasses a substantial portion of continental Europe, the majority of Anatolia, and the northwest of North Africa. The mosaic product, acquired from *EOX*[1], utilizes Level-2A data, providing Bottom of the Atmosphere (BOA) reflectance values across the red, green, blue, and near-infrared bands. The data is provided in GeoTIFF format with a `uint16` bit depth, projected in the WGS 84 geographic coordinate system (EPSG: 4326). The product comprises $328,000$ individual tiles, each with dimensions of $1024 \times 1024 \times 4$ pixels. Because the final model architecture (detailed in Chapter 6) uses a feature extractor initialized with weights pre-trained on 3-channel imagery, it requires an RGB input. Consequently, the near-infrared band provided by the mosaic is discarded prior to inference. The total data volume is approximately 1.6 TB.

The full Sentinel-2 mosaic is composed of hundreds of thousands of individual GeoTIFF tiles. To manage this collection as a continuous map without data replication, a Virtual Raster Tile (VRT) is utilized. The VRT serves as a metadata structure containing geographic information and pointers to underlying source files, allowing the processing

[1] `https://eox.at/`

pipeline to read only necessary data windows on demand. Input patches are extracted at a size of 512×512 pixels with a stride calculated to maintain an overlap of 128 pixels. This overlap is critical for the final blending step to prevent abrupt seams between adjacent predictions. Prior to model input, raw `uint16` reflectance values are normalized by a factor of 10000.0 to scale the data into the expected floating-point range. Inference is executed on a GPU in batches of 64. To ensure memory stability given the large spatial extent, predictions and associated weights are accumulated into large disk-backed memory map (memmap) arrays rather than system RAM. This optimized workflow handles the high data volume efficiently; the final map, with dimensions of $703,125 \times 478,125$ pixels, is processed in approximately 670 minutes.

Two layers of preprocessing filters are applied to maximize computational efficiency. First, a tile-level water skip filter is applied offline to the file list. Initially, 328,000 tiles are inspected; after filtering for overlap with the land area of interest, the processing queue is reduced to 160,173 tiles. Second, a runtime patch-level filter is implemented. A land mask, rasterized from administrative bounds to match the mosaic resolution, is queried during deployment. If a patch is found to be 100% water, it is skipped. Additionally, any patch disjoint from the study area bounding box is removed from the batch prior to inference.

Output patches passing these filters are blended using a weighted accumulation method on the disk-backed memory map. A Gaussian feathering mask (M_{Gauss}) is applied to each prediction to ensure seamless transitions. This mask, generated as a 2D Gaussian function centered on the patch, ensures prediction influence decays smoothly toward the edges. The accumulation of predictions (A) and weights (W) is calculated across all overlapping patches P at location (x, y), where P_{pred} is the predicted value and $\text{Mask}_{\text{valid}}$ denotes valid (non-NaN) pixels:

$$\begin{aligned} A[x, y] &= \sum_{P} P_{\text{pred}} \cdot M_{\text{Gauss}} \cdot \text{Mask}_{\text{valid}} \\ W[x, y] &= \sum_{P} M_{\text{Gauss}} \cdot \text{Mask}_{\text{valid}} \end{aligned} \tag{7.1}$$

The accumulation is weighted by the Gaussian mask, prioritizing predictions from the patch center while suppressing edge artifacts. The final mosaic prediction, M_{final}, is computed by dividing the accumulated predictions by the accumulated weights:

$$M_{\text{final}}[x, y] = \frac{A[x, y]}{W[x, y]}, \quad \text{where } W[x, y] \neq 0 \tag{7.2}$$

The final continuous output is scaled by a factor of 200.0 and exported as a Cloud Optimized GeoTIFF (COG) in `uint8` format, using 255 as the NoData value. The COG format includes internal overviews (pyramids) at levels 2^i for $i = 1$ to 10. This structure allows Geographic Information Systems (GIS) to render the $\sim 30\,\text{GB}$ map remotely via

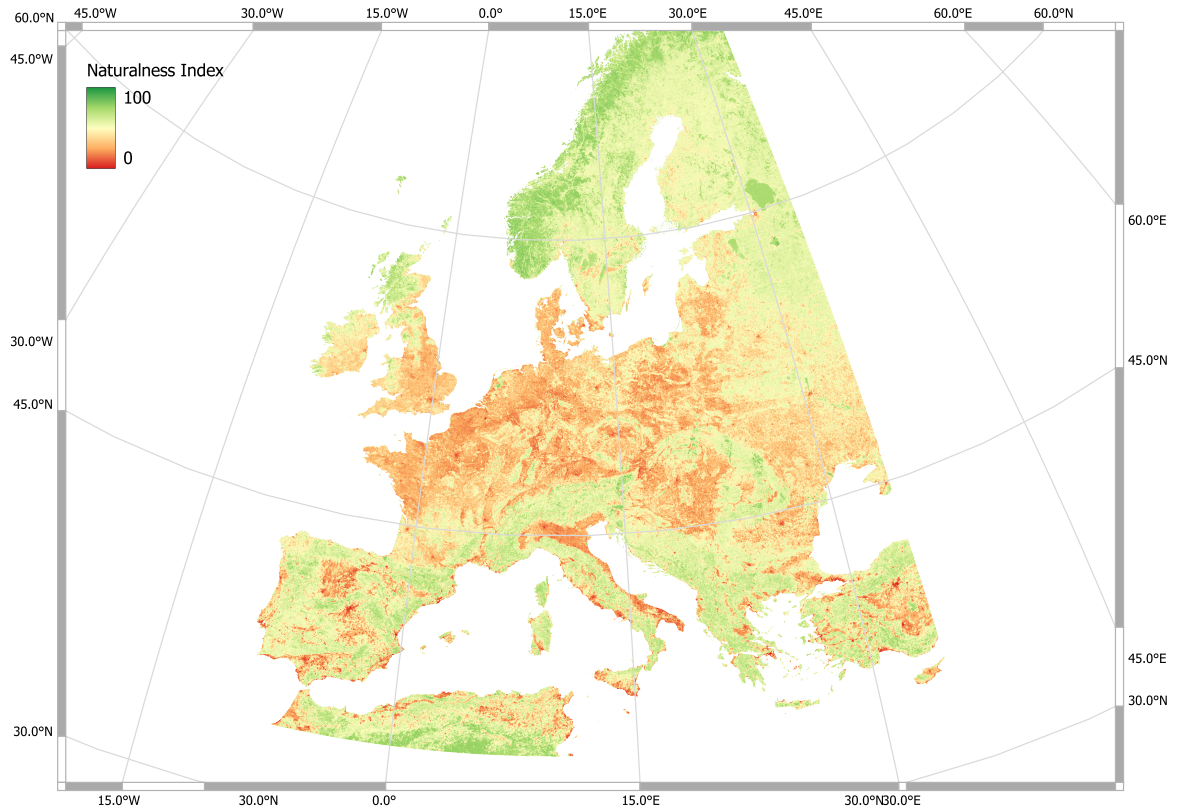


Figure 7.1: The resulting Naturalness Map of Europe from 2024, produced from Sentinel-2. This map provides a 10 m resolution view of naturalness across the continent.

HTTP without downloading the full file. The resulting Naturalness Map of Europe is presented in Figure 7.1.

Visually, the resulting DL-NI map immediately highlights key continental patterns: areas corresponding to the densely forested Scandinavian Peninsula, the Alps, and high mountain ranges show high naturalness scores, while the intensive agricultural zones of Central Europe and major urban corridors exhibit correspondingly low scores. Crucially, given the task definition of naturalness as the absence of human influence, regions like the forests of Northern Europe show scores of similar magnitude to remote, arid and semi-arid landscapes like the North African desert. This initial visual inspection confirms that the map accurately projects the model's core objective: assessing remoteness from anthropogenic pressure, irrespective of land cover type. This large scale output provides the basis for the quantitative and qualitative benchmarking detailed in the following section.

7.2 Comparison to Existing Human Influence Maps

This section evaluates the proposed Deep Learning Naturalness Index (DL-NI) by benchmarking it against three key reference products: the second-generation Human Foot-

print Index (HFP) [35] (Figure 7.2), the Global Human Modification map (gHM) [3] (Figure 7.3), and the Deep Learning-based Human Footprint Index (DL-HFI) [41] (Figure 7.4).

The evaluation presents both quantitative and qualitative methods. Quantitatively, density plots and statistical correlation metrics assess the alignment of the DL-NI with the three reference maps. Predictive performance is further evaluated by analyzing score distributions across distinct land-use types, specifically comparing urban environments (high anthropogenic pressure) against protected areas classified under the WDPA. Qualitatively, the maps are analyzed across multiple spatial levels, spanning from the continental perspective ($> 150\,\text{km}$) down to regional, city-level, and local details ($< 500\,\text{m}$), to thoroughly assess the impact of the 10 m resolution across diverse regions.

To ensure a robust comparative analysis, a unified valid-data mask is applied, restricting the evaluation exclusively to pixels with valid observations across all considered products. Given the heterogeneous numerical scales of the reference indices, a standardization step is required. However, applying a standard empirical min-max normalization—based on the observed sample extrema within the European spatial extent—would impose an artificial affine transformation that stretches the data to strictly occupy the $[0, 1]$ interval. This would inadvertently obfuscate the systematic biases, distributional skewness, and inherent calibration intrinsic to each product. To circumvent this, a theoretical normalization strategy is adopted. Each map is scaled relative to its mathematically predefined global limits (e.g., defined bounds of $[0, 100]$ for the DL-NI and $[0, 5000]$ for the HFP). By dividing by the theoretical maximum, this approach linearly projects the scores into a shared unified space of $[0, 1]$ while strictly preserving the relative magnitudes, internal variance, and structural biases of the original spatial distributions. Furthermore, because the HFP, gHM, and DL-HFI strictly quantify human modification, whereas the DL-NI quantifies naturalness (defined herein as the absence of such pressure), the normalized reference scores ($I_{\text{Ref, Norm}}$) require inversion. To ensure a monotonically consistent interpretation across all indices—where values approaching 1.0 denote high naturalness—the reference maps are inverted according to the following equation:

$$I_{\text{Ref, Inv}} = 1 - I_{\text{Ref, Norm}} \tag{7.3}$$

To analyze the statistical relationship, density plots are generated. Since the DL-NI is derived from an 8-bit integer scale, it possesses a practical precision of 0.01. To prevent this discretization from penalizing linear correlation metrics, the continuous indices of the inverted reference maps are binned to match this precision. A random sampling approach is adopted to ensure representative coverage, using a budget of 1 million points, which results in $n = 834,456$ valid coordinate points across the study area. No spatial resampling is performed during extraction. The statistical analysis reveals a strong alignment between the DL-NI and established maps. The strongest correlation is observed

with the gHM, yielding a Pearson correlation coefficient (r) of 0.85, an R^2 of 0.72, and a low RMSE of 0.10. Comparison with the HFP shows a similarly strong relationship ($r = 0.82$, $R^2 = 0.67$, RMSE = 0.15). The DL-HFI also demonstrates high consistency with the DL-NI ($r = 0.81$, $R^2 = 0.66$, RMSE = 0.12). These density plots and metrics are illustrated in Figures 7.5, 7.6, and 7.7.

To further benchmark against the gHM, HFP, and DL-HFI, the study analyzes mean normalized scores for regions defined by the WDPA and major urban centers. The analysis encompasses 32 countries and queries WDPA features in December 2025, yielding 2,340 distinct areas: 585 IUCN Category II (National Parks), 256 Category Ib (Wilderness Areas), 622 Category Ia (Strict Nature Reserves), and 877 Category III (Natural Monuments), totaling over $475,000\,\mathrm{km}^2$. To represent high anthropogenic pressure, administrative boundaries for 56 major cities (e.g., Tunis, Istanbul, London, Berlin) are acquired via OSMnx [144]. These geometries are used to crop the normalized rasters, calculating the mean pixel value per feature. A distinct and consistent separation between the high naturalness scores of IUCN categories and the low scores of built-up areas is evident across all four maps, as shown in Figure 7.8. Additionally, the figure includes an "Overall Average" column, which illustrates the aggregated distribution of mean scores across all evaluated geographic features, providing a macro-level summary of the analysis.

The comparison of mean normalized scores across these regions provides a quantitative validation of the DL-NI's performance, as shown in Figure 7.8. The analysis confirms a clear and expected separation between high-naturalness and high-pressure zones across all four indices. For all three strict conservation categories (Ia, Ib, and II), the DL-NI reports high mean naturalness scores (median values ranging from approximately 0.65 to 0.80), placing it statistically close to the HFP and gHM indices. Conversely, built-up areas exhibit significantly lower scores (median values below 0.35), confirming that the model accurately captures the inverse relationship between human pressure and naturalness. The DL-NI generally exhibits lower score variance (tighter interquartile ranges) than the traditional rule-based indices (HFP and gHM) within the IUCN categories. This suggests that the deep learning model, having been trained on global heterogeneity, provides a more consistent and less noisy estimate of naturalness across geographically diverse protected areas. Furthermore, the DL-NI shows slightly lower central tendencies (median scores) than the HFP in IUCN categories Ia and II. This outcome is consistent with the expectation that the DL-NI's 10 m resolution captures fine-scale infrastructure and fragmentation within protected area boundaries that the original 1 km resolution HFP may overlook, thus yielding a marginally more pessimistic, yet potentially more accurate, assessment of true intactness.

The DL-HFI's significantly coarser 1 km resolution (compared to the DL-NI's 10 m) indicates a fundamental limitation of coarse-scale aggregation in urban environments. The 1 km pixels blur the distinction between areas of high human influence and areas of

low human influence, resulting in a spatially diluted score where urban pixels are inaccurately classified with relatively high naturalness (near 0.4). This behavior, evident at both the high- and low-pressure extremes, confirms that the DL-NI's high spatial resolution successfully eliminates the spatial blurring inherent in coarser-resolution indices, providing a more reliable and precise assessment of anthropogenic pressure.

The continental-scale DL-NI is visually benchmarked against its counterparts across diverse locations and spatial levels. At the continental scale, the DL-NI exhibits high agreement in the density and general distribution of naturalness values with the gHM, HFP, and DL-HFI. This consistency holds at national and regional levels, as seen in comparisons for Austria and Bavaria (Figures 7.9–7.13).

The advantages of the DL-NI are most apparent at finer scales, as detailed through a comprehensive series of visual evaluations in Figures 7.14–7.22. Zooming into these city-level and sub-regional views reveals the true utility of the 10 m resolution: the ability to resolve fine-scale anthropogenic fragmentation that is otherwise lost to spatial aggregation. As visually evident in comparisons over Gabès and Métlaoui (Figures 7.17 and 7.18), traditional coarse-resolution indices render the landscape as homogeneous, highly pixelated blocks, effectively blurring the boundaries between distinct human and natural features. In contrast, the DL-NI captures fine-grained spatial structures, delineating exact urban edges, localized infrastructure, and transitional gradients of human pressure. This precision is particularly valuable in dense urban environments; in Munich and Vilnius (Figures 7.14 and 7.15), the DL-NI resolves intra-urban green spaces, river corridors, and suburban road networks, whereas the reference maps aggregate the metropolitan areas into monolithic high-pressure zones. The model also maps sharp topographical boundaries where development meets natural terrain, as seen in Tetouan (Figure 7.22). Furthermore, the DL-NI identifies isolated human activity: it cleanly outlines a specific quarry in Taghramt (Figure 7.21) that traditional maps fail to resolve. Similarly, near Negrine, Algeria (Figure 7.19), the HFP map incorrectly characterizes arable lands as high naturalness due to spatial dilution, while the DL-NI correctly reflects the localized anthropogenic pressure in these managed agricultural landscapes.

