Doctoral Dissertation
Doctoral Program in Computer Engineering (38.th cycle)

Spatio-Temporal Machine Learning for Ecology and Crisis Management

Daniele Rege Cambrin
* * * * *

Supervisors
Prof. Garza Paolo, Supervisor
Prof. Apiletti Daniele, Co-supervisor

Doctoral examination committee
Prof. Silvia Anna Chiusano, Politecnico di Torino
Prof. Attilio Fiandrotti, Referee, Università degli Studi di Torino
Prof. Dino Ienco, Referee, INRAE, UMR TETIS, University of Montpellier

Politecnico di Torino
January 30, 2026

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . . . .

Daniele Rege Cambrin

Turin, January 30, 2026

Summary

The exponential growth of spatio-temporal data from satellites and digital communications presents a great opportunity to address planetary-scale challenges. However, standard deep learning models, trained on conventional datasets, are often suboptimal for the specialized, context-dense, and scarce data that characterize domains like ecology and crisis management. This difference creates a significant bottleneck, limiting the conversion of raw data into actionable intelligence.

This thesis argues that effectively monitoring our planet requires a unified framework capable of handling systems in two distinct but interconnected states: long-term, gradual evolution (Ecology) and short-term, acute shocks (Crisis Management). Challenging the current trend of relying on monolithic, general-purpose models, this work demonstrates that a portfolio of targeted, data-efficient, and interpretable architectures provides a more robust, scalable, and trustworthy solution.

To achieve this, the thesis introduces a suite of novel machine learning models and foundational, public datasets designed to tackle the core challenges of specialized data. The contributions include: 1) data-efficient methods that achieve state-of-the-art performance on scarce or sparsely labeled data; 2) architectures that synthesize dynamic, multi-modal data streams into a coherent operational picture; and 3) explainable models that build trust with domain experts.

The proposed models demonstrate state-of-the-art performance across all tasks. In ecological monitoring, the proposed architecture for cross-modal land use classification achieved 54% greater accuracy than existing methods. In crisis management, the real-time tracker provides a significant computational advantage, operating at a constant cost regardless of data scale (in contrast to the linear complexity of competitors), while maintaining state-of-the-art accuracy. The creation of these four new benchmark datasets was essential to both achieving and validating these performance gains.

Ultimately, this thesis establishes that specialized, efficient systems are essential for unlocking the full potential of complex spatio-temporal data. The developed models and datasets offer immediate practical benefits for ecologists and crisis managers, enhancing decision-making in precision agriculture, water management, and disaster response to help build a more resilient and sustainable future.

Acknowledgements

This dissertation would not have been possible without the support and encouragement of many people. I would like to thank them sincerely.

First, I thank my parents, Claudia and Riccardo, for their unconditional love and for supporting me in every choice throughout my life. To my brother, Federico, thank you for always making me smile, especially during the last year we have spent together. I also hold dear the childhood memories shared with my grandparents, Alessandro, Anna, Domenico, and Lucia, my aunt Stefania, and my cousins Giada and Alessio.

My journey was enriched by my friends: my hometown friends, with whom I shared so many incredible trips, and my university friends, partners in many absurd events.

My daily life in our often grey lab was made significantly brighter by my colleagues. I am also grateful to the TorchGeo community for providing invaluable opportunities. The many people I met at conferences also deserve thanks for the insightful conversations that helped shape my future path.

I would like to thank my supervisor, Paolo. His support during these years was fundamental. I am grateful for the freedom he gave me to explore my ideas and for his encouragement during the most challenging moments.

To everyone mentioned and to the many others who have been a positive force in my life, thank you.

This thesis is dedicated to the incredible people in my life: to the memory of those who have passed, to those who are here, and to those who will come.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Today, science is defined by two major trends: an exponential growth in the volume and variety of spatio-temporal data in conjunction with an escalating urgency to address planetary-scale challenges. The proliferation of earth observation satellites, in-situ sensors, and digital communication platforms provides a huge amount of data, offering an unprecedented, high-resolution view of our world [108, 2, 154, 50, 91]. However, this volume of information presents an incredible challenge: to convert these vast, raw data streams into actionable intelligence requires a new generation of information systems and machine learning tools capable of navigating their complexity. This thesis is motivated by the critical need to understand and manage our planet's complex systems across their full spectrum of behavior. To do so, we must develop tools capable of monitoring both long-term, gradual changes and responding to short-term, acute shocks. This work focuses on two domains that represent these fundamental states: Ecology, the science of systems in dynamic equilibrium, and Crisis Management, the science of systems under sudden, intense stress. However, converting these raw data streams into actionable intelligence is limited by the highly specialized, context-dependent, and often scarce nature of the data itself. This thesis argues that the central challenge is learning from this 'non-standard' data. To solve this, we develop a new generation of data-efficient methods designed to translate complex spatio-temporal information into the clear intelligence needed to manage crises and guide our ecological future.

1.1 A common challenge

Despite significant advancements in data acquisition, the methods used to analyze this information in ecology and crisis management often proved suboptimal. Traditional statistical models and early machine learning approaches can struggle with the non-linear, dynamic, and multi-faceted nature of spatio-temporal data [163, 182, 197]. While large, general-purpose deep learning models are promising,

their computational and data requirements often render them impractical for many real-world applications, where resources or data are limited [158]. Additionally, the models that excel at interpreting everyday data often fail when confronted with the peculiar characteristics or rarity of these domains. Standard deep learning methodologies, often developed for domains with vast, well-curated datasets like ImageNet, do not easily transfer to the unique complexities of remote sensing data [203], which is often denser and contains richer context than a typical natural image. A photograph of a goldfish is not a 12-band multispectral image of a forest, as shown in Figure 1.1. This challenge of learning from dense, specialized data is not an isolated problem; it is the visual manifestation of the same fundamental bottleneck we observe in textual crisis data.

(a) Image from ImageNet [44] (b) Image from BEN [39]

Figure 1.1: Comparison between a natural image (ImageNet) and a remotely sensed image (BEN) in RGB

For example, when dealing with the rare context of a crisis, even common data types like text present unique challenges. Crisis text is often characterized by non-standard syntax, emergent jargon, and an urgent, implicit context that is not understandable without specialized knowledge. A standard model, lacking this context, fails to distinguish between noise and critical signal. An effective model must learn to differentiate personal anecdotes from specific, factual reports of damage. A casual tweet is not an actionable report from a disaster zone, as shown in Table 1.1

This thesis argues that solving these challenges requires a holistic approach that addresses planetary systems in two distinct but interconnected states. The first is the state of gradual evolution, where the goal is to monitor slow changes, understand long-term trends, and manage resources sustainably. This is the domain of Ecology. The second is the state of 'shock', where the goal is to detect, assess, and respond to sudden events that destabilize systems. This is the domain of Crisis Management. A truly robust spatio-temporal methodology must be effective in both regimes. By developing and validating specialized machine learning models for each, this thesis builds a comprehensive framework for planetary monitoring. Additionally, the "black box" nature of deep complex models can limit trust and

Table 1.1: Discerning Signal from Noise during a Hurricane

NOISE (Appears Relevant)	**SIGNAL (Actually Relevant)**
"The wind is howling and the power just flickered again here in South Tampa. So scary watching the trees bend like that. Praying for everyone in the path of #HurricaneLeo."	*"Just saw the power lines go down on Bay Street near MacDill. A transformer blew. The whole block is dark now in South Tampa. #HurricaneLeo."*
Analysis: Why It's Difficult to Discern	
Both tweets use identical keywords: "power," "South Tampa," and the hashtag "#HurricaneLeo." Both convey fear and describe the storm's effects. A simple keyword-based model would flag both as relevant. However, the tweet on the left is a personal experience with no verifiable, specific information for responders. The tweet on the right, in contrast, provides a precise location ("Bay Street near MacDill") and a cause ("a transformer blew") for a critical infrastructure failure. This is actionable data for utility companies and emergency services. An effective model must therefore learn to differentiate personal anecdotes from specific, factual reports of damage.	

adoption by domain experts who require a clear understanding of a model's reasoning to make impactful decisions [61, 146]. This thesis directly confronts these limitations, addressing several research gaps, providing a suite of novel machine learning models and foundational datasets that significantly advance the state-of-the-art in data-driven ecological monitoring and crisis management. This thesis challenges the current trend of relying on monolithic, general-purpose models for planetary-scale problems. It argues that, given highly specialized and often scarce data, a portfolio of targeted, computationally efficient, and more interpretable architectures provides a more robust, scalable, and trustworthy path for monitoring planetary systems through both long-term ecological change and rapid crisis events. By creating novel models and foundational public datasets, this work transforms complex, multi-modal satellite and text data into actionable, real-world intelligence, setting new performance benchmarks in several key applications.

This thesis contributes to the advancement in three main ways:

1. Learning from Specialized Data. We directly tackle the core challenge by developing foundational datasets (CaBuAr [25], QuakeSet [137], HydroChronos [28], and CrisisLandMark [29]) and data-efficient methods (DAC [136], Magnifier [26]) that demonstrate how to achieve state-of-the-art performance when training data is scarce, specialized, or lacks conventional labels.

2. Synthesizing Dynamic, Sparse Streams. We address the temporal dimension of data specialization, building architectures that synthesize evolving, multi-modal data streams into a coherent picture. This is crucial for both the slow evolution of ecological systems (ACTU [28] and CLOSP [29]) and the rapid, sparse information flow of a crisis (DQNC2S [135]).