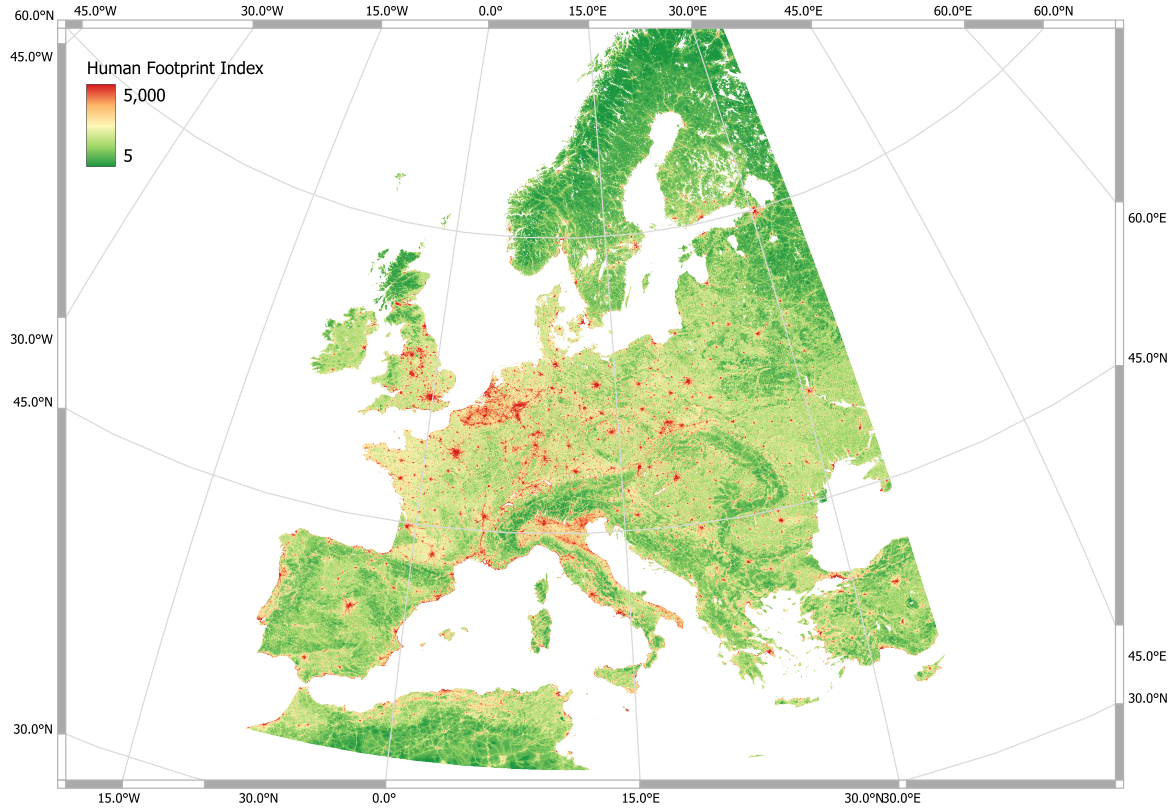


Figure 7.2: The Human Footprint Index (HFP) map from 2020, with 300 m native resolution.

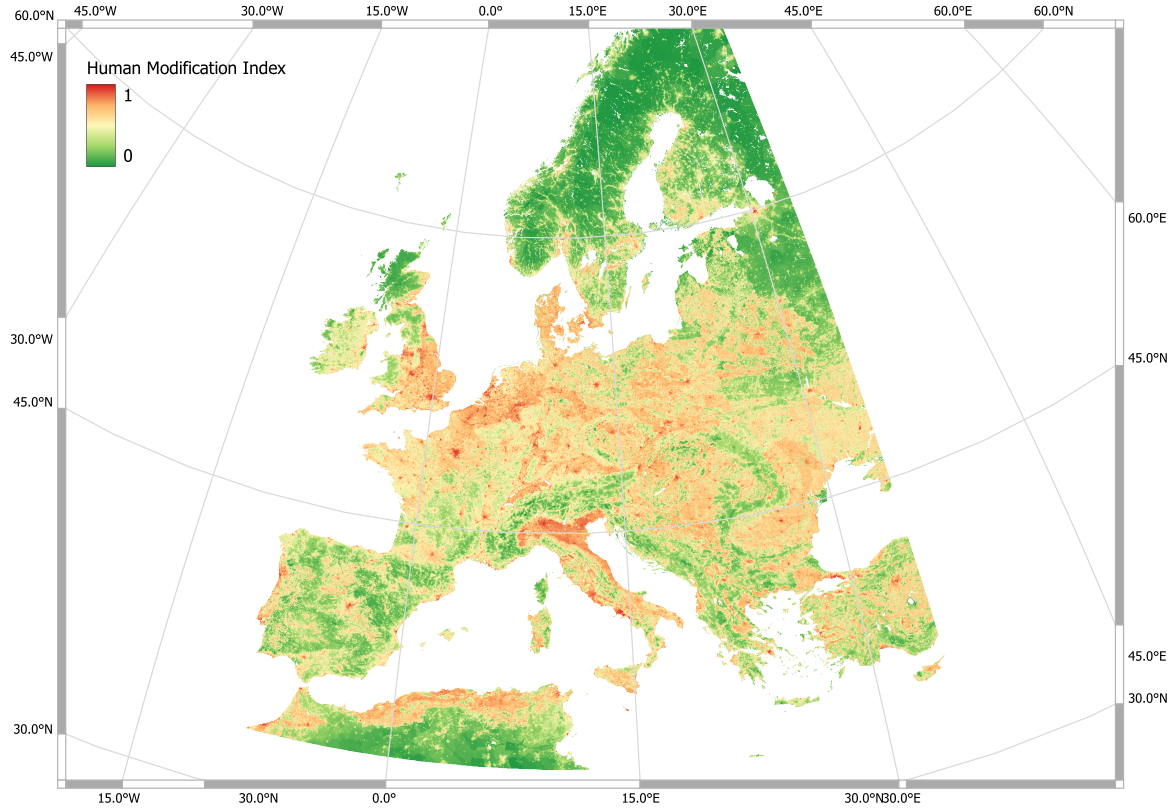


Figure 7.3: The Global Human Modification map (gHM) from 2016 with 1.1 km native resolution.

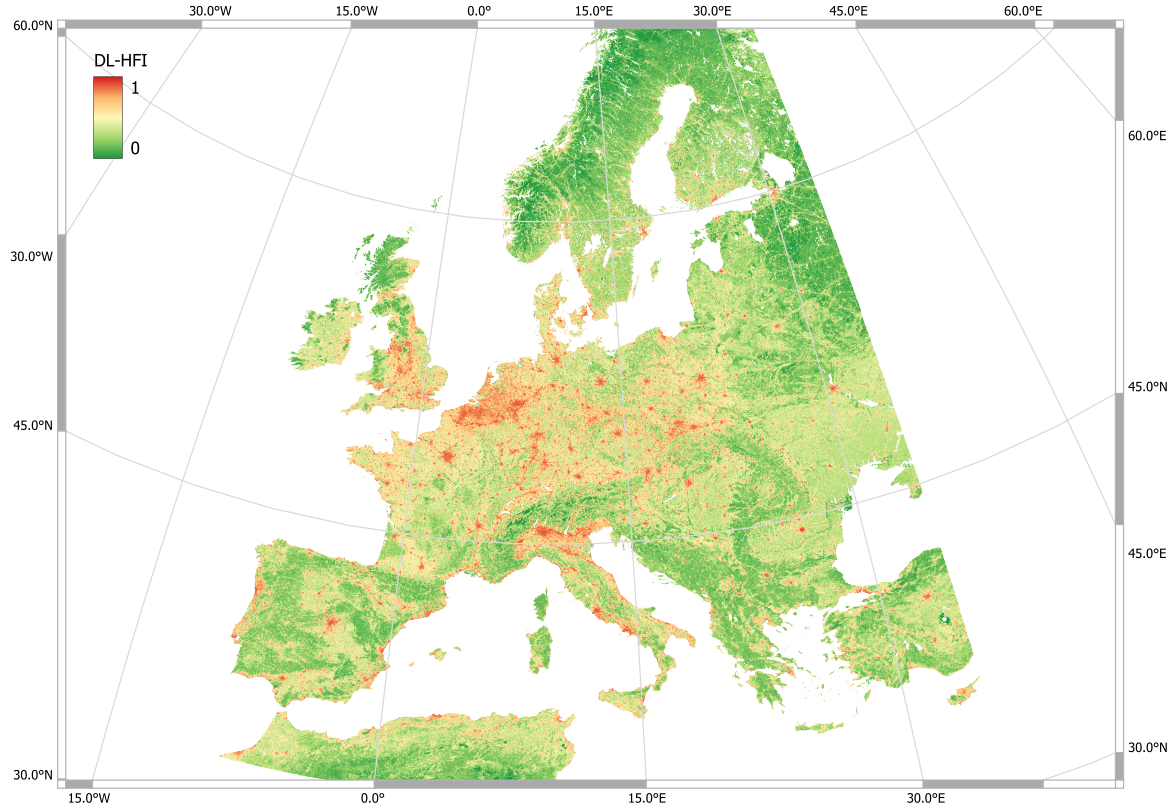


Figure 7.4: The Deep Learning-Human Footprint Index (DL-HFI) from 2019, with 1.1 km native resolution.

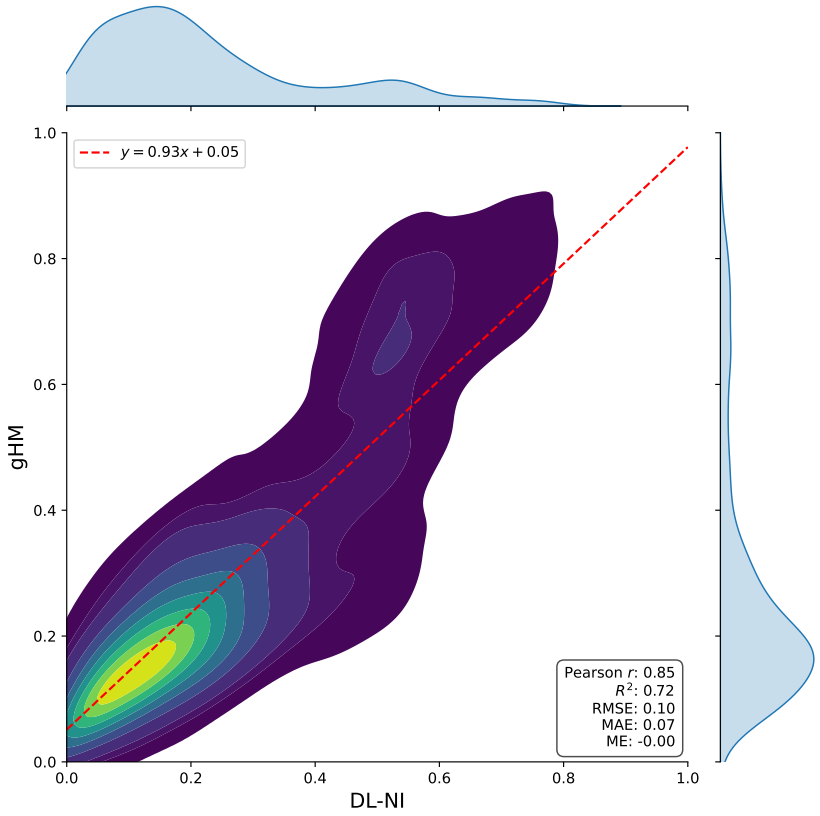


Figure 7.5: Aggregated pixel-level comparison between the DL-NI and Inverted gHM. $n = 834,456$ samples.

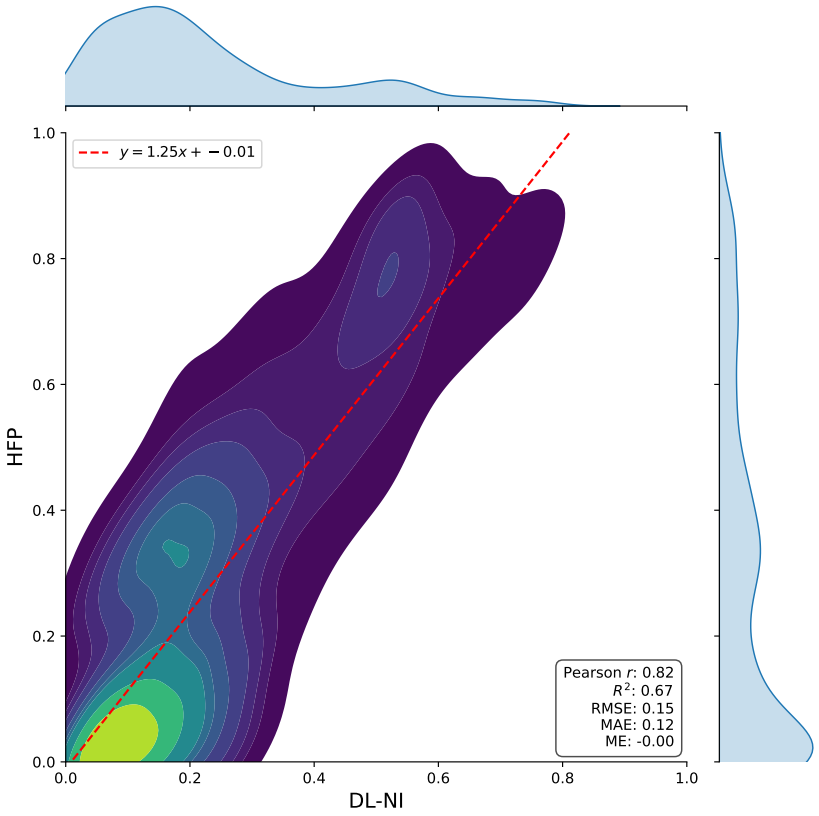


Figure 7.6: Aggregated pixel-level comparison between the DL-NI and Inverted HFP. $n = 834,456$ samples.

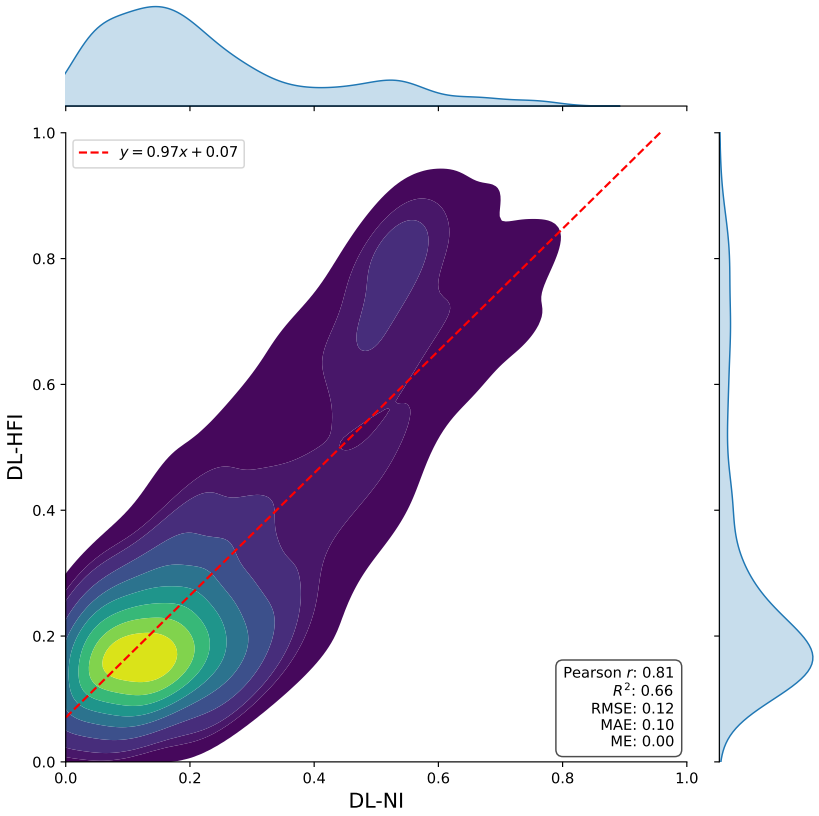


Figure 7.7: Aggregated pixel-level comparison between the DL-NI and Inverted DL-HFI. $n = 834,456$ samples.

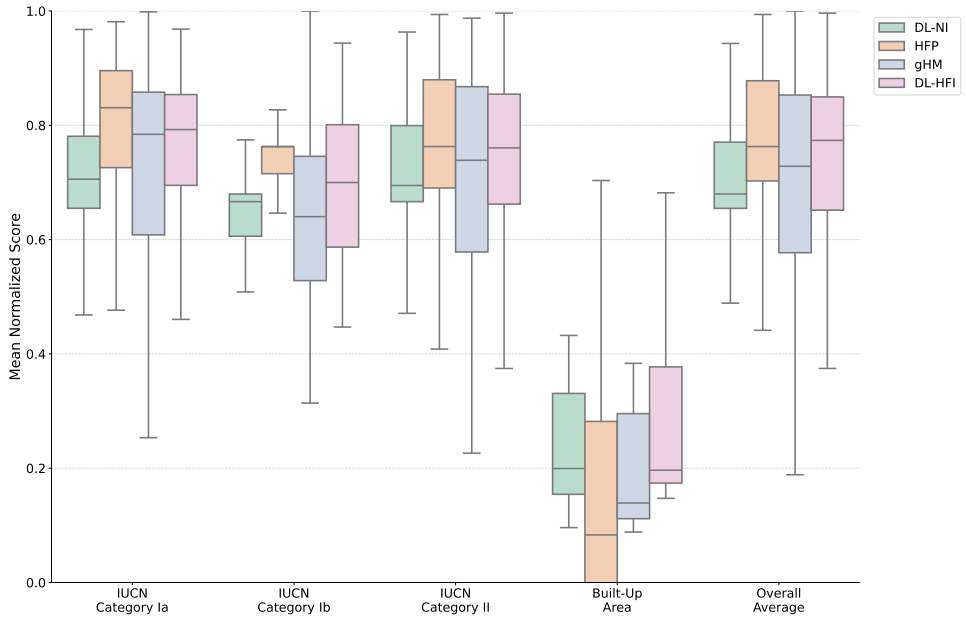


Figure 7.8: Comparison of the DL-NI with HFP, gHM, and DL-HFI across IUCN Protected Area categories (Ia, Ib, II) and built-up areas. The "Overall Average" column combines the scores from all evaluated categories.

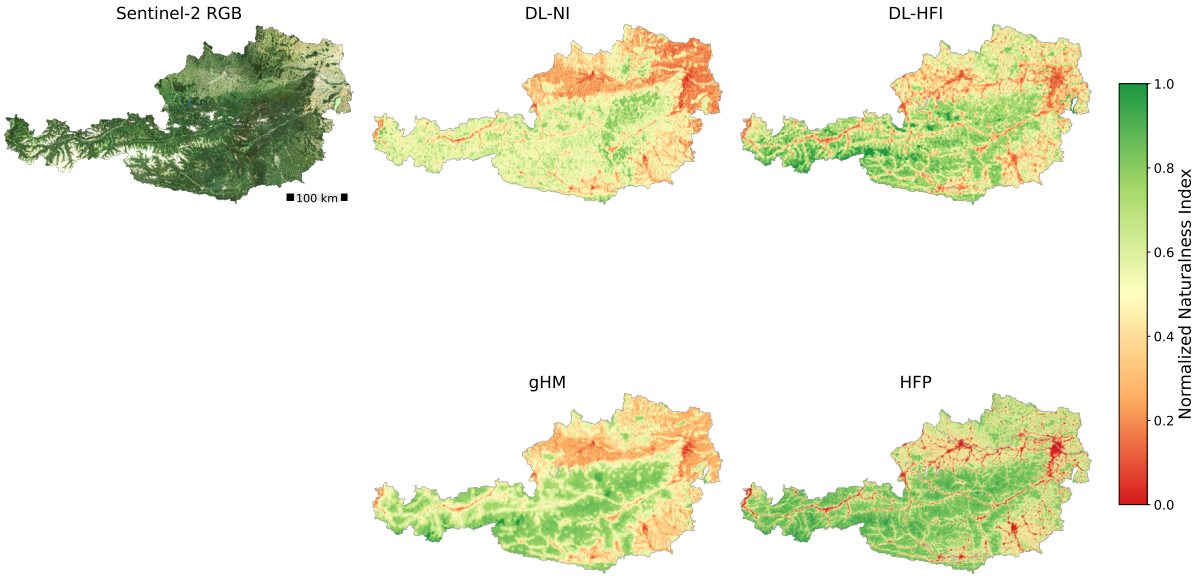


Figure 7.9: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Austria.

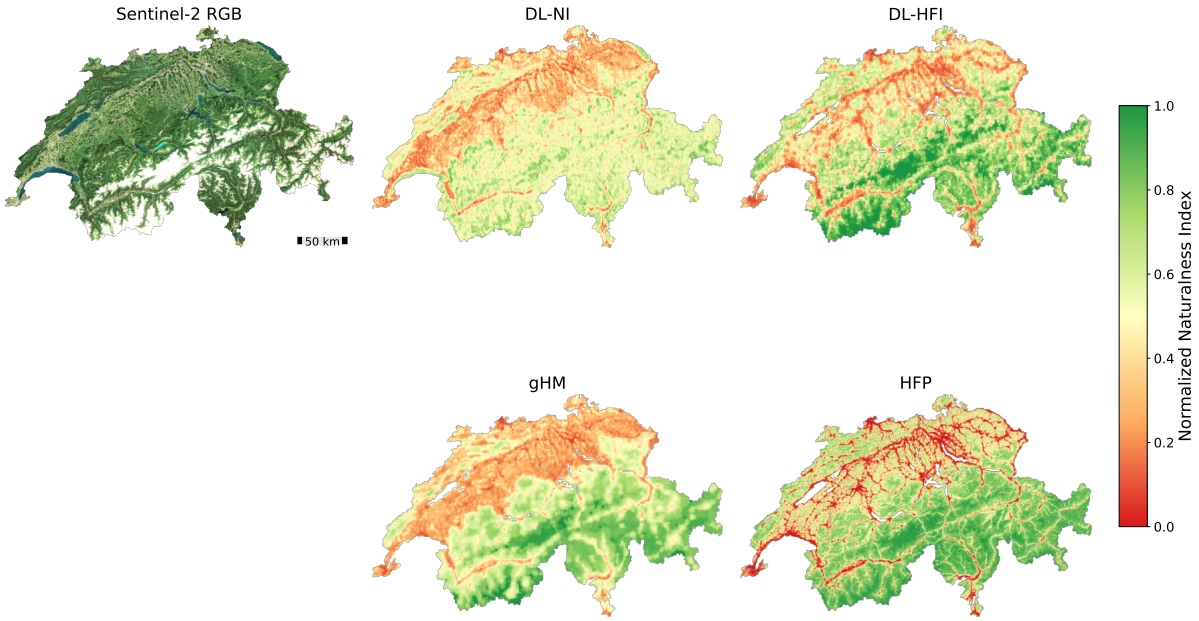


Figure 7.10: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Switzerland.

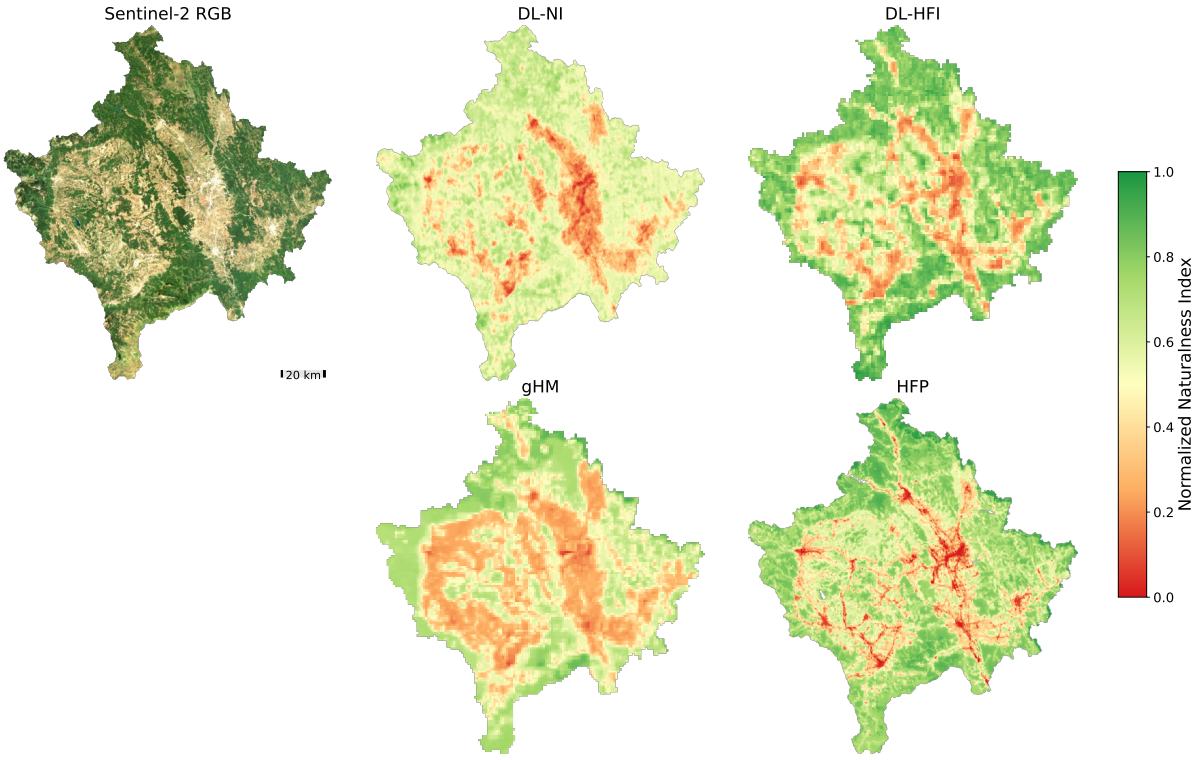


Figure 7.11: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Kosovo.

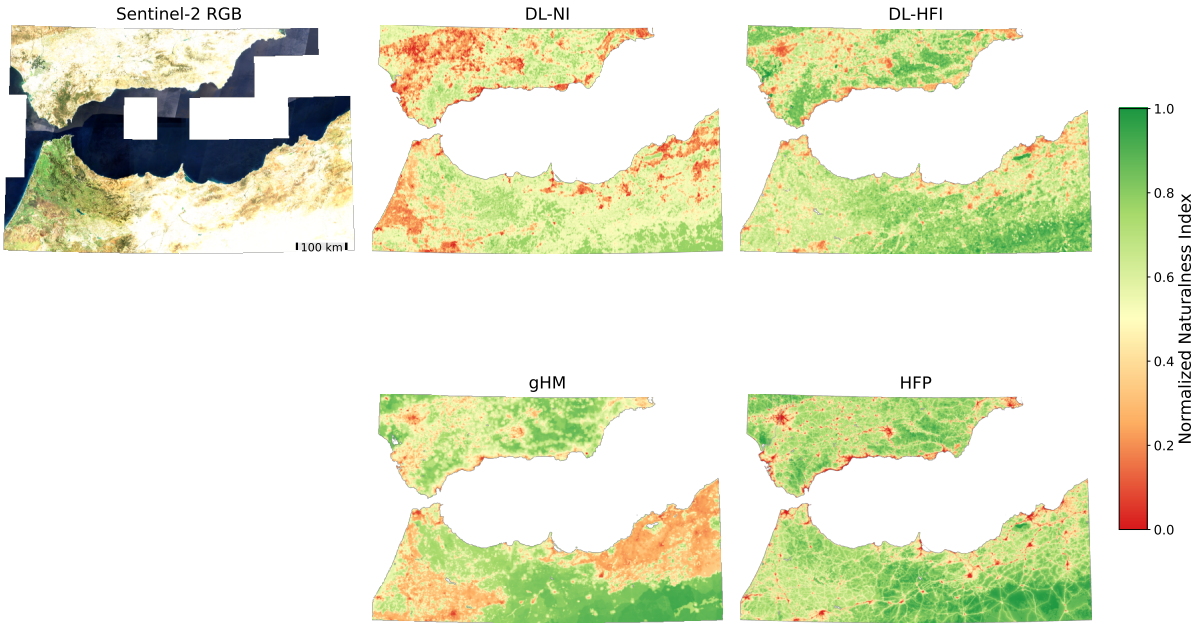


Figure 7.12: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Gibraltar.

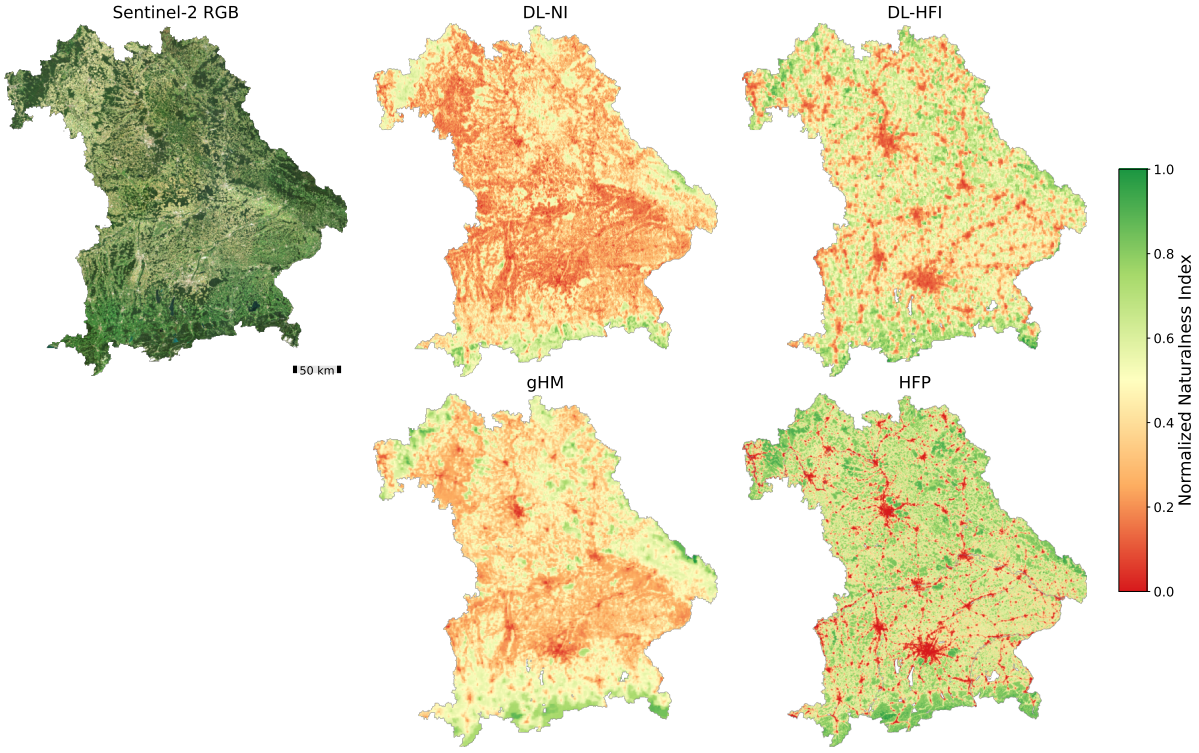


Figure 7.13: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Bavaria, Germany.