3. Building Trustworthy Models. We confront the "black box" problem by designing inherently explainable (U-KAN [138]) and transparent architectures that are critical for adoption by domain experts who must make high-stakes decisions based on limited evidence. A similar approach was adopted in HydroChronos [28] to foster transparency in hydrology and guide future research.

The thesis is organized as a step-by-step journey:

- Chapter 2: This chapter establishes the technical foundations for the following contributions. It provides a comprehensive overview of essential concepts for a broad audience, covering Semantic Segmentation, Depth Estimation, Contrastive Learning, Spatio-Temporal Forecasting, Information Retrieval, Deep Q Networks, Kolmogorov-Arnold Networks, the 'Black Box' problem, and Remote Sensing fundamentals.

- Chapter 3: This chapter details new methods for long-term ecological monitoring. It introduces new architectures and benchmarks designed to overcome data constraints and improve model interpretability for crop segmentation, forest canopy analysis, surface water forecasting, and land cover.

- Chapter 4: This chapter focuses on methods for short-term crisis events. It presents foundational datasets for assessing disaster damage from satellite imagery and introduces novel architectures for data-efficient mapping and real-time text summarization in crisis scenarios.

- Chapter 5: The final chapter synthesizes the findings, discusses the broader implications of the developed methods, states the limitations, and outlines future directions.

Chapter 2

Foundations

This chapter constructs our methodological foundation. We begin with a fundamental data source, Remote Sensing, establishing the nature of the spatio-temporal observations. From there, we explore core Perception tasks (Semantic Segmentation and Depth Estimation) to extract meaning from this raw data. We then introduce the Time dimension with Spatio-Temporal Forecasting. Recognizing that imagery alone is insufficient, we then introduce methods for processing textual data (Information Retrieval) and for fusing it with our visual understanding (Contrastive Learning). With this rich, multi-modal view of the world, we turn to Decision-Making using Deep Q-Networks. Finally, we address the critical need for Trust and Interpretability, exploring the 'Black Box' Problem and emerging architectures like KANs that aim to solve it.

2.1 Remote Sensing

The machine learning paradigms discussed in this thesis are mainly applied to remote sensing data, which can power large-scale analysis and have a more complex nature than natural imagery. Remote sensing is the science of acquiring information about an object, area, or phenomenon, analyzing data acquired by a device that is not in physical contact with the object under investigation. It provides the rich, multi-scale, spatio-temporal observations of the Earth's surface that are essential for monitoring ecological systems and managing crisis events. Understanding the foundational principles of how this data is generated is necessary to effectively apply machine learning models for its analysis.

2.1.1 The Physics of Observation

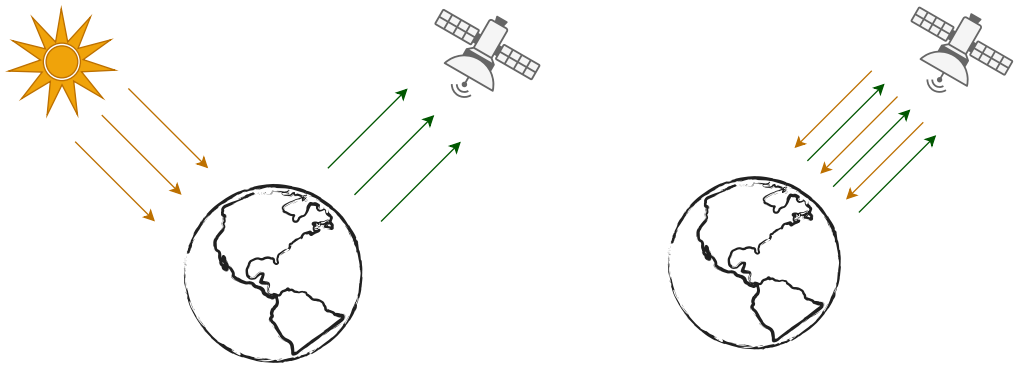
In most remote sensing systems, sensors are designed to capture electromagnetic radiation that has interacted with the Earth's surface. These interactions (reflection, absorption, and emission) vary depending on the physical and chemical

properties of the surface materials. This principle allows us to infer characteristics of the surface from the recorded radiation.

Remote sensing systems can be broadly categorized into two classes based on their energy source:

Passive Sensors These systems measure naturally available energy. Most commonly, they detect solar radiation that is reflected by the Earth's surface as shown in Figure 2.1. Optical sensors, such as those on the Landsat [186] or Sentinel-2 [104] satellites, are common examples. Landsat also measures thermal infrared radiation emitted by the Earth itself, which is a function of surface temperature.

Active Sensors These systems provide their own source of energy to illuminate the target. The sensor emits a controlled pulse of radiation and measures the portion that is scattered back from the surface. With this approach is able to acquire data at any time of day or night and under most weather conditions. Key examples are Light Detection and Ranging (LiDAR), which uses laser pulses, and Synthetic Aperture Radar (SAR) [171], which uses microwave pulses as shown in Figure 2.1.



(a) Passive Sensor (b) Active sensor

Figure 2.1: (a) Passive sensors depend on the Sun for illumination and measure reflected solar energy. (b) Active sensors emit their own pulses of energy and record the backscattered signal.

2.1.2 Core Data Characteristics: The Four Resolutions

The information content of remotely sensed data is defined by four fundamental resolutions. These dimensions affect the nature and scale of phenomena that can be observed and constitute the primary axes of the data space.

Spatial Resolution This refers to the size of the smallest feature that can be resolved by the sensor, determined by the ground area represented by a single pixel. A high spatial resolution (e.g., 0.5 meters) allows for the detection of fine-scale objects like individual trees or buildings. A lower spatial resolution (e.g., 30 meters) covers a larger area per pixel and allows only to see larger features like agricultural fields. An image of the same area from two satellites with different resolutions can be seen in Figure 2.2.

(a) Sentinel-2 [104] at 10m resolution (b) WorldView [107] at $<$ 1m resolution

Figure 2.2: A comparison of medium and very high spatial resolution satellite imagery. The Sentinel-2 scene (a) is at 10 m resolution, while (b) shows the same area from a WorldView satellite at $<$1 m resolution. The increased resolution in (b) significantly enhances the ability to distinguish small-scale ground features.

Spectral Resolution This defines the number and width of the specific wavelength intervals in the electromagnetic spectrum (EMS) to which a sensor is sensitive. A sensor with low spectral resolution (panchromatic) might have a single wide band, whereas a high spectral resolution sensor (hyperspectral) can have hundreds of narrow, contiguous bands as shown in Figure 2.3. The spectral signature acts as a feature vector for each pixel. This high-dimensional feature space enables discrimination between materials with subtle differences.

Temporal Resolution This is the revisit frequency of a sensor for a specific location on the Earth's surface. A high temporal resolution (e.g., daily revisits) is essential for monitoring quickly evolving dynamic processes. It provides the critical time dimension in the spatio-temporal data cube as shown in Figure 2.4.

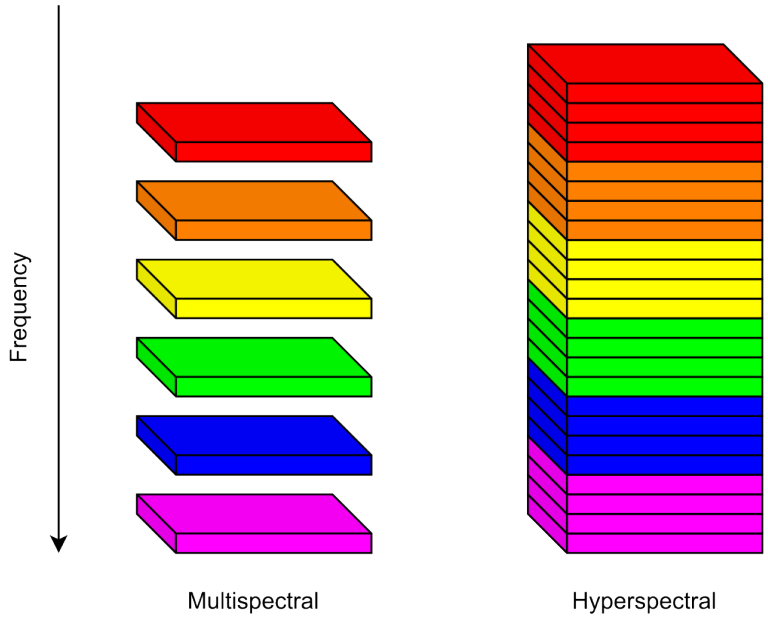


Figure 2.3: A conceptual illustration of spectral resolution. Multispectral imaging (left) captures data across a small number of broad, discrete spectral bands. In contrast, hyperspectral imaging (right) samples the electromagnetic spectrum using hundreds of narrow, contiguous bands.

(a) January (b) March (c) June (d) September

Figure 2.4: An illustration of temporal resolution using a multi-temporal series of satellite images. The images show the same geographic area in (a) January, (b) March, (c) June, and (d) September. They show the seasonal phenological cycles, highlighting the role of high temporal resolution data in monitoring dynamic processes over time. Images from SSL4EO [177]

Radiometric Resolution This describes the sensor's ability to distinguish fine gradations in energy levels, which means differences in the intensity of the electromagnetic signal. It is typically quantified by the number of bits used to store

the signal value for each pixel (e.g., 8-bit, 12-bit, or 16-bit). Higher radiometric resolution allows for the detection of even minor energy variations.