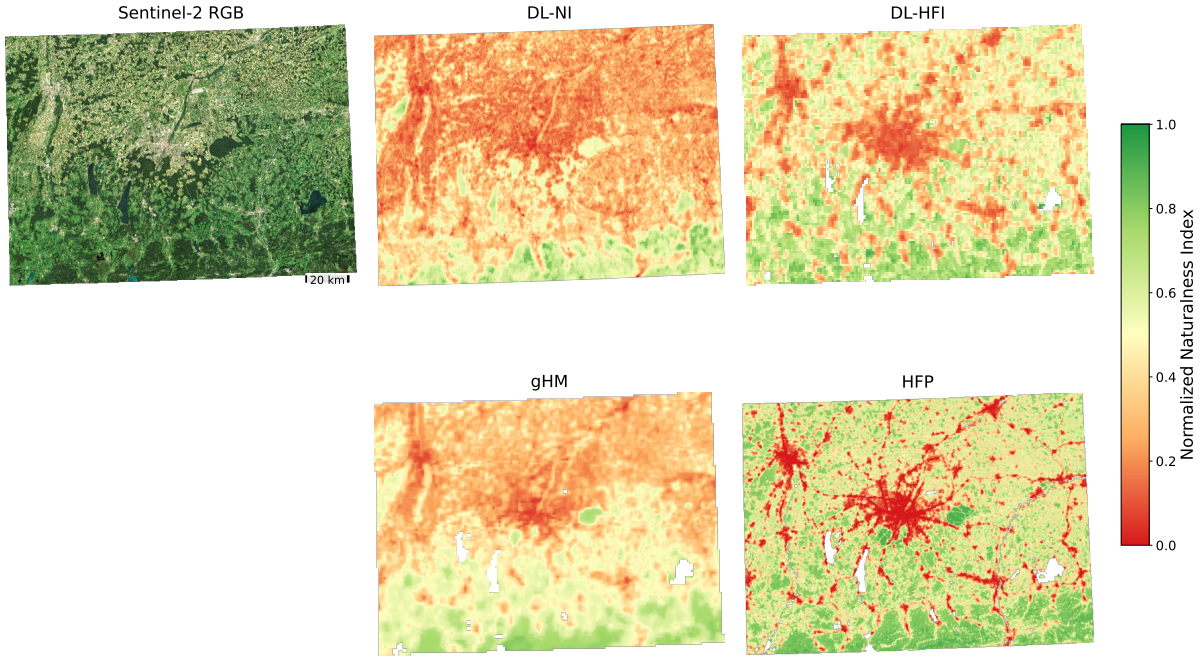


Figure 7.14: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Munich, Germany.

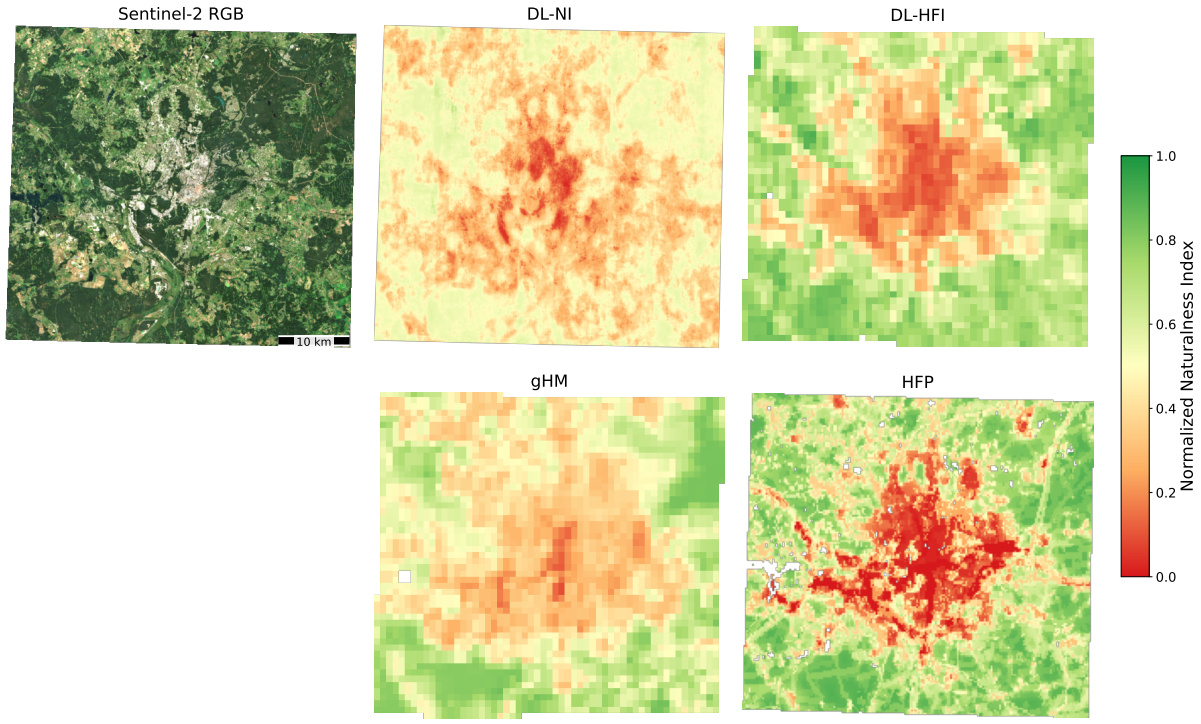


Figure 7.15: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Vilnius, Lithuania.

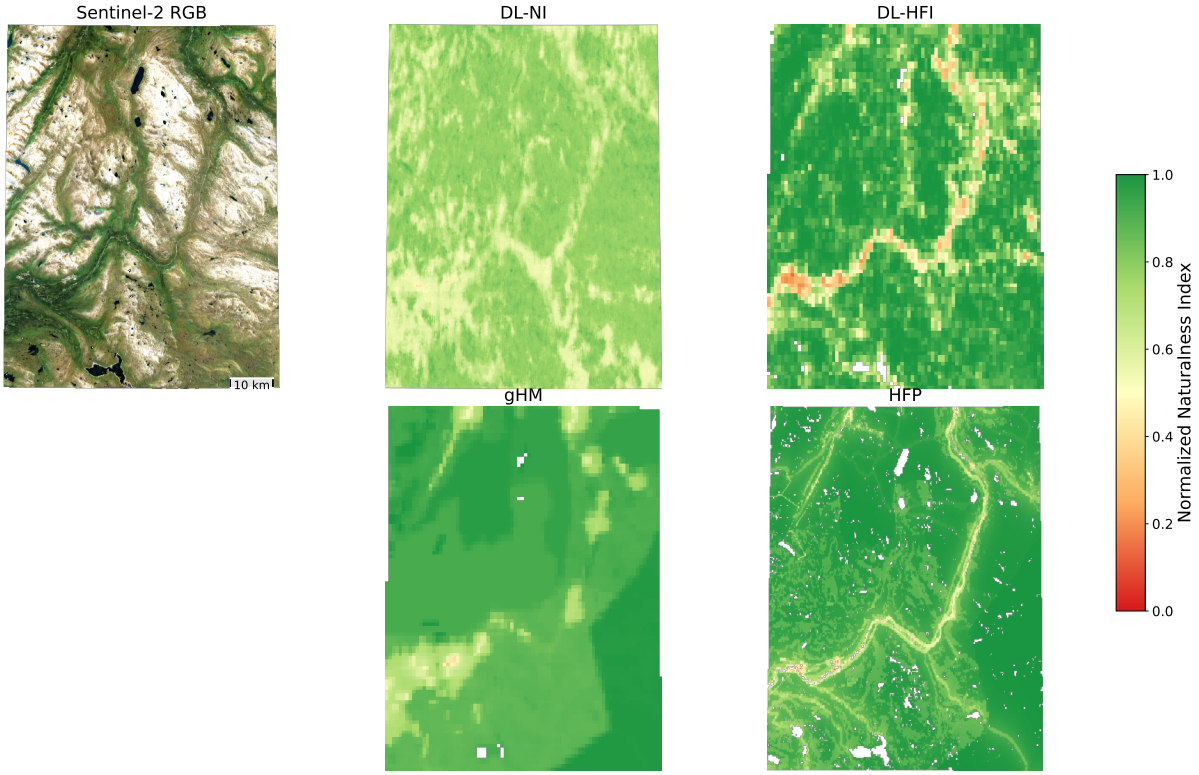


Figure 7.16: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Vindelfjällen Natural Reserve, Sweden.

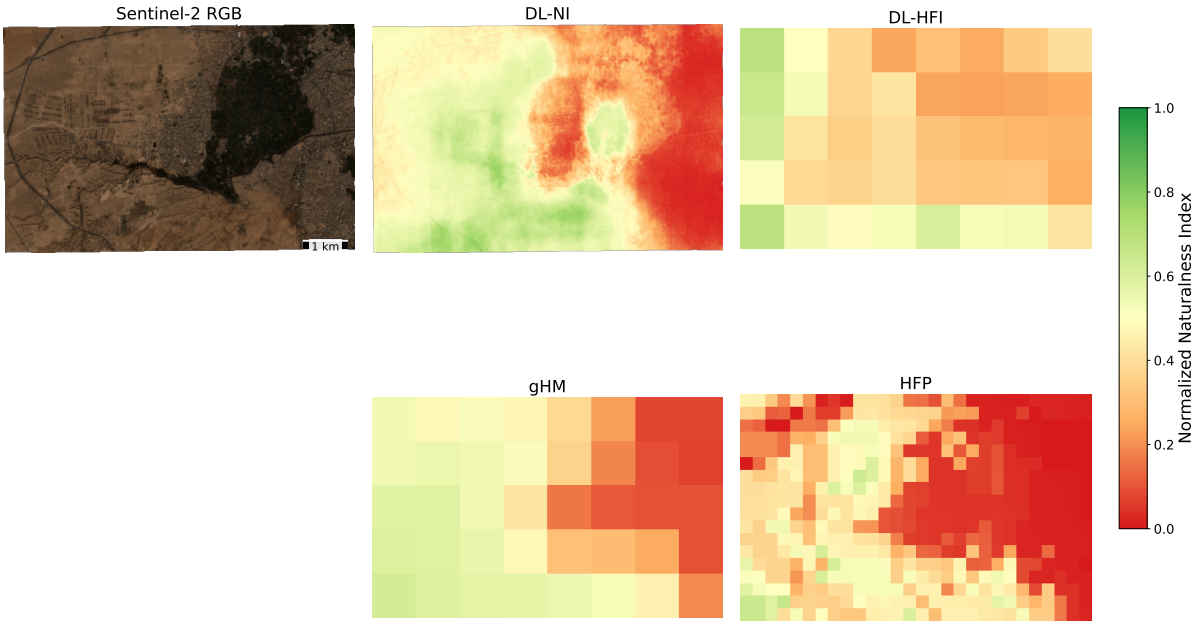


Figure 7.17: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Gabès, Tunisia.

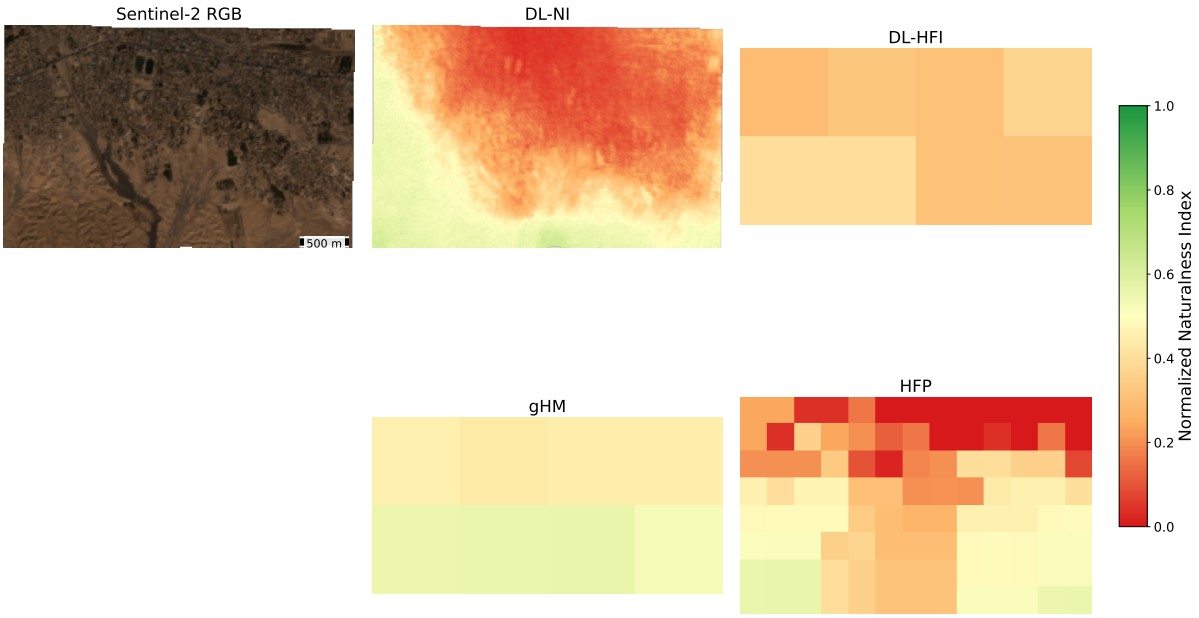


Figure 7.18: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Métlaoui, Tunisia.

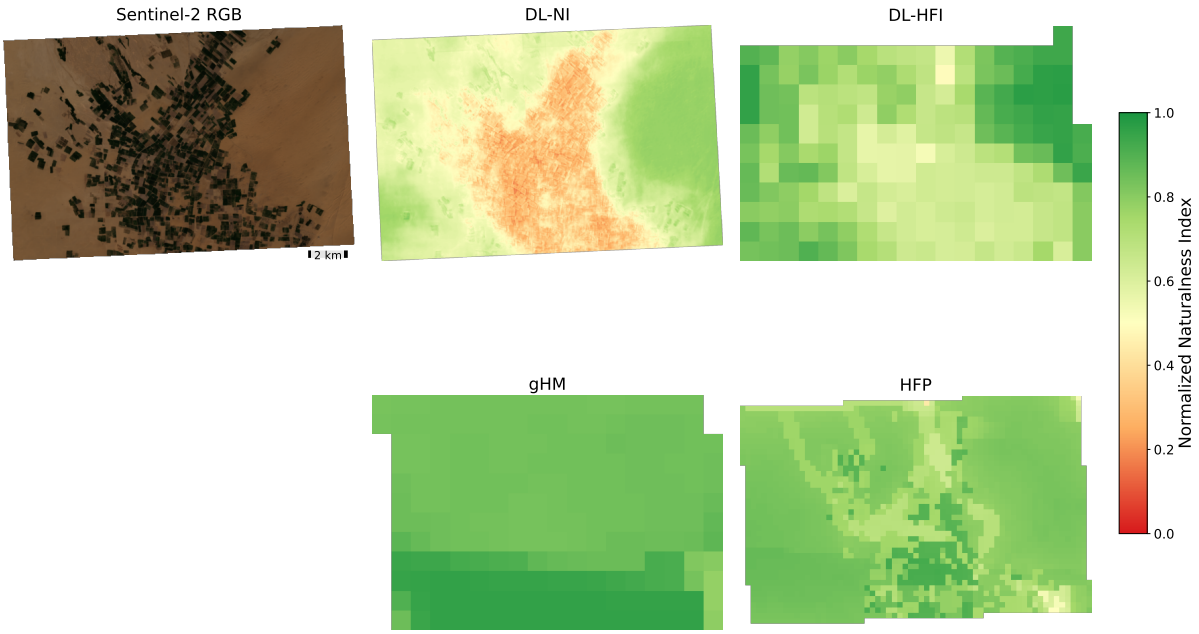


Figure 7.19: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Negrine, Algeria.

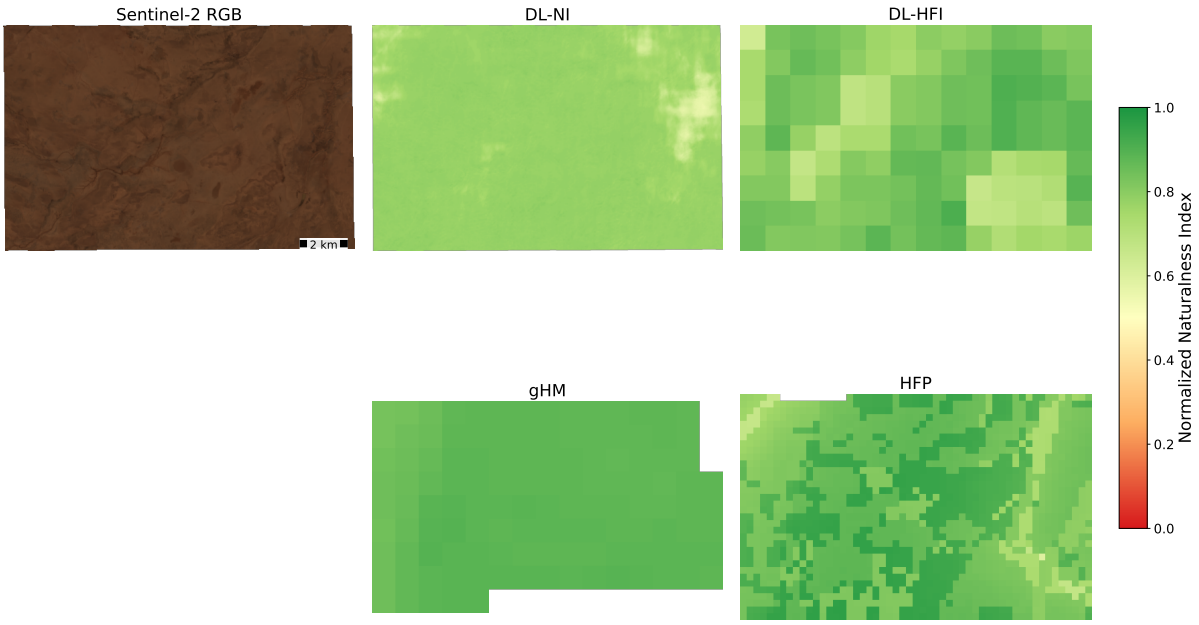


Figure 7.20: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over the Oriental Region, Morocco.