2.1.3 Platforms and Data Products

Remote sensing data is acquired from a variety of platforms, each involving a trade-off between spatial resolution and geographic coverage. Satellites provide global coverage with systematic, repeated observations, making them ideal for large-scale, long-term monitoring. Aircraft and Unmanned Aerial Vehicles (UAVs) offer much higher spatial resolution and acquisition flexibility for smaller, targeted areas, which is often required for immediate crisis response or localized ecological studies. The data from these platforms is processed into different "levels." Lower levels mean rawer data, whereas as the level grows, the data become more processed and analysis-ready (ARD).

These four resolutions (spatial, spectral, temporal, and radiometric) define the characteristics of the common data we work with. The challenge, and the focus of the following sections, is to convert this vast stream of raw pixel values into actionable knowledge.

2.2 Semantic Segmentation

To extract meaningful information from the multispectral data provided by remote sensing, our first task is to assign a semantic label to every pixel. Semantic segmentation provides the fundamental computer vision framework for many tasks. It classifies each pixel of an image, as shown in Figure 2.5. This is different from image classification, which assigns a single label to an entire image. It is suited for applications where a complete understanding of the image content is needed, like environmental monitoring (e.g., delineating forests [12] or water bodies [95]), precision agriculture (e.g., mapping crop fields [24, 11]), disaster response (e.g., assessing burned areas [86, 18]), and medical image analysis [144, 74, 196].

2.2.1 Foundational Architectures

The dominant paradigm for semantic segmentation is the encoder-decoder architecture. In this architecture, an encoder network progressively downsamples the input image. This process reduces spatial resolution to create a more abstract representation to capture high-level semantic context. The decoder network then takes these low-resolution, semantically rich features and progressively upsamples them to construct a full-resolution segmentation map where each pixel is assigned a class label.

(a) RGB Image (b) Segmentation Mask

Figure 2.5: An illustration of the semantic segmentation task. The input is a standard RGB image, and the objective is to produce a segmentation mask, where each pixel is assigned a label corresponding to its semantic class (e.g., road, building, vegetation). Images from ADE20K [200]

A well-known, foundational implementation of this paradigm is the U-Net architecture [144]. The U-Net is characterized by its symmetric, U-shaped structure comprising a contracting path (the encoder) and an expansive path (the decoder) as shown in Figure 2.6. It uses skip connections to concatenate feature maps from the encoder with the corresponding feature maps in the decoder. These connections allow the decoder to leverage both deep, semantic features from the later encoder layers and fine-grained, high-resolution features from the earlier layers. This fusion avoids losing information during the downsampling process.

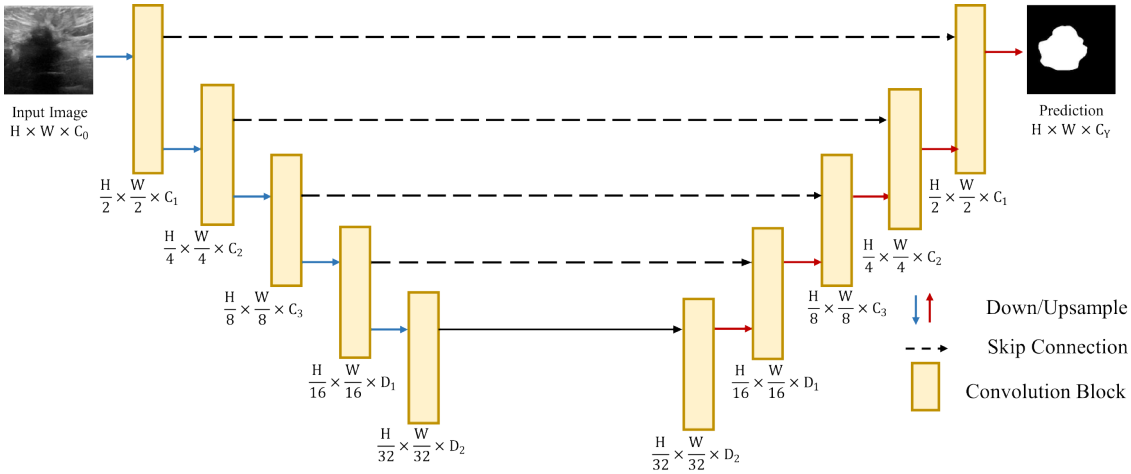


Figure 2.6: A schematic of the U-Net architecture [144, 139]. The model comprises a contracting path (encoder) and an expansive path (decoder). Skip connections pass feature maps from the encoder to the decoder, combining deep information with shallow details to generate the final segmentation map. The dimensions at each level indicate the change in spatial resolution (H×W) and channels (C_i, D_i).

To better capture multi-scale context, which is effective in segmenting objects of varying sizes, various solutions were proposed. One notable technique is the use of atrous (or dilated) convolutions, which increase the receptive field of filters without increasing the number of parameters or losing spatial resolution. DeepLab [32, 34, 33] processes the features using atrous convolutions in parallel at different dilation rates with multiple fields of view simultaneously.

More recently, Vision Transformers (ViTs) [48] have been adapted for dense prediction tasks. By dividing an image into patches and using a self-attention mechanism, ViTs can model long-range dependencies across the entire image more effectively than the localized receptive fields of CNNs. However, the original ViT architecture produces a single-resolution feature map without multi-scale features.

To address this, hierarchical transformers like the Swin Transformer [100, 99] were developed. The Swin Transformer introduces a hierarchical structure by merging image patches in deeper layers, creating feature maps at different scales (Figure 2.7). Additionally, it computes self-attention within local windows that are shifted across the image in successive layers (Figure 2.7). In this way, it can model long-range dependencies while maintaining a linear computational complexity with respect to the image size.

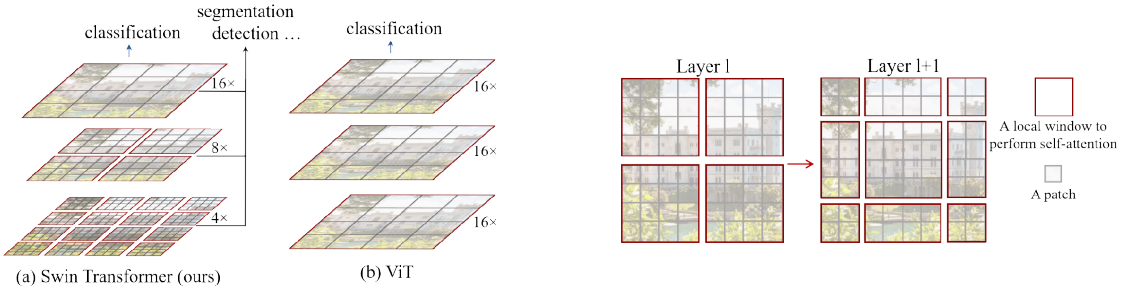


Figure 2.7: An overview of the Swin Transformer architecture from [100]. (Left) Swin builds a hierarchical feature pyramid, in contrast to the standard Vision Transformer (ViT), which maintains a single feature resolution. (Right) In the shifted window approach, the self-attention is computed in local windows (Layer l) that are then shifted in the next layer (Layer l+1) to enable cross-window connections.

Another popular transformer-based architecture is SegFormer [190]. It features a hierarchical transformer encoder and a lightweight MLP decoder. The encoder is designed to produce multi-scale features without the need for positional encodings, which makes it robust to variations in image resolution during inference. The MLP decoder aggregates information from the different scales of the encoder to produce the final segmentation map. It is a popular choice for real-time semantic segmentation.

The Dense Prediction Transformer (DPT) [132] is another influential architecture that adapts the ViT for dense prediction tasks. DPT maintains a high-resolution representation in the transformer encoder. It then employs a convolutional decoder to reassemble the transformer's output tokens into a full-resolution prediction. This approach allows DPT to capture fine-grained details and global context simultaneously, leading to highly coherent and accurate segmentation maps.

While transformer-based models have shown impressive performance, there is also a strong need for efficient models that can run on resource-constrained devices. MobileNets [71, 153, 72, 128] are a family of lightweight CNN architectures that use depthwise separable convolutions. They factorize a standard convolution into a depthwise convolution and a pointwise convolution as shown in Figure 2.8. This reduces the number of parameters and computational cost, making them an excellent choice for mobile and real-time semantic segmentation applications.

Modern ConvNeXt [183, 101] architecture represents a reworking of traditional CNNs, inspired by the design of Vision Transformers. They can achieve performance competitive with transformers on various vision tasks, thanks to the incorporation of architectural choices from transformers, such as larger kernel sizes, inverted bottleneck structures, and layer normalization. ConvNeXt demonstrates that with the right design choices, CNNs can provide a strong and efficient alternative to transformer-based architectures.

2.2.2 Key Components and Loss Functions

The choice of loss function is critical to ensure consistent shapes in segmentation, particularly in scenarios with significant class imbalance. This is common in satellite imagery where a target class (e.g., burned area) may occupy only small, isolated areas. Cross-entropy loss is a common baseline; however, it can be biased towards the majority class, and it does not directly optimize for region-based properties like shapes. To address these issues, the Dice Loss [112], derived from the Dice-Sørensen coefficient, and Focal Loss [96] are widely used.

Dice Loss The Dice coefficient is a measure of overlap between two sets, similar to the F1-score, and is defined for a predicted segmentation mask P and a ground truth mask T as:

$$DSC = \frac{2|P \cap T|}{|P| + |T|}$$

By directly optimizing this overlap metric, Dice Loss [112] (which is a surrogate differentiable version of the metric) can produce more spatially consistent and accurate shapes, especially for smaller objects.