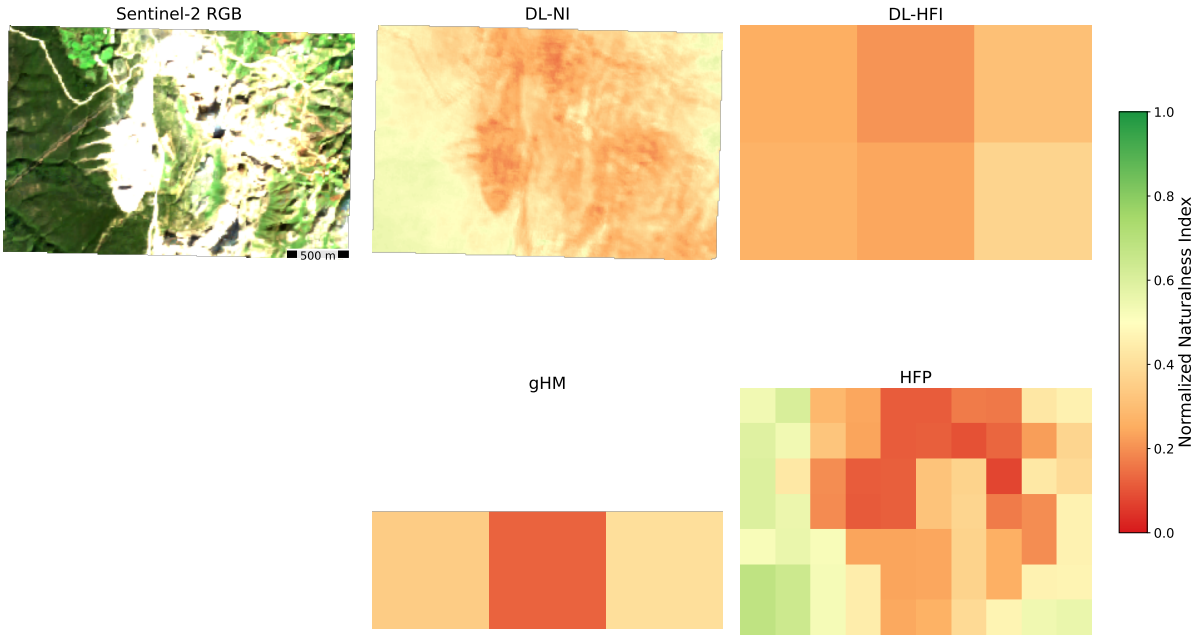


Figure 7.21: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over a quarry near Taghramt, Morocco.

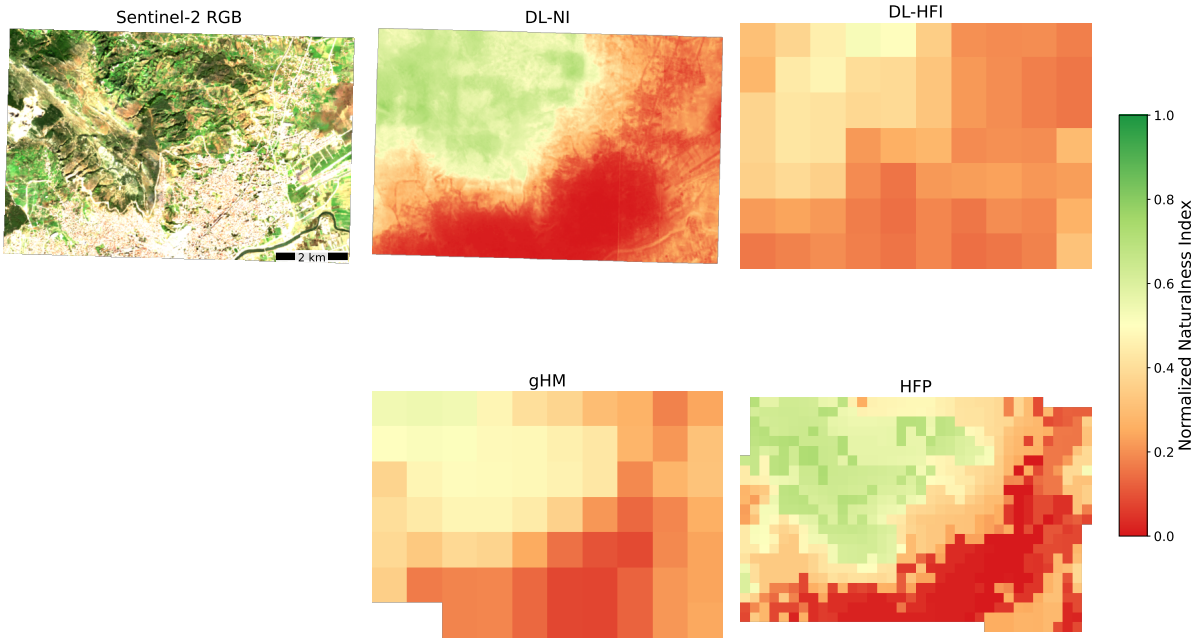


Figure 7.22: The input Sentinel-2 imagery, the DL-NI prediction map, DL-HFI, gHM, and HFP over Tetouan, Morocco.

7.3 Large Scale Deployment: Prediction Reliability Map

The distribution shift detection methodology detailed in Section 6.5 is applied to the final model during its continental-scale deployment. Consequently, for each input inference patch, the model concurrently generates a primary task prediction and a corresponding distribution shift score within the continuous interval $[0, 1]$. This metric, referred to as the g-score, quantifies the degree of covariate shift associated with the input data. The resulting spatial visualization is termed the *Prediction Reliability Map*, as it provides operational insights into the expected reliability of the model's predictions.

Specifically, the process is defined as follows: ID data is sourced directly from the final model's training set, which is utilized for continental-scale deployment as detailed in Section 7.1. The WILD data is prepared by randomly sampling Sentinel-2 patches from the deployment mosaic. Both ID and WILD patches possess dimensions of $1024 \times 1024 \times 3$ pixels. Following the procedure outlined in Section 5.3, the $1,000$ ID and $1,640$ WILD samples (representing 0.1% of the total mosaic) are processed by the frozen model, with internal activations extracted from the bottleneck layer `model.encoder.layer4`. These activations are spatially downsampled to yield one-dimensional feature vectors. The combined activation space of ID and WILD samples is clustered into 164 groups (calculated as $0.1\% \times \text{len(WILD)}$) using k-means. Proximity to known ID samples subsequently serves as a proxy for assigning surrogate labels to the WILD data. For each cluster, if the proportion of ID samples meets or exceeds a threshold of 0.1, all constituent samples are assigned the surrogate label 0 (ID); otherwise, they are assigned the label 1 (OOD). This yields a labeled training set used to fit a binary classifier that maps feature representations to their respective distribution labels. During deployment, activations are extracted from unseen samples to predict a distribution score ranging from 0 to 1, referred to as the Reliability Score.

The resulting spatial distribution of these scores is visualized in Figure 7.23, which displays randomly sampled Sentinel-2 input patches across the continental study area. A distinct pattern is observable regarding offshore maritime regions: patches located over the sea are predominantly categorized as "Low Reliability". This indicates that, regardless of the primary task prediction quality, these input patches exhibit a high covariate distribution shift relative to the training data, rendering the model's output less reliable in these zones.

Complementing the spatial analysis, quantitative results characterizing the global distribution of these scores are presented in Figure 7.24. As shown by the kernel density estimate, the distribution of g-scores is heavily skewed toward the low-reliability regime. Based on the statistical properties of this plot, specific thresholds were derived to define the "Low", "Medium", and "High" reliability categories visualized in Figure 7.23.

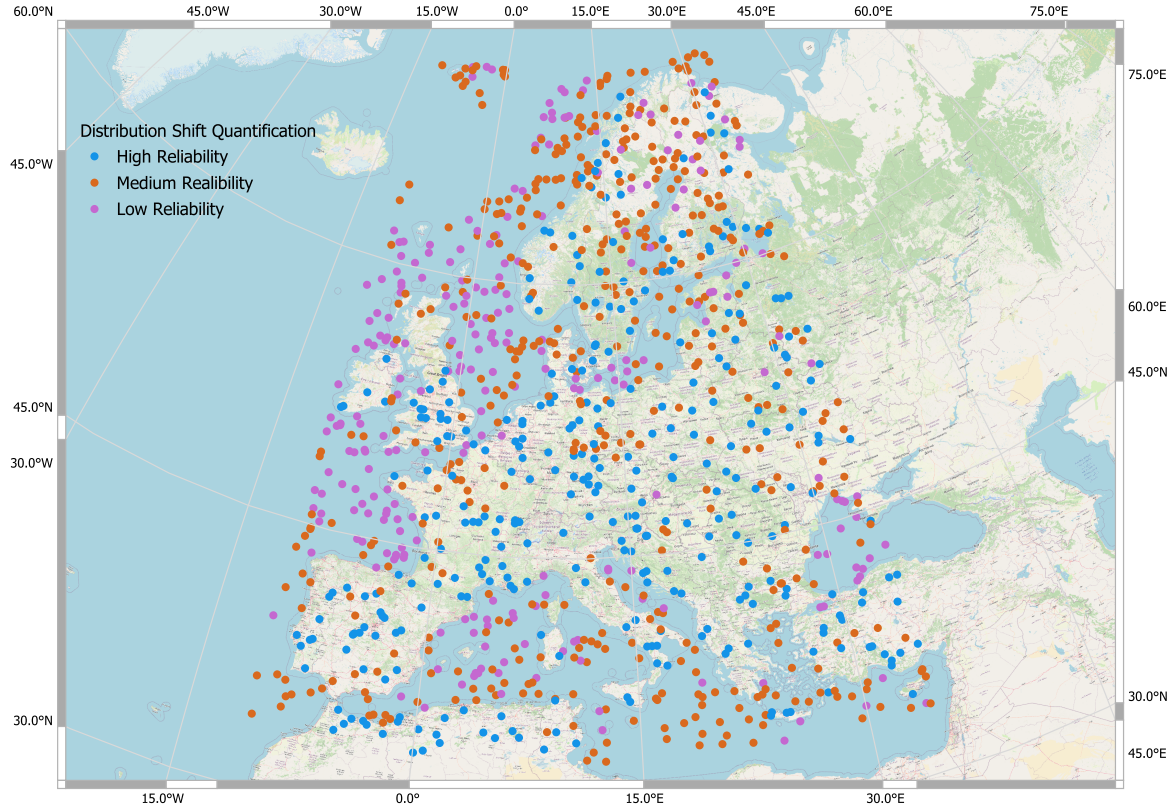


Figure 7.23: Continental-level Prediction Reliability Map. Inference tiles are categorized into distribution shift categories based on their TARDIS g-score. *Base map data copyrighted OpenStreetMap [4] contributors and available from* ***https://www.openstreetmap.org***.

It is important to note that because these results represent global statistics over the entire map, they do not necessarily quantify specific model performance or map accuracy. Rather, they serve as a diagnostic tool, enabling users to interpret the utility of the distribution shift detector alongside the main prediction output. Qualitative results demonstrating this variance are illustrated in Figure 7.25, which presents input Sentinel-2 imagery alongside their corresponding DL-NI predictions and reliability scores.

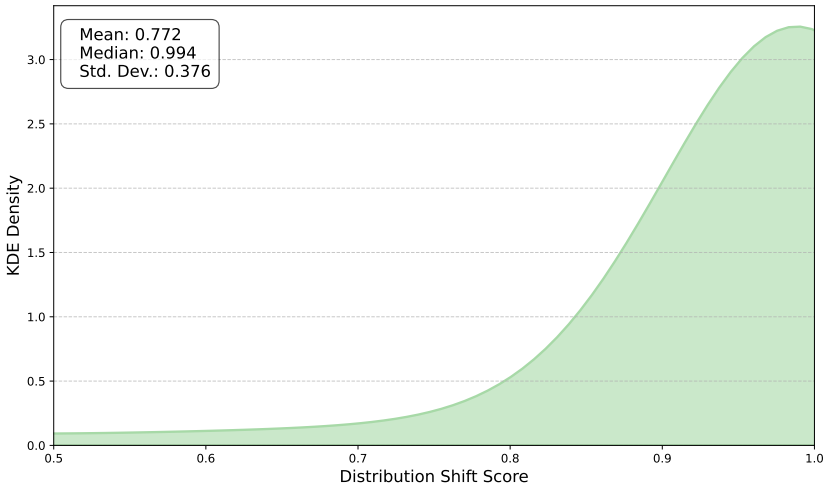


Figure 7.24: Distribution of OOD Shift Scores Across European Patches. This kernel density estimate shows the distribution of the TARDIS g-scores. The distribution is heavily skewed toward the high distribution shift regime.

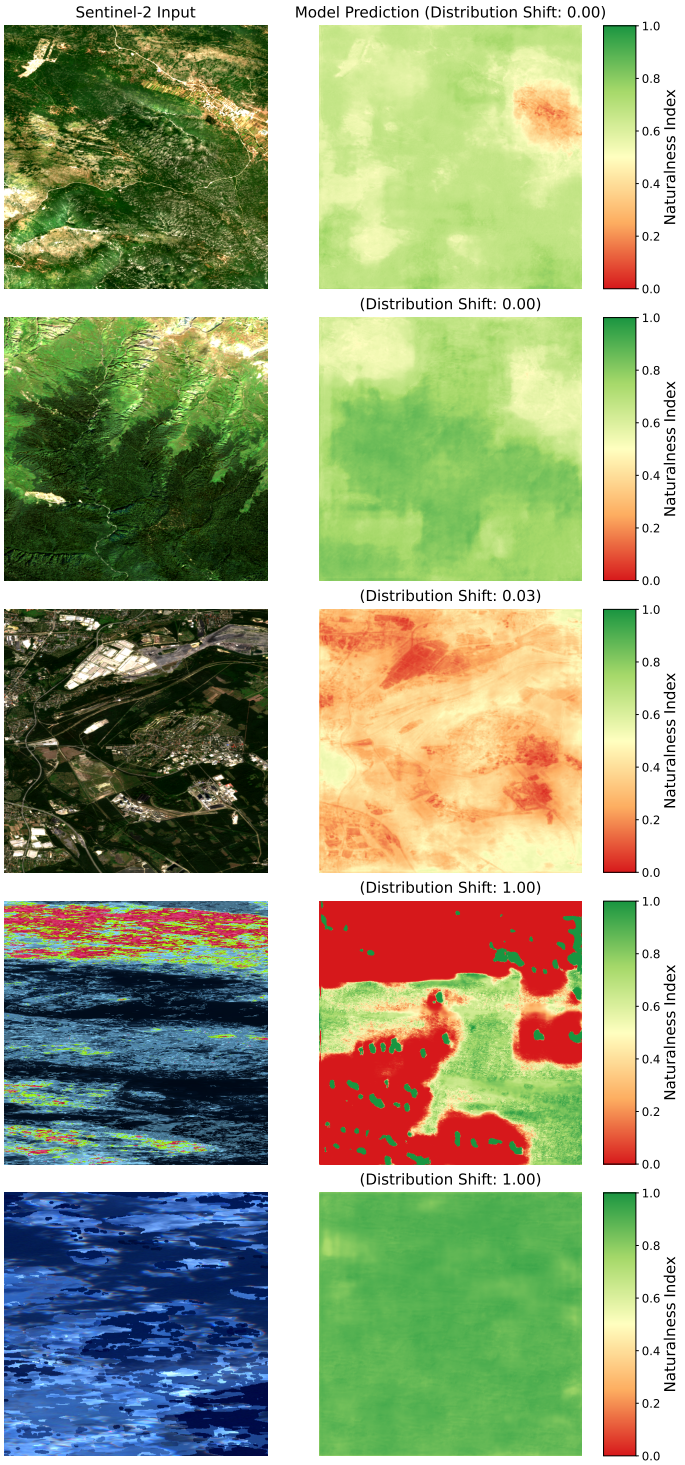


Figure 7.25: Model and reliability predictions for WILD samples. The figure illustrates five example model predictions across the mosaic. Predictions whose input images appear visually distinct or corrupted are assigned high distribution shift scores and low reliability. In contrast, predictions with inputs similar to the training set are assigned low shift scores and high reliability.