Focal Loss Focal Loss [96] modifies the standard cross-entropy loss to down-weight the contribution of easy-to-classify examples, thereby forcing the model to

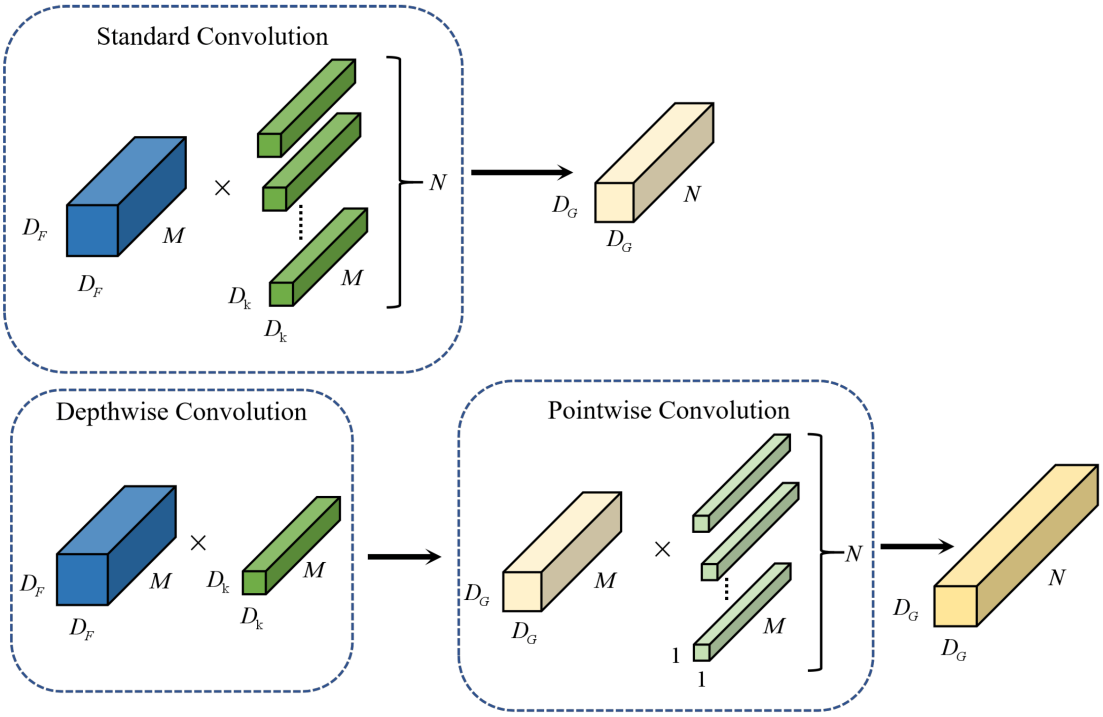


Figure 2.8: A visual comparison of a standard convolution and a depthwise separable convolution. **(Top)** A standard convolution operation where an input tensor of size $D_F \times D_F \times M$ is transformed by N filters, each of size $D_k \times D_k \times M$. This operation simultaneously processes spatial and cross-channel correlations to produce an output feature map of size $D_G \times D_G \times N$. **(Bottom)** A depthwise separable convolution, which factorizes the standard operation into two distinct steps. First, a depthwise convolution applies a single spatial filter of size $D_k \times D_k \times 1$ to each of the M input channels independently. Second, a pointwise convolution uses N filters of size $1 \times 1 \times M$ to linearly combine the output channels from the depthwise step. Image adapted from [75].

focus on hard negatives. The Focal Loss is defined as:

$$L_{\text{Focal}}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

Here, p_t is the model's estimated probability for the ground truth class, α_t is a balancing parameter to address class imbalance directly, and $\gamma \geq 0$ is the tunable focusing parameter. As γ increases, the modulating factor $(1 - p_t)^\gamma$ reduces the loss for well-classified examples (i.e., where $p_t \to 1$), shifting the model's attention towards misclassified examples.

Hybrid Loss Functions In practice, in many segmentation tasks, a single loss function may not be sufficient to address all challenges simultaneously. Therefore,

hybrid loss [167, 196, 77] functions that combine the strengths of multiple formulations are often employed. A popular and effective strategy is to create a linear combination of a distribution-based loss (like Cross-Entropy or Focal Loss) and a region-based loss (like Dice Loss):

$$L = \lambda_1 L_1 + \lambda_2 L_2$$

The hyperparameters λ_1 and λ_2 balance the influence of each component.

Semantic segmentation helps generate detailed 2D maps that answer the question, 'What is on the ground?'. While this two-dimensional understanding is critical, many applications require a three-dimensional understanding of the area.

2.3 Depth Estimation

Understanding the three-dimensional structure of an environment, for applications like estimating forest biomass or assessing building damage, is the domain of depth estimation. Monocular depth estimation (MDE) is a computer vision task that involves inferring a dense depth map from a single 2D image, as shown in Figure 2.9. Applying MDE to remote sensing data presents unique challenges. Unlike ground-level imagery, where strong cues like perspective and object occlusion exist, top-down views lack these features. Depth must instead be inferred from subtle cues like solar illumination angles and cast shadows. The objective is to assign a continuous depth or distance value to every pixel to convert a flat image into a 3D representation of the scene from the camera's viewpoint. Alternative data acquisition methods like direct 3D measurements using technologies like Light Detection and Ranging (LiDAR) are often prohibitively expensive and logistically complex to deploy at a global scale. Consequently, estimating depth from 2D imagery has become an efficient and scalable alternative for many applications, including autonomous navigation, robotics, 3D scene reconstruction, and environmental analysis [189, 113, 136].

The task is typically framed as a pixel-wise regression problem, where a model learns a mapping from an input image $I \in \mathbb{R}^{H \times W \times C}$ to an output depth map $D \in \mathbb{R}^{H \times W}$. The output map is composed of continuous values and not discrete classes. Additionally, a distinction in this field is between relative depth estimation, which predicts ordinal scale-agnostic relationships, and monocular metric depth estimation (MMDE), which predicts absolute depth in physical units (e.g., meters).

2.3.1 Foundational Architectures

The evolution of monocular depth estimation has been linked to advancements in deep learning architectures. These models have progressively improved, moving from early convolutional designs to sophisticated generative frameworks.

(a) RGB Image (b) Depth Map

Figure 2.9: An illustration of monocular depth estimation. The model infers a dense depth map from a single RGB image. In the depth map, warmer colors (e.g., red) represent closer objects, while cooler colors (e.g., purple) indicate objects farther from the camera. Adapted from [136]

The initial transition from traditional geometric methods to learning-based approaches introduced a multi-scale Convolutional Neural Network (CNN) [54] to extract both the global scene layout and the fine-scale details. After this work, encoder-decoder architectures became a dominant paradigm in the field. This structure proved effective for generating high-resolution depth maps while maintaining both global structural consistency and local detail.

To address the inherent ambiguity and ill-posed nature of MDE, researchers developed more specialized architectures. A significant innovation was the move from direct depth regression to adaptive binning strategies [56, 17]. Rather than predicting a continuous depth value for each pixel, these methods discretize the depth range into a set of bins and predict a probability distribution over them. The final depth is computed as a linear combination of the bin centers.

More recently, the field has adopted generative models, such as diffusion models. Generative architectures like Marigold [83] frame depth estimation as an iterative denoising generation process. This approach has proven effective for complex geometric structures.

2.3.2 Training Paradigms and Data

The early deep learning methods were trained in a supervised manner. This approach requires large-scale datasets containing images paired with pixel-aligned ground-truth depth maps, typically captured using specialized hardware like LiDAR sensors or stereo cameras. Acquiring and annotating this data is costly. To overcome these limitations, the field shifted towards self-supervised learning. In this way, it is possible to effectively employ vast quantities of unlabeled data as a supervision signal without the need for explicit depth labels.

To produce a universal model capable of performing well across a wide range of unseen environments without any fine-tuning requires training on a mixture of diverse datasets. Additionally, specialized loss functions, such as scale-and-shift-invariant losses, are necessary to allow learning relative depth relationships that are consistent across different settings. The MiDAS [90] model demonstrates that training on a heterogeneous collection of datasets provides cross-domain generalizability.

Real-world datasets provide genuine scene distributions and textures, but they often suffer from imperfect labels, including blurred object boundaries, missing depth values, and sensor noise. Instead, synthetic datasets, generated using photorealistic rendering engines, offer a powerful alternative. They provide pixel-perfect, noise-free ground truth for depth, surface normals, and other geometric properties. This is particularly advantageous for supervising complex structures and challenging optical phenomena like transparency and reflection, where real-world sensors fail. However, models trained exclusively on synthetic data often struggle to generalize to real-world images due to the "domain gap". Models like Depth Anything V2 [195] are trained on a combination of real and synthetic data. This often involves techniques to bridge the domain gap, such as pseudo-labeling, where a "teacher" model (often trained on high-quality synthetic data) generates depth labels for a large corpus of unlabeled real-world images. This allows a "student" model to learn from the precise geometric supervision of synthetic data while also adapting to the rich visual diversity of real-world scenes. The advantages of the various approaches are summarized in Table 2.1.

Table 2.1: Comparison of training paradigms for Monocular Depth Estimation.

Paradigm	**Data Requirement**	**Advantages**	**Challenges**
Supervised	Paired RGB images and depth maps	High accuracy. The learning objective is straightforward.	Expensive, labor-intensive. Suffers from sensor noise.
SSL	Large quantities of unlabeled monocular image pairs.	No specialized depth sensors.	Relies on photometric consistency assumptions.
Hybrid	Photorealistic synthetic data and real-world images.	Synthetic pixel-perfect depthmaps. Visual diversity of the real world.	The "synth-to-real domain gap" must be bridged.

2.3.3 Loss Functions for Depth Regression

Initially, depth estimation was framed as a pixel-wise regression problem, employing loss functions such as the L2 or L1 loss. However, the L2 loss penalty for large errors can lead to the suppression of fine details and result in overly smooth depth maps. The L1 loss, while robust to outliers, may struggle in sharp discontinuities at object boundaries.

A major challenge in monocular depth estimation is the inherent ambiguity of absolute scale. A model trained on one dataset with a specific camera and scene scale may fail to generalize to another. Scale-invariant loss functions [54] were introduced to address this issue. They are independent of the absolute scale and shift of the depth maps. To solve the issue of blurred edges and loss of fine-grained detail, researchers developed gradient-based losses [195]. These functions operate on the gradients of the depth map, explicitly penalizing inconsistencies in local geometric structure. Nowadays, in practice, most state-of-the-art models employ a hybrid loss function composed of multiple components.

These methods allow us to infer a third dimension from flat imagery. However, both ecological systems and crisis events are not static; they are dynamic processes that evolve over time, demanding methods that can model this temporal dimension.

2.4 Spatio-Temporal Forecasting

The static 2D and 3D representations derived from segmentation and depth estimation offer a snapshot of a given moment. However, to move from reactive monitoring to proactive management, we must be able to predict how these snapshots will evolve. This is the core challenge of spatio-temporal forecasting. It is the task of predicting the future state of a system across a set of spatial locations, given the history of its past states. A model has to capture the intricate, underlying dynamics governing the system's evolution. This requires simultaneously considering two forms of dependency: temporal dependency, which describes how the state at a single location evolves over time, and spatial dependency, which describes how the states of different locations influence one another at any given instant. In ecology, the challenge is often modeling slow processes like the gradual urbanization over many years. In crisis management, the focus shifts to rapid event propagation, such as modeling the evolution of a wildfire. To formalize this task, we can represent the state of the entire system at a specific time T as a "snapshot" that contains the feature values for all N spatial locations. The forecasting problem then becomes: given a sequence of historical snapshots, predict a sequence of future snapshots. This process is conditioned on the underlying spatial structure, which defines the relationships between the locations.

2.4.1 Foundational Architectures

Classical time-series models like the AutoRegressive Integrated Moving Average (ARIMA) were generalized to Vector AutoRegression (VAR) for multivariate time series, and further to Space-Time ARIMA (STARIMA). These models incorporate spatial and temporal autoregressive and moving average terms. However, they typically rely on strong assumptions, such as linearity and stationarity (i.e., that the statistical properties of the system do not change over time or space), which are often violated in complex systems. The limitations of classical models and the advent of machine learning and deep learning changed the landscape of spatio-temporal forecasting.

Temporal Dependency Recurrent Neural Networks (RNNs), especially variants like Long Short-Term Memory (LSTM) [70] and Gated Recurrent Units (GRU) [36], are designed to process sequential data. They maintain an internal state that captures information from past time steps, making them a natural choice for modeling the temporal dependency. More recently also transformers are also employed to model long sequences.

Spatial Dependency Convolutional Neural Networks (CNNs) are the standard for processing grid-based data like images. Their convolutional filters can learn spatial hierarchies of features, from simple edges to complex shapes. They provide an effective solution to model local-temporal dependencies. While ViT [48] can be employed, many solutions still use convolution, eventually combined with transformers. Graph Neural Networks (GNNs) [156] are also used to model spatial dependencies. GNNs operate via a message-passing mechanism, where nodes iteratively aggregate information from their neighbors. This allows the model to learn representations that are explicitly aware of the topology.

Spatio-Temporal Dependencies The most powerful models handle both spatial and temporal dependencies within a unified architecture. A seminal approach for grid-based data is the Convolutional LSTM (ConvLSTM) [164]. It fuses the capabilities of CNNs and LSTMs by replacing the matrix multiplication operations inside the LSTM gates with convolutional operations. This modification allows the model's recurrent state to maintain a 2D spatial structure, enabling it to process spatial features at each time step while modeling their temporal evolution as shown in Figure 2.10.

Another strategy involves extending convolutions directly into the time dimension, employing 3D Convolutional Neural Networks. This principle is often integrated into encoder-decoder structures, such as the 3D U-Net [37]. For data structured as graphs, Spatio-Temporal GNNs [194] have become the standard. They typically alternate between a GNN layer (for spatial dependencies) and a

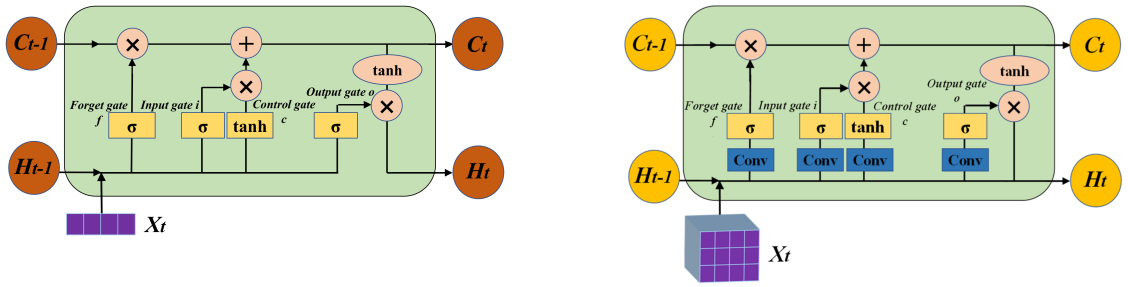


Figure 2.10: A comparison between LSTM and Convolutional LSTM (ConvLSTM) architectures. On the left, an LSTM cell, which processes 1D vector inputs (X_t) through gates regulated by fully connected operations. On the right, a ConvLSTM cell, where the matrix multiplications are replaced with convolution operations. Adapted from [43]

temporal modeling layer that captures the evolution of the node states over time. More recently, Spatio-Temporal Transformers [6, 192] have emerged. These models tokenize the input data into a sequence of spatio-temporal patches and apply a self-attention mechanism to learn global dependencies between any two patches in space and time, as shown in Figure 2.11. This makes them particularly adept at capturing the long-range interactions.

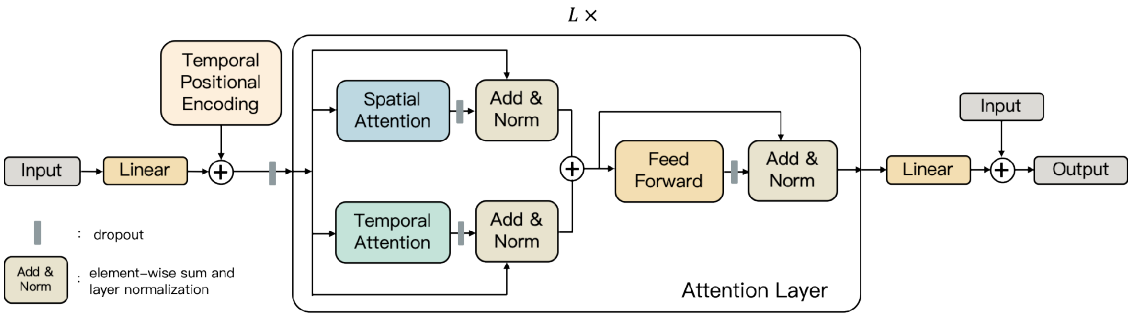


Figure 2.11: A Spatio-Temporal Transformer block. Input embeddings are first combined with temporal positional encodings. The block then processes the data through two parallel streams: a Spatial Attention module to capture dependencies across spatial locations and a Temporal Attention module for dependencies over time. Adapted from [6]

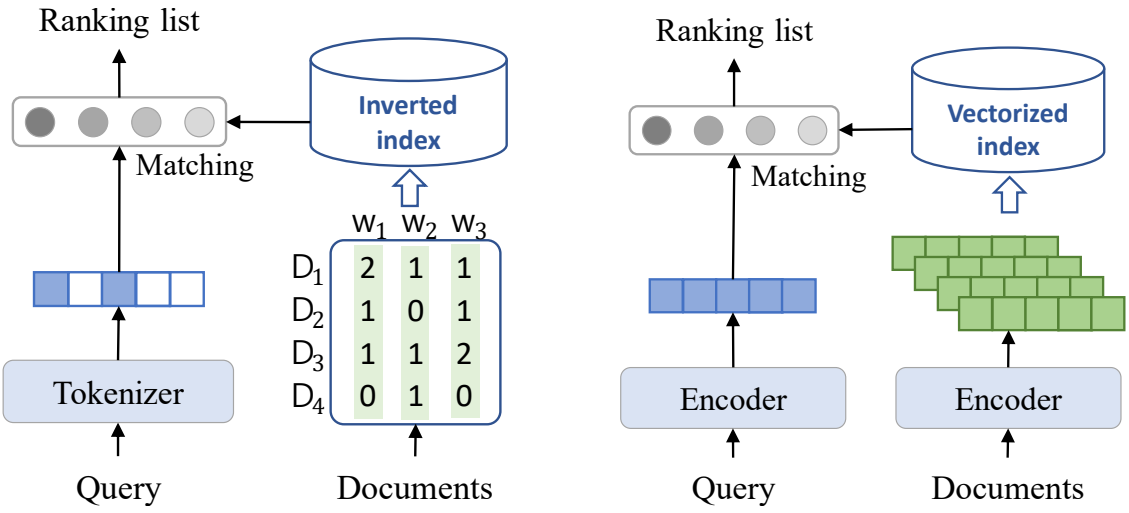
These architectures allow us to learn the underlying dynamics of a system from past observations to forecast its future state. While forecasting models can predict what will happen based on past observations, they cannot explain why without external context. To bridge this gap, we must incorporate methods for processing context provided by human observation and reporting, generally in the form of unstructured textual data.