Chapter 8

Discussion

This chapter presents a critical review and synthesis of the thesis's work, structured around the fundamental limitations encountered in developing and deploying a data-driven approach to naturalness mapping.

Acknowledging the NI as a proxy of human influence. It is important to acknowledge that the Naturalness Index remains an approximation rather than an absolute measure of human influence on land. Human influence is a dynamic and multi-faceted concept, influenced by countless interacting factors. Incorporating every possible driver is impractical and limited by the challenge of defining naturalness in a spatially and temporally consistent way. Thought-provoking questions highlight this complexity: At what point does the conversion of a single tree into a forest reveal naturalness characteristics? How much should naturalness vary from day to day and season to season? How natural should the meadow in the middle of a football stadium be considered? Is a national park next to a bustling city less natural than one that is remote? These questions illustrate that mapping human influence is a task that lingers between being conceptually fluid and temporally variable, meaning the annotation is, at best, an approximation of something that is not rigidly defined.

How to read the maps? It is crucial to recognize that the Naturalness Index and similar human influence indices are proxies of anthropogenic pressure, not ecological impact per se. The index identifies areas where human influence is present or absent based on features observable in satellite imagery, such as infrastructure and land use change. The presence of high influence does not inevitably translate to harmful ecological outcomes, and conversely, a low influence score does not guarantee pristine ecological health. Importantly, the human influence indices treat the land surface as if it were a blank slate on which human influence is written, but this is not the case; the pre-existing ecological sensitivity and history of a location fundamentally shape the effect of any new human activity.

The ability to interpret patterns of human influence that are based on geographic features is further constrained by the complexities of human interactions with nature

and our limited understanding of them. For instance, human behaviors—such as the distance people travel from roads and rivers—vary significantly across different biomes, being less in the temperate zone than in the tropics. Similarly, per capita consumption in the developed world results in ecological impacts not just locally, but globally. Current modeling capabilities do not yet allow us to assess these complex, nuanced interactions globally in a spatially and temporally consistent manner. Achieving such a comprehensive, oracle-like assessment across both space and time represents the guiding principle for the future of human-nature modeling. As such, the index should not be interpreted to mean that the lowest possible NI score (highest naturalness) implies an untouched or maximally functional ecosystem, nor does the highest score (lowest naturalness) automatically dictate complete ecological ruin. Instead, the NI provides a quantitative surrogate for understanding the relative distribution of human pressure across the landscape.

There are many ways of measuring human influence, and geospatial data is only one of them. A core assumption of this thesis is that coupling the NI annotations with satellite observations forms a supervised learning task, assuming a mapping can be found between the annotations and the satellite imagery. This suggests that the satellite imagery of an area is expected to harbor distinct visual notions of naturalness. Although there is a certain level of quantitative disagreement with the existing traditional maps (HFP and gHM), the qualitative analysis supports that this is indeed a mapping that can be learned from the visual data.

On inter-map agreement and disagreement: scale. The high statistical agreement ($r > 0.80$) between the DL-NI and coarse-resolution benchmarks (gHM, HFP, DL-HFI) validates that the DL-NI successfully captures the same broad, continental-scale patterns of human influence defined by traditional methods. However, the true utility of the DL-NI lies in the spatial scale where disagreement emerges. The established indices suffer from spatial limitations, as the averaging effect of 1 km pixels hinders the distinction between areas of high human influence and areas of low human influence, leading to spatially diluted scores in urban environments. The DL-NI's 10 m resolution resolves this issue, providing a more constrained and accurate score in high-pressure areas. This confirms that while all maps agree globally, the DL-NI's higher spatial resolution successfully eliminates the spatial blurring inherent in coarser-resolution indices, providing a more reliable and precise assessment of anthropogenic pressure across the entire human influence continuum. This demonstrates the model's capacity to deliver reliable, unmixed signals on a continental scale, where each 10 m $\times$ 10 m pixel requires accurate measurement under diverse geographic conditions.

On inter-map agreement and disagreement: temporal disparity. The comparison between the DL-NI (a 2024 map) and its benchmarks (dated between 2016 and 2020) necessitates accounting for a significant temporal lag of several years. For a coarse-scale analysis, this disparity is often overlooked. However, when observing environmental

change at the DL-NI's high 10 m resolution, this time difference becomes critical. Infrastructure expansion, small-scale deforestation, and other dynamic land use changes can manifest significantly over a four-to-eight-year period. Consequently, areas of low agreement between the DL-NI and older maps may reflect genuine, recent changes in anthropogenic pressure that the 10 m resolution successfully captures, rather than simple model error. This suggests that the comparison serves not just as a broad validation, but as an empirical demonstration of the DL-NI's superior temporal relevance for monitoring rapidly changing environments. This challenge is compounded by the pre-existing problem of consensus in the field: as detailed in Section 3.1.1, studies like Riggio et al. (2020) [45] show that traditional maps inherently exhibit major discrepancies due to subjective methodologies. The fact that the DL-NI achieves high correlation despite the temporal gap and the lack of consensus among the benchmarks validates the robustness of the data-driven model.

The Naturalness Index is limited by the validity of its input data. A fundamental constraint of both traditional and Deep Learning-driven Naturalness Index maps is that their validity is fundamentally tied to the completeness and accuracy of the geospatial data sources used to calculate the underlying human pressure layers. The NI, which serves as the supervision signal for the model, is built upon a combination of diverse public datasets, including OpenStreetMap, land use maps, population density grids, and nighttime lights. The accuracy of these source layers is heterogeneous; while OSM is an invaluable, crowd-sourced resource, its coverage and detail vary dramatically between regions, often suffering from sparsity or incompleteness, particularly in remote areas or parts of the Global South. Moreover, other pressure layers, such as those depicting livestock density or nighttime lights, are also imperfect proxies subject to their own inherent measurement errors and temporal lags. Therefore, the resulting NI, and consequently the deep learning model trained on it, can only be as good as this foundational input data. Any quantitative disagreement with established indices may reflect not just modeling differences, but fundamental discrepancies in the input pressure layers used by each respective framework.

With greater resolution comes greater noise. The high 10 m spatial resolution is a great asset for detecting fine-scale disturbances. However, the fact that human influence mapping is a task whose characteristic spatial resolution is unknown means that increasing the resolution is not always a marginal advantage. The influence dynamics can be intertwined and conflicting at different scales, depending on the level of coarseness required for the policy application. This challenge is exacerbated by the observation that the annotation source is at best offering weak supervision; thus at finer resolutions, the training signals will likely be less diluted. Lastly, if the underlying human influence is fundamentally a regional or landscape-scale phenomenon, then increasing the spatial

resolution risks introducing fine-grained noise that may not correlate to the conceptual scale of the influence being addressed.

Deployment on Sentinel-2 MGRS tiles versus mosaic. The initial approach considered using individual Sentinel-2 Military Grid Reference System (MGRS) tiles for deployment and stitching the predictions together. However, using temporally aggregated mosaics is preferred over individual static tiles. This decision is based on the assumption that the naturalness of an area is not expected to inherently change over the course of a year, even though surface appearance varies due to the phenology cycle (e.g., flora appearance across seasons, crops between planting and harvesting seasons, or snow cover on mountains). Since these transient visual changes are assumed not to inherently alter the state of human-induced naturalness of a land, the temporally collapsed mosaic provides a more stable and representative input for the prediction task.

On low-resource environments. The majority of existing large scale deep learning driven geospatial maps use cloud-based services for deployment. However, while cloud services provide convenience, they often make the resulting map products and underlying models less accessible to those operating under low-resource environments and in the Global South. The framework developed in this thesis accounts for this constraint. The entire pipeline, from architecture selection to the computation of large scale map creation, has been optimized to run with a minimal memory and computational footprint (i.e., memmap, Virtual Raster Tiles, Cloud Optimized GeoTIFFs (COGs) etc.).

(Geospatial) Deep Learning. This thesis chooses to use the wording *Geospatial Deep Learning* to underscore the need of designing deep learning solutions that are intrinsically aware of the unique characteristics of Earth Observation data. The inherent limitations of applying models designed for other data modalities to satellite imagery call for a specialized approach. Satellite data is a distinct modality, primarily because of the unique characteristics and challenges it presents. Observations are acquired from an overhead perspective and lack a natural orientation, unlike typical images, which means models must account for rotation. Target sizes span a logarithmic scale, from individual trees (less than 1 m) to large geographical features (greater than 1 km), demanding highly adaptive spatial resolution capabilities. Temporally, patterns range from hours to decades, requiring specialized handling of vast time series data. Spectral richness is another factor, as sensors capture diverse and characteristic signatures. Furthermore, the comprehensive view of Earth often requires fusing multiple sensors and derived products, creating complex multi-modal learning problems. Finally, the geospatial data's non-Euclidean, spherical distribution makes encoding its spatio-temporal position a unique challenge for architectures. All these raise the need for purpose-driven solutions rather than simple adaptation of models designed for natural images.

Take it to the end. This thesis offers an end-to-end framework that starts from the initial concept, proceeds through the creation of proxy annotations, dataset curation,

specialized model architecture design, large scale deployment, and concludes with measuring prediction reliability at a continental scale. Such complete, end-to-end flows are needed to move *Geospatial Deep Learning* from theoretical demonstrations to impactful, operational systems capable of informing policy and conservation efforts at scale.

Chapter 9

Conclusion and Perspectives

This final chapter concludes the thesis by summarizing its core contributions and offering perspectives on future research directions and the ethical implications of the work.

9.1 Summary and Conclusion

The central challenge addressed here was the development of a robust and scalable methodology for mapping human influence on the environment directly from satellite imagery. While traditional methods provided a valuable foundation for the field, their reliance on coarse, heterogeneous, and often temporally lagged geo-indicators has fundamentally limited their spatial consistency, accuracy, and temporal relevance.

Confronting these limitations, the work proposed an end-to-end deep learning framework, establishing a task termed *Naturalness Mapping* to quantify the absence of modern human influence. The research demonstrated a multi-stage approach that navigates the key challenges of *Geospatial Deep Learning*. The framework began by establishing a foundational data architecture, including the rule-based *Naturalness Index* for weak supervision annotation and the globally-sampled *MapInWild-NI* dataset. The proposed method uses this Naturalness Index as the annotation to train a supervised model that learns to map directly from satellite imagery to naturalness scores. Building upon the argument that satellite data constitutes a distinct modality, the methodology introduced a technique that first identified the most salient input data modality (Sentinel-2) through feature attribution, then tailored a geospatially-aware model architecture incorporating both contextual and geographic location priors. The deployment of this model resulted in a high-resolution map that achieves correlation with existing traditional indices at continental scales, while providing far superior spatial detail (10 m resolution) that successfully delineates fine-scale disturbances where traditional maps fail. Importantly, the proposed framework enables the creation of such maps for any given satellite imagery, removing the reliance on external geo-indicators.

This methodology was explicitly designed for real-life application, tackling the challenges of *Geospatial Deep Learning* head-on. The key components—feature attribution for data efficiency and a tailored architecture for geospatial priors—were developed with computational lightness and real-world scalability in mind.

The resulting Naturalness Map of Europe provides a spatial resolution of 10 m, representing an over 110-fold improvement compared to the existing deep learning baseline. Crucially, it maintains strong statistical alignment with established indices, yielding high Pearson correlation coefficients (r) exceeding 0.80 across major human influence benchmarks. To address the practical challenges of real-world deployment and the inherent ambiguity of the naturalness mapping task, the framework was augmented with a post-hoc out-of-distribution detector. This culminates in the production of two key outputs at a continental scale: a deep learning-derived *Naturalness Map* and its accompanying *Reliability Map*, which provides a crucial layer of transparency and trust for end-users by quantifying model prediction reliability. To conclude, this work provides not only a novel large scale map product but also a methodological blueprint for developing and deploying robust and reliable *Geospatial Deep Learning* solutions. By tackling the pipeline from data creation and architectural design to large scale deployment and responsible quantification of prediction reliability, the effort contributes to the broader goal of leveraging machine learning for planetary-scale environmental stewardship and provides a framework for future research to build upon.

9.2 Future Work

The demonstration of the Naturalness Map across the continental scale of Europe presents several clear directions for future research.

An obvious extension of this work is to deploy the final model globally and to generate a time series. Deploying the model across continents with distinct environmental characteristics will likely expose areas of significant distribution shift. Although the *Prediction Reliability Map* addresses this challenge, this scenario calls for future work on domain adaptation techniques. This need is motivated by several factors, including the distinction between the data processing levels used for training and deployment. Adopting the Copernicus terminology, the model is trained on Level-2A data, while the large scale deployment takes place on a mosaic of Level 3/4 products which are formed by temporally collapsing individual tiles. This change in processing level introduces a systematic domain shift apart from the inherent geographical differences. Additionally, the Sentinel-2 mission changed its processing baseline in early 2021, causing data acquired before and after the shift to be treated differently due to radiometric inconsistencies. This offset introduces a significant covariate shift that would render the model's predictions less reliable, even within the European domain. These technical data shifts are

considered separate from the traditional challenges expected in Earth Observation, such as acquisition artifacts, cloud contamination, and seasonal changes.

To address these technical domain shifts, future work should focus on integrating domain adaptation techniques, specifically those targeting data-level or pipeline shifts, to ensure the model remains robust across different Sentinel-2 processing baselines and aggregation levels. Furthermore, the distribution shift detection tool can be enhanced by incorporating additional measures of prediction uncertainty, such as epistemic uncertainty, to provide a more comprehensive assessment of model reliability in unseen regions.

9.3 Ethical Considerations

The ethical dimensions of this research have been a central consideration throughout its development. This section aims to transparently address the key ethical challenges and limitations inherent in developing a geospatial framework and deploying a large scale map of naturalness.

The definition of Naturalness and Human Influence. The most significant ethical consideration of this work lies in the very definition and interpretation of *naturalness*. This thesis operationalizes the concept as the *absence of modern anthropogenic pressure*, as detectable through geo-proxies like industrial infrastructure, large scale agriculture, and nighttime lights. It is critically important to state what this is not: it is not a measure of pristine wilderness, nor is it a judgment on the value of different forms of human-land interaction.

A core ethical principle guiding this work is that *human influence is not inherently negative*. Many landscapes that would register as highly natural on the map produced by this thesis are, in fact, the ancestral and current homes of Indigenous Peoples and local communities. These communities have been the stewards of their lands for millennia, shaping ecosystems in ways that are often sustainable and biodiversity-enhancing. The model developed here is not designed to detect, nor should it be interpreted as erasing, this form of human presence.

Indigenous rights and the risk of misinterpretation. The primary risk associated with this research is the potential for misinterpretation and misuse of the final map products. Equating high naturalness scores with "untouched", "unoccupied", or "empty" land is a dangerous oversimplification that has historically been used to justify the displacement of communities and the dispossession of their lands in the name of conservation ("fortress conservation").

This misinterpretation poses a direct risk to Indigenous Peoples and local communities. Therefore, it must be stated unequivocally that the map products from this thesis must not be used to undermine the rights, claims, or sovereignty of Indigenous Peoples, as recognized in frameworks such as the UN Declaration on the Rights of Indigenous

Peoples. The maps are intended to identify areas under pressure from industrial-scale activities, not to invalidate the presence and stewardship of communities who live in harmony with their environment.