2.5 Information Retrieval

Forecasting models built on sensor data can capture physical dynamics, but they often miss context that explains the 'why'. During a natural disaster, for instance, this context is contained in high-velocity streams of unstructured information, from official field reports to eyewitness accounts. Filtering this data to find actionable intelligence is a critical challenge for Information Retrieval (IR). It aims to find material in a large collection of an unstructured nature that satisfies a user's information need [105]. IR aims to connect users with relevant information efficiently and effectively. Its applications are widespread, ranging from web search engines and digital libraries to more specialized domains such as e-discovery, patent search, and evidence-based medicine.

2.5.1 Retrieval Paradigms

Modern IR systems have largely evolved from sparse retrieval methods to dense retrieval, which leverages the semantic power of deep learning models. Their functioning is summarized in Figure 2.12.



(a) Sparse Retrieval (b) Dense Retrieval

Figure 2.12: A schematic of sparse and dense retrieval. On the left, sparse retrieval systems operate on lexical overlap. Queries are tokenized and matched against an inverted index that stores term-document occurrences. This results in a sparse representation. On the right, dense retrieval systems use neural encoders to embed both queries and documents into a common low-dimensional, dense vector space. Retrieval is performed by searching for the nearest document vectors to the query vector. Adapted from [201]

Sparse Retrieval Okapi BM25 [142] (Best Match 25) sparse retriever is a classical foundation of IR. BM25 is a probabilistic ranking function [143] that scores documents based on the query terms they contain. It operates on a "bag-of-words" representation, where documents and queries are seen as high-dimensional, sparse vectors. The score for a document D given a query Q (containing terms $q_1, \ldots, q_n$) is calculated as follows:

$$\text{Score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

The formula's components are:

- $\text{IDF}(q_i)$: It is the Inverse Document Frequency and assigns a higher value to terms that are rare in the collection, as they are considered more informative. N is the total number of documents and $n(q_i)$ is the number of documents containing the term q_i.
- $f(q_i, D)$: The frequency of the term q_i in the document D.
- $|D|$ and avgdl: The length of document D and the average document length in the collection, respectively.
- k_1 (typically $\in [1.2, 2.0]$) calibrates how the score scales with term frequency, preventing documents with many occurrences of a term from dominating.
- b (typically 0.75) normalizes for document length, reducing the bias towards longer documents, which have a higher chance of containing query terms.

Despite its simplicity, BM25 remains an exceptionally strong and efficient baseline, making it a standard benchmark.

Dense Retrieval Dense retrieval [82] represents a paradigm shift in IR. Both queries and documents are encoded into low-dimensional dense vectors, known as embeddings, generated by deep neural networks. They are designed to capture the semantic meaning of the content. The retrieval process is framed as a nearest neighbor search problem within this embedding space, using a similarity function to measure the proximity between the query vector and the document vectors. Documents with the highest similarity scores are considered the most relevant. This approach allows for the retrieval of documents that are semantically related to the query, even if they do not share common keywords.

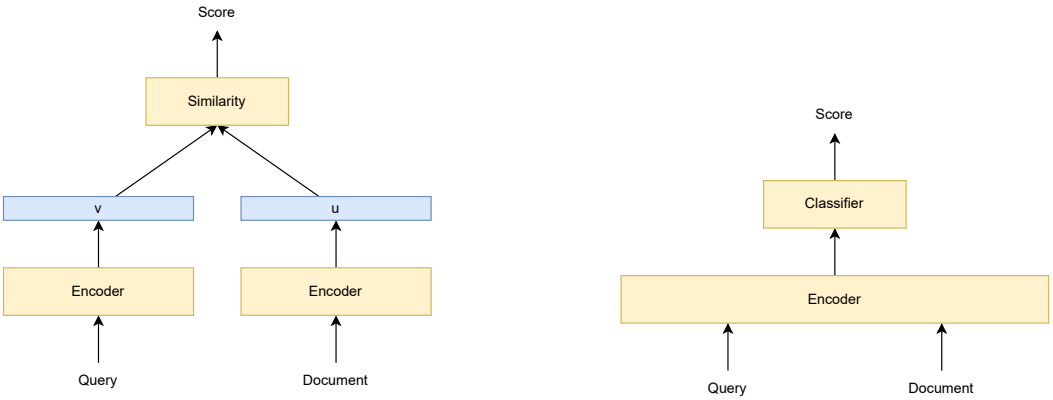
Two Stage Retrieval To balance computational efficiency with retrieval accuracy, a common architectural pattern is the Retrieve & Re-rank pipeline [119]. This two-stage process involves:

1. Retrieval (First Stage): A highly efficient, scalable retriever (like BM25) is used to quickly scan the entire collection and retrieve a large set of candidate documents (e.g., the top-k candidates).

2. Re-ranking (Second Stage): A more powerful, computationally intensive model is then applied to this smaller set of candidates. This is often a cross-encoder [119] or a specialized architecture like ColBERT [84].

2.5.2 Foundational Architectures

The shift towards dense retrieval was enabled by deep learning, particularly by transformers [45, 174]. In the context of information retrieval, they are primarily used in two distinct architectural patterns: bi-encoders and cross-encoders (shown in Figure 2.13).



(a) Bi-Encoder (b) Cross-Encoder

Figure 2.13: A comparison of Bi-Encoder and Cross-Encoder. (a) The Bi-Encoder architecture uses two separate encoders (typically with shared weights) to generate dense embeddings for the query (v) and the document (u) independently. The relevance score is then calculated via a late-interaction step. (b) The Cross-Encoder receives the query and document as a single concatenated input to one encoder. This allows understanding the interactions between the query and document early.

Bi-Encoders A bi-encoder architecture uses two independent transformations to generate embeddings for the query and the document separately. Retrieval is performed by calculating a similarity score between the query embedding and the pre-computed embeddings of all documents in the corpus. Since document embeddings can be pre-computed and stored in an efficient index, bi-encoders can be used for the first-stage retrieval.

Cross-Encoders A cross-encoder processes the query and a document together. The model outputs a single score that directly represents the relevance of the document to the query. It can better capture the interactions between a query and a document, but this approach is computationally expensive, as it requires a full forward pass for every query-document pair. They are generally used as second-stage rerankers.

Cross-Modal Architectures The principles of dense retrieval can be extended beyond a single modality. Cross-modal retrieval aims to find items in one modality (e.g., images, video) using a query from another modality (e.g., text). This is achieved by creating a shared embedding space where different data types can be compared, typically using an encoder for each modality, such as in CLIP [130].

2.5.3 Evaluation Metrics

Evaluation is often focused on the top-k results, as users are most likely to interact with the first page of results.

Some metrics, like Precision@K and Recall@K, evaluate the quality of the top-k set of retrieved documents without considering their internal order. Normalized Discounted Cumulative Gain@K (nDCG@K) instead can handle multiple levels of relevance (e.g., "perfect," "good," "fair") and its use of a logarithmic discount factor emphasizes the importance of the very top ranks:

$$nDCG@k = \frac{DCG@k}{IDCG@k}$$

$$DCG@k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$$

where the Ideal DCG ($IDCG@k$) is the $DCG@k$ for a perfectly ranked list and rel_i is the relevance scores of the i-th document in the ranking. The logarithm in $DCG@k$ penalizes the relevant items ranked low. An example is shown in Table 2.2. The ideal ranking for this query would be documents with relevance scores of $[3, 3, 2, 1, 0]$, which leads to $IDCG@5 = 7.129$, while the current ranking gives $DCG@5 = 6.149$. The ranking system obtained $nDCG@5 = 0.863$.

Rank (i)	rel_i	$1/\log_2(i+1)$	$rel_i/\log_2(i+1)$
1	3	1.000	3.000
2	2	0.631	1.262
3	3	0.500	1.500
4	0	0.431	0.000
5	1	0.387	0.387

Table 2.2: An example calculation of Discounted Cumulative Gain (DCG) at rank 5 (DCG@5).

With IR, we can efficiently surface relevant documents from vast collections. We now have access to two powerful but separate streams of information: a quantitative understanding from imagery, and a qualitative understanding from text. The next critical step is to fuse them.

2.6 Contrastive Learning

Manually creating labeled datasets to connect every image patch with a corresponding report is intractable. To bridge this modality gap, modern solutions employ Contrastive Learning (CL), a self-supervised (SSL) approach that learns to align data from different sources in a shared semantic space. SSL has emerged as a powerful paradigm for training deep learning models, particularly in domains where large-scale labeled datasets are scarce or expensive to create. SSL methods generate supervisory signals directly from the data, without relying on manual annotations or unsupervised approaches. Contrastive Learning (CL) is a well-known and highly effective approach in the SSL domain that learns representations by enforcing similarity between related data samples while simultaneously enforcing dissimilarity between unrelated ones. The fundamental principle is to "contrast" a positive pair against a set of negative pairs in the embedding space. This process encourages the model to learn robust and generalizable features that capture the underlying semantic structure of the data.

2.6.1 Core Mechanism

The objective of contrastive learning is to learn an encoder function, f_θ, that maps an input sample x to a feature vector, or embedding, $z = f_\theta(x)$. The objective is to organize the embedding space such that semantically similar samples are pulled closer together, while dissimilar samples are pushed farther apart, as shown in Figure 2.14. This is achieved through three key components: pair generation, shared embedding space, and contrastive loss.

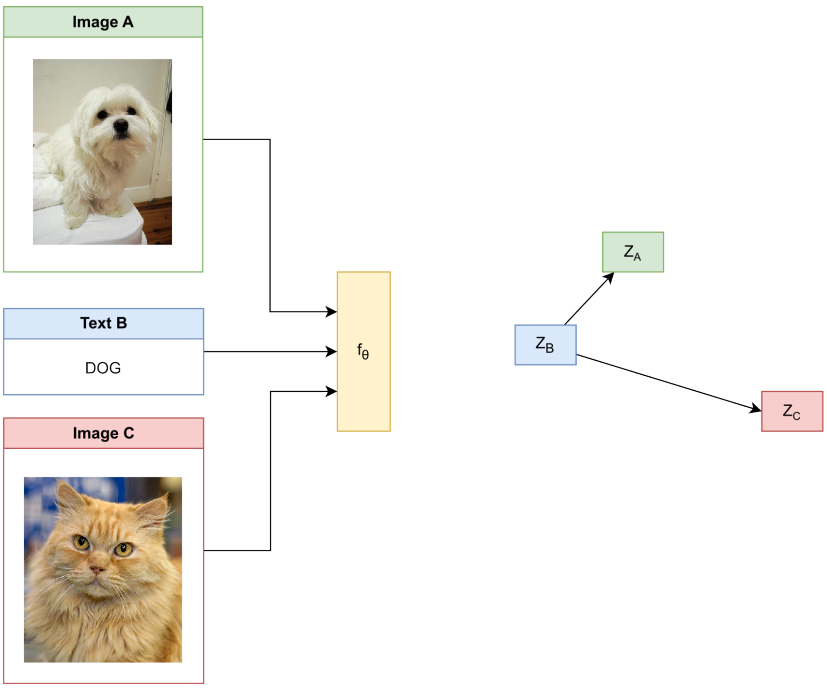


Figure 2.14: A function f_θ learns to map inputs into a common vector space. An image of a dog (Image A) and its corresponding text description (Text B) are encoded into nearby embeddings, z_A and z_B, respectively. The cat (Image C) is encoded into an embedding, z_C, that is distant from the others, because it is semantically distant.

Pair Generation. The first step is the creation of "positive" and "negative" pairs. A positive pair can be two mined similar samples, but also two augmented views of the same data sample. In a multi-modal context, a positive pair could be an image and its corresponding textual description. Negative pairs consist of samples that should be considered dissimilar, such as two entirely different images or an image paired with a random, incorrect text description. While more robust mined solutions should provide a stronger signal, even loosely selected samples proved effective [130].

Encoders and Shared Embedding Space. An encoder network is used to transform the input samples into embeddings. In multi-modal settings, separate encoders are used for each modality (e.g., an image encoder and a text encoder). The goal is to project these embeddings into a shared latent space where their proximity reflects their semantic relationship.

Contrastive Loss Function. The training is driven by a contrastive loss function. Early approaches introduced the Contrastive Loss [66], which operates on pairs, and the Triplet Loss [157], which uses an anchor, positive, and negative sample to enforce a margin-based separation in the embedding space. Now, the InfoNCE (Noise Contrastive Estimation) loss is one of the most widely used variants thanks to its efficiency. Given an "anchor" sample embedding z_i, its positive counterpart z_j, and a set of $N-1$ negative sample embeddings $\{z_k\}_{k\neq i,j}$, the loss aims to maximize the similarity of the positive pair relative to all negative pairs. The InfoNCE loss for a positive pair (i, j) is formulated as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{k\neq i} \exp(\text{sim}(z_i, z_k)/\tau)}$$

Here, $\text{sim}(u, v)$ is the cosine similarity between two vectors, and τ is a temperature hyperparameter that controls the sharpness of the distribution, influencing how well the model distinguishes between difficult negative samples.

2.6.2 Foundational Applications

Contrastive learning has been successfully applied to develop many foundation models across various domains.

Multi-modal Alignment: The Contrastive Language-Image Pre-training (CLIP) model [130] is a canonical example of multi-modal contrastive learning. It was trained on hundreds of millions of (image, text) pairs from the internet, as shown in Figure 2.15. By training separate image and text encoders to align their respective embeddings in a shared space, CLIP learns rich representations that enable zero-shot classification capabilities. However, these models are trained on generic web data, creating a significant domain gap when applied to our problems. The visual concepts and vocabulary in ecology and crisis response are highly specialized; for example, the textual concept of 'high soil moisture content' or 'post-fire burn scar' corresponds to specific multi-spectral signatures in a satellite image that are absent from web photos.

Self-Supervised Visual Learning: In the unimodal visual domain, contrastive methods such as SimCLR [35] and MoCo [68] have been used to learn powerful features from images alone. The core intuition is to treat two different augmented views of the same image as a "positive pair", while views of all other images are treated as "negative pairs". SimCLR relies on extremely large in-memory batches, using all other samples in a given batch as negatives. In contrast, MoCo introduced a more memory-efficient approach by maintaining a dynamic queue of negatives from preceding batches, which are encoded by a slowly evolving momentum encoder.

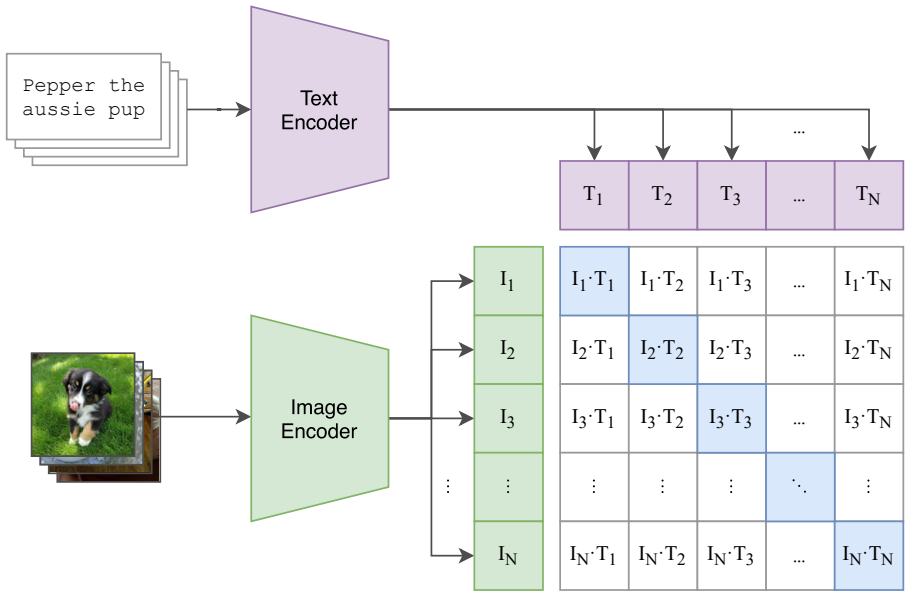


Figure 2.15: Schematic adapted from CLIP [130]. During training, a batch of N image-text pairs is processed. An image encoder and a text encoder independently map the raw inputs into feature embeddings (I_i and T_i, respectively) within a common vector space. An N×N matrix is then constructed, where each element is the cosine similarity between an image embedding I_i and a text embedding T_j. The training objective aims to maximize the similarity of the diagonal positive pairs and minimize the similarity of the off-diagonal negative pairs.

Contrastive learning allows us to compare and combine visual and textual information using a single state representation. Now, we can move from passive understanding to active decision-making.

2.7 Deep Q-Networks

With the ability to construct a rich state representation of an environment, we can try to address the question: 'Given the current situation, what is the best course of action?' This is the domain of Reinforcement Learning, where agents learn optimal decision-making policies through interaction. Deep Q-Networks (DQN) provide a foundational algorithm for this task. DQNs are foundational architectures in the field of Deep Reinforcement Learning (DRL). DRL combines deep neural networks with the decision-making framework of reinforcement learning. The DQN algorithm demonstrated the ability to learn complex control policies directly from high-dimensional sensory inputs, famously achieving human-level performance across a suite of Atari 2600 games [114]. This success established a powerful paradigm for training agents to solve a wide variety of sequential decision-making tasks, with

applications ranging from robotics and game playing to resource management.