Data bias and algorithmic limitations. The methodology itself is not free from biases. The *Naturalness Index*, which serves as the annotation source, is built upon geospatial proxies (e.g., OpenStreetMap, nighttime lights) that are known to have geographic biases, often being more complete in data-rich, more developed regions (i.e., Global North). This can lead to the model having a systematically different understanding of human influence in data-rich versus data-poor regions of the world.

The development of the accompanying *Reliability Map* is a direct response to this ethical challenge. It is a tool for epistemic humility, designed to be a transparent acknowledgment of the model's limitations. By highlighting areas where the model is encountering out-of-distribution data and its predictions are therefore less trustworthy, the reliability map serves as a crucial guardrail against naive or overconfident interpretation of the naturalness map. It is an essential component for enabling more responsible and informed decision-making.

A note on administrative boundaries. All administrative and national boundaries used for analysis or visualization in this thesis are derived from the OpenStreetMap (OSM) dataset. As OSM is a community-sourced project, the representation of these boundaries may not reflect the official position of all national governments and does not imply any endorsement of a particular geopolitical claim by the author.

Intended use and dual-use potential. The intended application of this work is to support scientific research, guide conservation planning, and inform policy aimed at mitigating the pressures of large scale industrial development on natural ecosystems. However, it is crucial to acknowledge the potential for dual-use. Like any large scale monitoring technology, these maps could potentially be misused for surveillance or to identify areas for resource extraction under the guise of low *human* influence. As the developer of this methodology, I advocate for its use exclusively in applications that are transparent, equitable, and aligned with international conservation and human rights standards.

Publications

The research presented in this thesis has led to several publications and the core chapters are primarily based on the works listed below. Additional related publications resulting from research conducted are also included for completeness.

Core Publications

- **Ekim B**, Tadesse G A, Robinson C, Hacheme G, Schmitt M, Dodhia R, Lavista Ferres J M (2025). "Distribution Shifts at Scale: Out-of-Distribution Detection in Earth Observation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2290–2299.

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

- **Ekim B**, Schmitt M (2024). "Multi-Scale Context Fusion for Pixel-Level Naturalness Mapping Using Sentinel-2 Imagery". *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, pp. 403–406.

- **Ekim B**, Schmitt M (2024). "Mapping Land Naturalness from Sentinel-2 Using Deep Contextual and Geographical Priors". *arXiv preprint arXiv:2406.19302*. Presented at Machine Learning for Remote Sensing Workshop, International Conference on Learning Representations, 2024.

- **Ekim B**, Stomberg T T, Roscher R, Schmitt M (2023). "MapInWild: A remote sensing dataset to address the question of what makes nature wild [Software and Data Sets]". *IEEE Geoscience and Remote Sensing Magazine* 11 (1), pp. 103–114.

- **Ekim B**, Schmitt M (2023). "Explaining Multimodal Data Fusion: Occlusion Analysis for Wilderness Mapping". *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, pp. 962–965.

- **Ekim B**, Schmitt M (2022). "Mapinwild: A Dataset for Global Wilderness Mapping". *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, pp. 231–234.

- **Ekim B**, Dong Z, Rashkovetsky D, Schmitt M (2021). "The naturalness index for the identification of natural areas on regional scale". *International Journal of Applied Earth Observation and Geoinformation* 105, 102622.

Related Publications

- Kochupillai M, Albrecht C M, **Ekim B**, Kahl M, Schmitt M, Taubenböck H (2024). "A Framework for Implementing Ethics-Aware Machine Intelligence: Insights from Earth Observation". *(Under Submission).*

- Landmann T, Schmitt M, **Ekim B**, Jandouwe V, Faith A, Jan C H, Henri E Z T (2023). "Insect Diversity is a Good Indicator of Biodiversity Status in Africa". *Communications Earth & Environment* 4, 234.

Bibliography

[1] Sanderson EW, Jaiteh M, Levy MA, Redford KH, Wannebo AV, Woolmer G (2002) The human footprint and the last of the wild: The human footprint is a global map of human influence on the land surface. *BioScience* 52(10): 891–904.

[2] Allan JR, Venter O, Watson JEM (2017) Temporally inter-comparable maps of terrestrial wilderness and the Last of the Wild. *Sci. Data* 4: 170187.

[3] Kennedy CM, Oakleaf JR, Theobald DM, Baruch-Mordo S, Kiesecker J (2019) Managing the middle: A shift in conservation priorities based on the global human modification gradient. *Glob. Chang. Biol.* 25(3): 811–826.

[4] OpenStreetMap contributors (2017) *Planet dump*. `https://planet.osm.org`.

[5] Torres R, Snoeij P, Geudtner D, Bibby D, Davidson M, Attema E, Potin P, Rommen B, Floury N, Brown M, Navas Traver I, Deghaye P, Duesmann B, Rosich B, Miranda N, Bruno C, L'Abbate M, Croci R, Pietropaolo A, Huchler M, Rostan F (2012) GMES Sentinel-1 mission. *Remote Sens. Environ.* 120: 9–24.

[6] Drusch M, Del Bello U, Carlier S, Colin O, Fernandez V, Gascon F, Hoersch B, Isola C, Laberinti P, Martimort P, Meygret A, Spoto F, Sy O, Marchese F, Bargellini P (2012) Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sens. Environ.* 120: 25–36.

[7] Elvidge CD, Zhizhin M, Ghosh T, Hsu F.-C, Taneja J (2021) Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sens.* 13(5): 922.

[8] Zanaga D, Van De Kerchove R, De Keersmaecker W, Souverijns N, Brockmann C, Quast R, Wevers J, Grosu A, Paccini A, Vergnaud S, Cartus O, Santoro M, Fritz S, Georgieva I, Lesiv M, Carter S, Herold M, Li L, Tsendbazar N.-E, Ramoino F, Arino O (2021) *ESA WorldCover 10 m 2020 v100*. Zenodo.

[9] Venter ZS, Barton DN, Chakraborty T, Simensen T, Singh G (2022) Global 10 m land use land cover datasets: A comparison of Dynamic World, World Cover and Esri Land Cover. *Remote Sens.* 14(16): 4101.

[10] Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L, Seal D, Shaffer S, Shimada J, Umland J, Werner M, Oskin M, Burbank D, Alsdorf D (2007) The Shuttle Radar Topography Mission. *Rev. Geophys.* 45(2): RG2004.

[11] Tadono T, Ishida H, Oda F, Naito S, Minakawa K, Iwamoto H (2014) Precise global DEM generation by ALOS PRISM. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* II-4: 71–76.

[12] Haklay M (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap ordnance survey datasets. *Environ. Plan. B: Plan. Des.* 37(4): 682–703.

[13] Barrington-Leigh C, Millard-Ball A (2017) The world's user-generated road map is more than 80% complete. *PLoS One* 12(8): e0180698.

[14] LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc. IEEE* 86(11): 2278–2324.

[15] LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553): 436–444.

[16] Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60(6): 84–90.

[17] Luo W, Li Y, Urtasun R, Zemel R (2017) Understanding the effective receptive field in deep convolutional neural networks. *arXiv preprint* 1701.04128.

[18] Mou L, Zhu XX (2019) Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* 4(2): 22–40.

[19] Rolf E, Klemmer K, Robinson C, Kerner H (2024) Mission critical: Satellite data is a distinct modality in machine learning. *arXiv preprint* 2402.01444.

[20] Tuia D, Persello C, Bruzzone L (2016) Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geosci. Remote Sens. Mag.* 4(2): 41–57.

[21] Rolf E (2023) Evaluation challenges for geospatial ML. *arXiv preprint* 2303.18087.

[22] McCloskey JM, Spalding H (1989) A reconnaissance-level inventory of the amount of wilderness remaining in the world. *Ambio* 18(4): 221–227.

[23] Hannah L, Carr JL, Lankerani A (1995) Human disturbance and natural habitat: a biome level analysis of a global data set. *Biodivers. Conserv.* 4(2): 128–155.

[24] Wackernagel M, Rees W (1996) *Our ecological footprint: Reducing human impact on the Earth.* Philadelphia, PA: New Society Publishers.

[25] Gassert F, Venter O, Watson JEM, Brumby S, Mazzariello JC, Atkinson S, Balch WE, Bennet A, Biggs P, Brancalion P, Broadbent EN, Burger JR, Butsic V, Clark ML, Cousins SAO, Creighton K, Du L, Ellis E, Ferreira J, Finer M, Gerber LR, Gorelick N, Hancher M, Hansen MC, Heiner M, Herold M, Hudson LN, Joppa L, Kennedy CM, Kiesecker J, Klemmer K, Kuemmerle T, Laurance WF, Li X, Lindsell JA, Lucas R, Magrach A, Malhi Y, McClure CJW, Meyfroidt P, Mitchard ETA, Moore NJ, Mu H, Navarro L, Nelson E, Newbold T, Oakleaf J, Oldekop JA, Pimm SL, Poff NL, Potapov P, Riggio J, Ripple WJ, Robinson C, Robinson JG, Roehrdanz PR, Rolf E, Rudel TK, Sanderson EW, Schneider FD, Scholes RJ, Seto KC, Sherman J, Simensen T, Smith SJ, Strassburg BBN, Theobald DM, Tilman D, Tulloch AIT, Venter Z, Visconti P, Wang Y, Ward M, Williams BA, Woodcock CE, Wu W, Young L, Zinda JA, Robinson N (2023) An operational approach to near real time global high resolution mapping of the terrestrial human footprint. *Front. Remote Sens.* 4.

[26] Venter O, Sanderson EW, Magrach A, Allan JR, Beher J, Jones KR, Possingham HP, Laurance WF, Wood P, Fekete BM, Levy MA, Watson JEM (2016) Global terrestrial human footprint maps for 1993 and 2009. *Sci. Data* 3(1): 1–10.

[27] Stevens WK (1995) *Miracle under the oaks: The revival of nature in America*. New York: Simon and Schuster.

[28] Rojstaczer S, Sterling SM, Moore NJ (2001) Human appropriation of photosynthesis products. *Science* 294(5551): 2549–2552.

[29] Chapin III FS, Zavaleta ES, Eviner VT, Naylor RL, Vitousek PM, Reynolds HL, Hooper DU, Lavorel S, Sala OE, Hobbie SE, Mack MC, Díaz S (2000) Consequences of changing biodiversity. *Nature* 405(6783): 234–242.

[30] Redford KH, Richter BD (1999) Conservation of biodiversity in a world of use. *Conserv. Biol.* 13(6): 1246–1256.

[31] Venter O, Sanderson EW, Magrach A, Allan JR, Beher J, Jones KR, Possingham HP, Laurance WF, Wood P, Fekete BM, Levy MA, Watson JEM (2016) Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nat. Commun.* 7(1): 12558.

[32] Watson JE, Shanahan DF, Di Marco M, Venter O, Sanderson J, Mackey B, Possingham HP (2016) Catastrophic declines in wilderness areas undermine global environment targets. *Curr. Biol.* 26(21): 2929–2934.

[33] Tulloch VJD, Tulloch AIT, Visconti P, Halpern BS, Watson JEM, Evans MC, Auerbach NA, Barnes M, Beger M, Chadès I, Giakoumi S, McDonald-Madden E, Murray NJ, Ringma J, Possingham HP (2015) Why do we map threats? Linking threat mapping with actions to make better conservation decisions. *Front. Ecol. Environ.* 13(2): 91–99.

[34] Garnett ST, Burgess ND, Fa JE, Fernández-Llamazares Á, Molnár Z, Robinson CJ, Watson JEM, Zander KK, Austin B, Brondizio ES, Collier NF, Duncan T, Ellis E, Geyle H, Jackson MV, Jonas H, Malmer P, McGowan B, Sivongxay A, Leiper I (2018) A spatial overview of the global importance of Indigenous lands for conservation. *Nat. Sustain.* 1(7): 369–374.

[35] Sanderson EW, Fisher K, Robinson N, Allan J, Bastianelli M, Bicknell J, Beyer H, Chaplin-Kramer R, Ferrier S, Foppen R, Gaughan A, Gibson L, Halpern B, Hoskins A, Jones K, Kiesecker J, Kuemmerle T, Melchiorri M, Mu H, Navarro L, O'Connor B, Osti M, Peixoto R, Potapov P, Riggio J, Robinson J, Rolf E, Simensen T, Stevens F, Strassburg B, Svenning J.-C, Theobald D, Venter O, Venter Z, Watson J, Wintle B (2022) The march of the human footprint. *EcoEvoRxiv.*

[36] Williams BA, Venter O, Allan JR, Atkinson SC, Rehbein JA, Ward M, Di Marco M, Grantham HS, Ervin J, Goetz SJ, Hansen AJ, Jantz P, Pillay R, Rodríguez-Buriticá S, Supples C, Virnig A, Watson JE (2020) Change in terrestrial human footprint drives continued loss of intact ecosystems. *One Earth* 3(3): 371–382.

[37] Skinner EB, Glidden CK, MacDonald AJ, Mordecai EA (2023) Human footprint is associated with shifts in the assemblages of major vector-borne diseases. *Nat. Sustain.* 6: 652–661.

[38] Theobald DM, Kennedy C, Chen B, Oakleaf J, Baruch-Mordo S, Kiesecker J (2020) Earth transformed: detailed mapping of global human modification from 1990 to 2017. *Earth Syst. Sci. Data* 12(3): 1953–1972.

[39] Mu H, Li X, Wen Y, Huang J, Du P, Su W, Miao S, Geng M, Biggs P (2022) A global record of annual terrestrial human footprint dataset from 2000 to 2018. *Sci. Data* 9(1): 176.

[40] Zhou J, Wu N, Lu T (2025) Mapping human footprint changes over Qingzang Plateau. *Geogr. Sustain.* 6(3): 100249.

[41] Keys PW, Barnes EA, Carter NH (2021) A machine-learning approach to human footprint index estimation with applications to sustainable development. *Environ. Res. Lett.* 16(4): 044061.

[42] Orihuela-Pinto B, Keys PW, Davenport FV, Barnes EA (2025) Uncovering patterns of converging human-induced pressure on global lands. *Mach. Learn.: Earth.*

[43] Ye J, Xiao C, Feng Z, Qiao T (2024) A review of global wilderness area identification since the 21st century. *J. Environ. Manage.* 358: 120946.

[44] Barrington-Leigh C, Millard-Ball A (2017) The world's user-generated road map is more than 80% complete. *PLoS ONE* 12(8): 1–20.

[45] Riggio J, Baillie JEM, Brumby S, Ellis E, Kennedy CM, Oakleaf JR, Tait A, Tepe T, Theobald DM, Venter O, Watson JEM, Jacobson AP (2020) Global human influence maps reveal clear opportunities in conserving Earth's remaining intact terrestrial ecosystems. *Glob. Chang. Biol.* 26(8): 4344–4356.

[46] Kim K.-H, Park M, Park T, Lee W.-K (2025) Spatial heterogeneity patterns along the human footprint gradient and their ecological implications: A case study in South Korea. *Ecol. Indic.* 177: 113771.

[47] Cheng H, Johansen K, Jin B, Ma Q, Li W (2025) Human footprint with machine learning identifies risks of the invasive weed Conyza sumatrensis across land-use types under climate change. *Glob. Ecol. Conserv.*: e03657.

[48] Presotto A, Hamilton S, Izar P (2025) A 10-meter resolution human footprint dataset to support biodiversity and conservation studies in Brazil. *Sci. Data* 12(1): 1754.

[49] Emam A, Stomberg TT, Roscher R (2024) Leveraging activation maximization and generative adversarial training to recognize and explain patterns in natural areas in satellite imagery. *IEEE Geosci. Remote Sens. Lett.* 21: 1–5.

[50] Emam A, Farag M, Roscher R (2024) Confident naturalness explanation (CNE): A framework to explain and assess patterns forming naturalness. *IEEE Geosci. Remote Sens. Lett.* 21: 1–5.

[51] Emam A, Roscher R (2025) Confidence-Filtered Relevance (CFR): An interpretable and uncertainty-aware machine learning framework for naturalness assessment in satellite imagery. *arXiv preprint arXiv:2507.13034.*

[52] Emam A, Farag M, Rußwurm M, Roscher R (2026) NaT-ReX: Naturalness assessment with transformer-based reliable explainability. In: Proc. of DAGM GCPR: 571–585.

[53] Sirko W, Kashubin S, Ritter M, Annkah A, Bouchareb YE, Dauphin Y, Keysers D, Neumann M, Quinn C, Quinn T (2021) Continental-scale building detection from high resolution satellite imagery. *arXiv preprint* 2107.12283.

[54] Lang N, Jetz W, Schindler K, Wegner JD (2023) A high-resolution canopy height model of the Earth. *Nat. Ecol. Evol.* 7(11): 1778–1789.

[55] Tolan J, Yang H.-I, Nosarzewski B, Couairon G, Vo HV, Brandt J, Spore J, Majumdar S, Haziza D, Vamaraju J, Moutakanni T, Bojanowski P, Johns T, White B, Tiecke T, Couprie C (2024) Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sens. Environ.* 300: 113888.

[56] Karaman K, Jiang Y, Robert D, Sainte Fare Garnot V, Santos MJ, Wegner JD (2025) GSR4B: Biomass map super-resolution with Sentinel-1/2 guidance. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* X-G-2025: 487–494.

[57] Jain M (2020) The benefits and pitfalls of using satellite data for causal inference. *Rev. Environ. Econ. Policy* 14(1): 157–169.

[58] Lu K, Kluger DM, Bates S, Wang S (2025) Regression coefficient estimation from remote sensing maps. *arXiv preprint* 2407.13659.

[59] The World Bank Group (2018) *Mainstreaming the use of remote sensing data and applications in operational contexts*. Tech. rep. Washington, D.C.: The World Bank.

[60] United Nations Satellite Imagery and Geospatial Data Task Team (2017) *Earth observations for official statistics*. Tech. rep. United Nations.

[61] Meyer H, Pebesma E (2021) Estimating the area of applicability of remote sensing-based machine learning models with limited training data. In: Proc. of IGARSS: 2028–2030.

[62] Meyer H, Pebesma E (2022) Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat. Commun.* 13(1): 2208.

[63] Cohrs K.-H, Osika Z, Gonzalez-Calabuig M, Murillo-Montoya S, Roscher R (2025) SHRUG-FM: Reliability-aware foundation models for Earth observation. *arXiv preprint* 2511.10370.

[64] Milà C, Mateu J, Pebesma E, Meyer H (2022) Nearest neighbour distance matching leave-one-out cross-validation for map validation. *Methods Ecol. Evol.* 13(6): 1304–1316.

[65] Meyer H, Reudenbach C, Wöllauer S, Nauss T (2019) Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecol. Model.* 411: 108815.

[66] Tong X.-Y, Wang S (2025) Invariant features for global crop type classification. *arXiv preprint* 2509.03497.

[67] Butsko C, Van Tricht K, Tseng G, Kerner H (2025) Deploying geospatial foundation models in the real world: Lessons from WorldCereal. *arXiv preprint* 2508.00858.

[68] Huang B, Reichman D, Collins LM, Bradbury K, Malof JM (2019) Tiling and stitching segmentation output for remote sensing: Basic challenges and recommendations. *arXiv preprint* 1805.12219.

[69] Neumann M, Raichuk A, Jiang Y, Stanimirova R, Sims M, Slay C, Harris N, Hansen M (2025) Natural forests of the world – a 2020 baseline for deforestation and degradation monitoring. *Sci. Data* 12(1): 1715.

[70] Abrahams E, Snow T, Siegfried MR, Pérez F (2024) A concise tiling strategy for preserving spatial context in Earth observation imagery. *arXiv preprint* 2404.10927.

[71] Roscher R, Russwurm M, Gevaert C, Kampffmeyer M, Dos Santos JA, Vakalopoulou M, Hansch R, Hansen S, Nogueira K, Prexl J, Tuia D (2024) Better, not just more: Data-centric machine learning for Earth observation. *IEEE Geosci. Remote Sens. Mag.* 12(4): 335–355.

[72] Roscher R, Bohn B, Duarte MF, Garcke J (2020) Explainable machine learning for scientific insights and discoveries. *IEEE Access* 8: 42200–42216.

[73] Chandrashekar G, Sahin F (2014) A survey on feature selection methods. *Comput. Electr. Eng.* 40(1): 16–28. 40th-year commemorative issue.

[74] Lundberg S, Lee S.-I (2017) A unified approach to interpreting model predictions. *arXiv preprint* 1705.07874.

[75] Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": Explaining the predictions of any classifier. *arXiv preprint* 1602.04938.

[76] Bastidas AA, Tang H (2019) Channel attention networks. In: Proc. of CVPR Workshops.

[77] Liang L, Zhang S, Li J, Plaza A, Cui Z (2023) Multi-scale spectral-spatial attention network for hyperspectral image classification combining 2D octave and 3D convolutional neural networks. *Remote Sens.* 15(7): 1758.

[78] Kakogeorgiou I, Karantzalos K (2021) Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* 103.

[79] Ibrahim S, Balzter H, Tansey K (2024) Machine learning feature importance selection for predicting aboveground biomass in African savannah with Landsat 8 and ALOS PALSAR data. *Mach. Learn. Appl.* 16: 100561.

[80] Hooker S, Erhan D, Kindermans P.-J, Kim B (2019) A benchmark for interpretability methods in deep neural networks. *arXiv preprint* 1806.10758.

[81] Najjar H, Nuske M, Dengel A (2024) Data-centric machine learning for Earth observation: Necessary and sufficient features. *arXiv preprint* 2408.11384.

[82] Zaryabi EH, Moradi L, Kalantar B, Ueda N, Halin AA (2022) Unboxing the black box of attention mechanisms in remote sensing big data using XAI. *Remote Sens.* 14(24): 6254.

[83] Rao A, Rolf E (2025) Using multiple input modalities can improve data-efficiency and O.O.D. generalization for ML with satellite imagery. *arXiv preprint* 2507.13385.

[84] Neto E, Faria FA, de Oliveira AAS, Fazenda ÁL (2024) A satellite band selection framework for Amazon forest deforestation detection task. *arXiv preprint* 2404.02659.

[85] Stomberg TT, Stone T, Leonhardt J, Roscher R (2022) Exploring wilderness characteristics using explainable machine learning in satellite imagery. *arXiv preprint* 2203.00379.

[86] Mena F, Arenas D, Miranda M, Dengel A (2025) On what depends the robustness of multi-source models to missing data in Earth observation? *arXiv preprint* 2503.19719.

[87] Gong Y, Xiao Z, Tan X, Qian H, Liu W (2020) Context-aware convolutional neural network for object detection in VHR remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* 58(1): 34–44.

[88] Zhu FW, Jerzak CT, Daoud A (2024) Optimizing multi-scale representations to detect effect heterogeneity using Earth observation and computer vision: Applications to two anti-poverty RCTs. *arXiv preprint* 2411.02134.

[89] Audebert N, Le Saux B, Lefèvre S (2017) Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks. In: Proc. of ACCV: 180–196.

[90] Benedetti P, Ienco D, Gaetano R, Ose K, Pensa RG, Dupuy S (2018) M^3Fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 11(12): 4939–4949.

[91] Danish MS, Munir MA, Shah SRA, Khan S, Khan FS (2025) TerraFM: A scalable foundation model for unified multisensor Earth observation. *arXiv preprint* 2506.06281.

[92] Huang T, Chen T, Wang Z, Liu S (2023) The counterattack of CNNs in self-supervised learning: Larger kernel size might be all you need. *arXiv preprint* 2312.05695.

[93] Gupta R, Li S, Zhu T, Malik J, Darrell T, Sangiovanni-Vincentelli A (2024) xt: Nested tokenization for larger context in large images. *arXiv preprint* 2403.01915.

[94] Ghassemi B, Fraga-Dantas C, Gaetano R, Ienco D (2025) Geographical context matters: Bridging fine and coarse spatial information to enhance continental land cover mapping. *arXiv preprint* 2504.12368.

[95] Tang K, Paluri M, Fei-Fei L, Fergus R, Bourdev L (2015) Improving image classification with location context. In: Proc. of ICCV: 1008–1016.

[96] Ayush K, Uzkent B, Meng C, Tanmay K, Burke M, Lobell D, Ermon S (2022) Geography-aware self-supervised learning. *arXiv preprint* 2011.09980.

[97] Madadikhaljan M, Schmitt M (2025) Geolocation-aware deep coding. *PFG* 93(1): 3–18.

[98] Mai G, Lao N, He Y, Song J, Ermon S (2023) CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations. *arXiv preprint* 2305.01118.

[99] Klemmer K, Rolf E, Robinson C, Mackey L, Rußwurm M (2023) SatCLIP: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint* 2311.17179.

[100] Dhakal A, Sastry S, Khanal S, Zelikman E, Green A, Jacobs N (2025) RANGE: Retrieval augmented neural fields for multi-resolution geo-embeddings. In: Proc. of CVPR: 24680–24689.

[101] Sastry S, Khanal S, Dhakal A, Diwan A, Jacobs N (2024) Birdsat: Cross-view contrastive masked autoencoders for bird species classification and mapping. In: Proc. of WACV: 7136–7145.

[102] Cepeda VV, Nayak GK, Shah M (2023) GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *arXiv preprint* 2309.16020.

[103] Liu Z, Zhang F, Jiao J, Wang Y, Wang Y (2025) GAIR: Improving multimodal geo-foundation model with geo-aligned implicit representations. *arXiv preprint* 2503.16683.

[104] Haas L, Skreta M, Alberti S, Finn C (2024) Pigeon: Predicting image geolocations. In: Proc. of CVPR: 12893–12902.

[105] He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proc. of ICCV: 1026–1034.

[106] Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60(6): 84–90.

[107] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. *arXiv preprint* 1312.6199.

[108] Hendrycks D, Gimpel K (2016) A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint* 1610.02136.

[109] Nguyen A, Yosinski J, Clune J (2015) Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proc. of CVPR: 427–436.

[110] Liang S, Li Y, Srikant R (2020) Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint* 1706.02690.

[111] Liu W, Wang X, Owens JD, Li Y (2021) Energy-based out-of-distribution detection. *arXiv preprint* 2010.03759.

[112] Lee K, Lee K, Lee H, Shin J (2018) A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *arXiv preprint* 1807.03888.

[113] Sun Y, Guo C, Li Y (2021) ReAct: Out-of-distribution detection with rectified activations. *arXiv preprint* 2111.12797.

[114] Sun Y, Li Y (2022) DICE: Leveraging sparsification for out-of-distribution detection. *arXiv preprint* 2111.09805.

[115] Djurisic A, Bozanic N, Ashok A, Liu R (2023) Extremely simple activation shaping for out-of-distribution detection. In: Proc. of ICLR. Published online: 1–13.

[116] Gulati A, Dong X, Hurtado C, Gan W (2024) Out-of-distribution detection through soft clustering with non-negative kernel regression. *arXiv preprint* 2407.13141.

[117] Sinhamahapatra P, Koner R, Roscher K, Günnemann S (2022) Is it all a cluster game? Exploring out-of-distribution detection based on clustering in the embedding space. *arXiv preprint* 2203.08549.

[118] Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, Hu W, Yasunaga M, Phillips RL, Gao I, Lee T, David E, Stavness I, Guo W, Earnshaw B, Haque I, Beery SM, Leskovec J, Kundaje A, Pierson E, Levine S, Finn C, Liang P (2021) Wilds: A benchmark of in-the-wild distribution shifts. In: Proc. of ICML: 5637–5664.

[119] Kerner H, Sundar S, Satish M (2024) Multi-region transfer learning for segmentation of crop field boundaries in satellite images with limited labels. *arXiv preprint* 2404.00179.

[120] Kerner H, Nakalembe C, Yang A, Zvonkov I, Tseng G, Becker-Reshef I (2024) How accurate are existing land cover maps for agriculture in Sub-Saharan Africa? *Sci. Data* 11(1): 486.

[121] Aiken E, Rolf E, Blumenstock J (2023) Fairness and representation in satellite-based poverty maps: Evidence of urban-rural disparities and their impacts on downstream policy. *arXiv preprint* 2305.01783.

[122] Gawlikowski J, Saha S, Kruspe A, Zhu XX (2021) Out-of-distribution detection in satellite image classification. *arXiv preprint* 2104.05442.

[123] Coca M, Coca Neagoe I, Datcu M (2022) Hybrid DNN-Dirichlet anomaly detection and ranking: Case of burned areas discovery. *IEEE Trans. Geosci. Remote Sens.* 60: 1–16.

[124] Le Bellier G, Audebert N (2024) Detecting out-of-distribution Earth observation images with diffusion models. *arXiv preprint* 2404.12667.

[125] Xie SM, Kumar A, Jones R, Khani F, Ma T, Liang P (2020) In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. *arXiv preprint* 2012.04550.

[126] Gawlikowski J, Saha S, Niebling J, Zhu XX (2023) Handling unexpected inputs: incorporating source-wise out-of-distribution detection into SAR-optical data fusion for scene classification. *EURASIP J. Adv. Signal Process.* 2023(1).

[127] Al-Emadi SA, Yang Y, Ofli F (2025) Analysing satellite imagery classification under spatial domain shift across geographic regions. *Int. J. Comput. Vis.* 133: 7672–7709.

[128] Crasto R (2025) Robustness to geographic distribution shift using location encoders. *arXiv preprint* 2503.02036.

[129] Zhang H, Klemmer K, Rolf E, Alvarez-Melis D (2025) Predicting out-of-domain performance under geographic distribution shifts. In: Proc. of ICLR Workshop on Tackling Climate Change with Machine Learning. Paper 20: 1–5.

[130] Ekim B, Dong Z, Rashkovetsky D, Schmitt M (2021) The naturalness index for the identification of natural areas on regional scale. *Int. J. Appl. Earth Obs. Geoinf.* 105: 102622.

[131] Ekim B, Schmitt M (2022) MapInWild: A dataset for global wilderness mapping. In: Proc. of IGARSS: 231–234.

[132] UNEP-WCMC, IUCN (2023) *Protected Planet: The World Database on Protected Areas (WDPA)*. Cambridge, UK.

[133] Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202: 18–27.

[134] Ekim B, Stomberg TT, Roscher R, Schmitt M (2023) MapInWild: A remote sensing dataset to address the question of what makes nature wild. *IEEE Geosci. Remote Sens. Mag.* 11(1): 103–114.

[135] Ekim B, Schmitt M (2023) Explaining multimodal data fusion: Occlusion analysis for wilderness mapping. *arXiv preprint* 2304.02407.

[136] Ekim B, Schmitt M (2024) Mapping land naturalness from Sentinel-2 using deep contextual and geographical priors. *arXiv preprint* 2406.19302.

[137] Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. *arXiv preprint* 1505.04597.

[138] Ekim B, Tadesse GA, Robinson C, Ortiz A, Dodhia R, Lavista Ferres JM (2025) Distribution shifts at scale: Out-of-distribution detection in Earth observation. In: Proc. of CVPR Workshops: 2290–2299.

[139] Santos LA, Ferreira KR, Camara G, Picoli M, Simoes R (2021) Quality control and class noise reduction of satellite image time series. *ISPRS J. Photogramm. Remote Sens.* 177: 75–88.

[140] Stewart AJ, Robinson C, Corley IA, Ortiz A, Lavista Ferres JM, Banerjee A (2025) TorchGeo: Deep learning with geospatial data. *ACM Trans. Spat. Algorithms Syst.* 11(4): 1–28.

[141] Bastani F, Wolters P, Gupta R, Ferdinan J, Kembhavi A (2023) Satlaspretrain: A large-scale dataset for remote sensing image understanding. In: Proc. of ICCV: 16772–16782.

[142] Helber P, Bischke B, Dengel A, Borth D (2019) EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *arXiv preprint* 1709.00029.

[143] Gupta R, Hosfelt R, Sajeev S, Patel N, Goodman B, Doshi J, Heim E, Choset H, Gaston M (2019) xBD: A dataset for assessing building damage from satellite imagery. *arXiv preprint* 1911.09296.

[144] Boeing G (2025) Modeling and analyzing urban networks and amenities with OSMnx. *Geogr. Anal.* 57(4): 567–577.