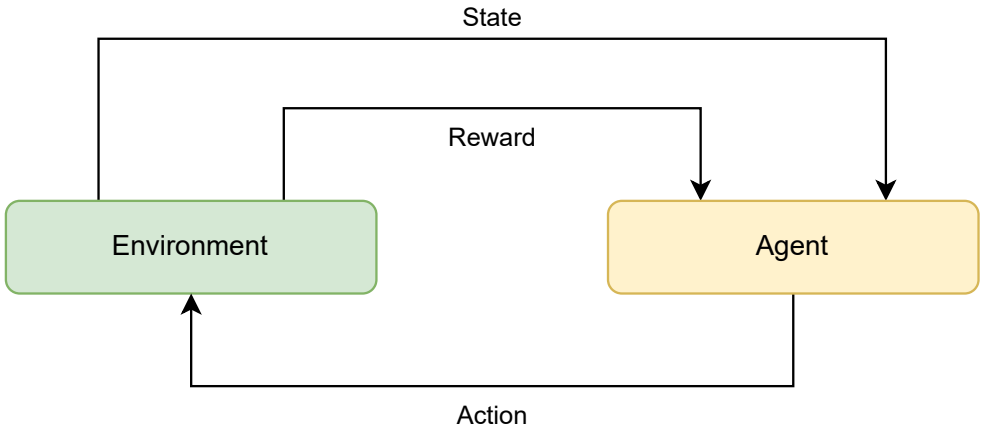


Figure 2.16: The feedback loop of a reinforcement learning system. The agent interacts with the environment by performing actions. The environment, in return, provides the agent with new states and reward signals, which the agent uses to plan its future actions.

2.7.1 Core Principles and Architecture

Reinforcement Learning (RL) involves an agent interacting with an environment over a sequence of discrete time steps. At each step t, the agent observes the environment's state $s_t \in S$, selects an action $a_t \in A$, and receives a scalar reward r_t and the next state s_{t+1}. The agent's goal is to learn a policy, π, that maximizes the cumulative discounted future reward, known as the return. The general loop in RL is shown in Figure 2.16.

The central component of a DQN is the action-value function or Q-function $Q(s, a)$. This function estimates the expected return for taking action a in state s and following the optimal policy thereafter. The optimal Q-function, $Q^*(s, a)$, adheres to the Bellman equation:

$$Q^*(s, a) = \mathbb{E}_{s' \sim \mathcal{E}} \left[r + \gamma \max_{a'} Q^*(s', a') | s, a \right]$$

where r is the reward received after taking action a in state s, s' is the resulting state, and $\gamma \in [0, 1]$ is a discount factor that balances the importance of immediate versus future rewards.

The key innovation of DQN is to approximate $Q^*(s, a)$ using a deep neural network (Q-network), with parameters θ. This network takes a state representation s as input and outputs a vector of Q-values, one for each possible action: $Q(s, \cdot; \theta) \approx$

$Q^*(s, \cdot)$. While the original work employed Convolutional Neural Networks (CNNs), the specific architecture of the Q-network is domain-dependent.

2.7.2 Training Mechanisms and Loss Function

Training the Q-network involves minimizing a sequence of loss functions, $L_i(\theta_i)$, that are updated at each iteration i. The objective is to make the predicted Q-values from the network align with the target values derived from the Bellman equation. The loss is typically a Mean Squared Error (MSE) between the predicted Q-value and a target value y:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')}\left[(y_i - Q(s, a; \theta_i))^2\right]$$

$$y_i = r + \gamma \max_{a'} Q(s', a'; \theta_{i-1})$$

To stabilize the learning process and improve data efficiency, DQN incorporates:

1. Experience Replay: Instead of training on consecutive samples as they are generated, the agent's experiences (s_t, a_t, r_t, s_{t+1}) are stored in a large replay buffer D. During training, mini-batches of experiences are sampled from this buffer. This technique breaks the strong temporal correlations inherent in sequential observations, reducing the variance of the updates.

2. Fixed Target Network: The standard Q-learning update can be unstable, because the same network is used to both estimate the current Q-value and the target value. To mitigate this, a separate target network with parameters θ^- is used to calculate the target y. The target network's parameters are updated every given number of steps using the main Q-network's parameters. This adds a delay between the time a Q-value is updated and the time its change affects the target values. The loss function using a target network is:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')}\left[\left(\left(r + \gamma \max_{a'} Q(s', a'; \theta_i^-)\right) - Q(s, a; \theta_i)\right)^2\right]$$

The full DQN pipeline is shown in Figure 2.17.

DRL offers a powerful framework for sequential decisions in complex environments. However, many tasks discussed before, as well as the policies of DQNs, are embedded in deep neural networks, making their reasoning opaque to humans. This 'black box' nature presents a significant barrier to trust in many impactful applications.

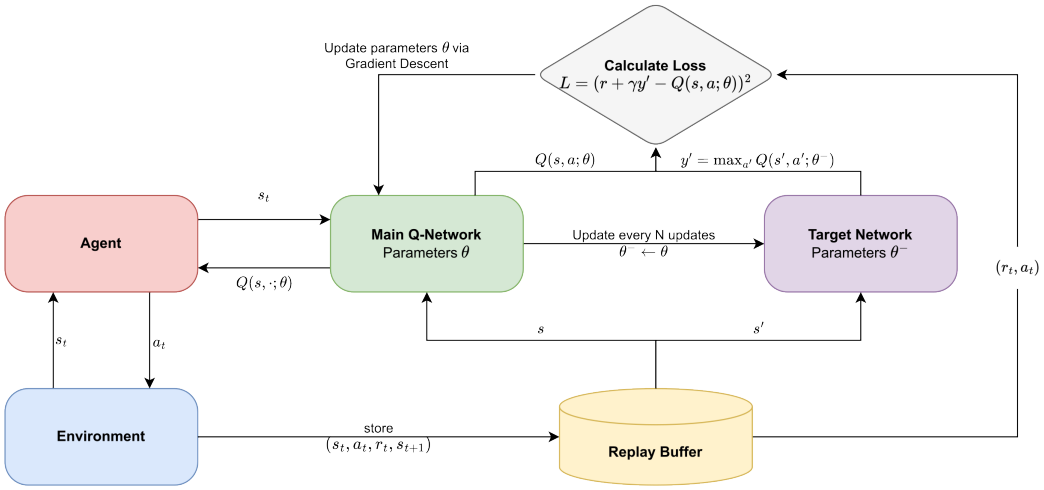


Figure 2.17: A Deep Q-Network with experience replay and a target network. The agent's interactions with the environment produce experience tuples (s_t, a_t, r_t, s_{t+1}), which are stored in a replay buffer. During training, minibatches are sampled from this buffer. The main Q-network, parameterized by θ, is used to select actions and to estimate the Q-value for the current state-action pair. A separate, periodically updated target network, parameterized by θ^-, is used to generate a stable target value for the loss function calculation.

2.8 The "Black Box" Problem

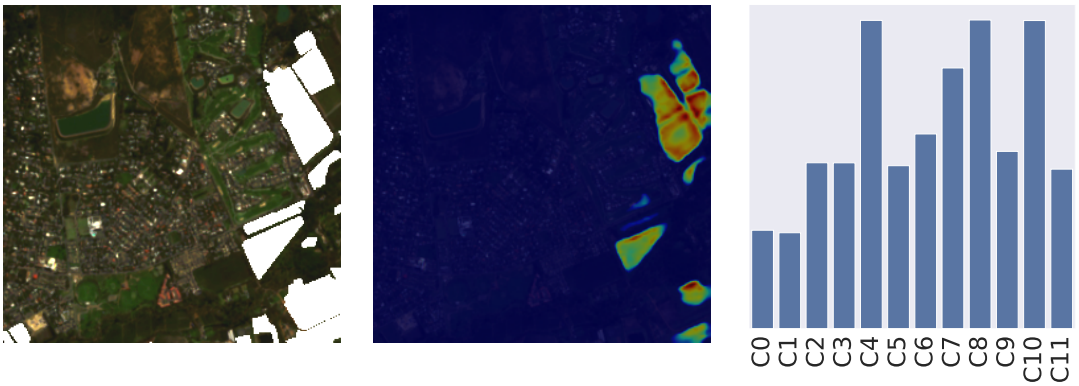
The deep learning models discussed before achieve remarkable performance at the cost of interpretability. When a model's prediction informs a decision to evacuate a town or allocate scarce resources, 'because the model said so' is an insufficient justification. Deep learning models learn intricate hierarchical features directly from vast amounts of data; however, this comes at the cost of lower interpretability. The complex, multi-layered, and non-linear structure of these models makes their internal decision-making process opaque to human observers. This phenomenon is widely known as the "black box" problem [**castelvecchi**]. The inability to understand why a model made a specific prediction is a significant barrier to its adoption in domains such as medical diagnostics, autonomous navigation, and Earth observation, where accountability, trust, and the verification of the model's reasoning are paramount [61]. Addressing this challenge has given rise to the field of Explainable AI (XAI), which seeks to develop methods to understand models.

2.8.1 Post-Hoc Saliency Methods

Saliency methods generate heatmaps that visualize the importance of different input features for a given prediction, as shown in Figure 2.18.

Grad-CAM One well-known technique is the Gradient-weighted Class Activation Mapping (Grad-CAM) [161]. It provides a visual explanation by producing a localization map that highlights the important regions in an image for a specific decision. This is achieved by using the gradients of the target class as weights for the feature maps of the final convolutional layer.

Perturbation-based Methods (Per-Channel Relevance) Perturbation-based methods systematically alter parts of the input and measure the impact on the model's output. A specific adaptation for multi-modal or multi-spectral data is per-channel relevance [23]. In this approach, entire input channels (e.g., specific spectral bands) are occluded and the resulting change in the model's prediction is measured. A significant change indicates that the occluded channel is highly relevant to the model's decision-making process.



(a) Image and Mask (b) Saliency Map (c) Relevance

Figure 2.18: An illustration of a post-hoc method on a satellite image. (a) The input image with the segmentation mask. (b) The corresponding saliency map, where brighter regions indicate pixels that are more influential for the model's prediction. (c) The relevance of each of the 12 multispectral input channels (C0-C11) for the output.

2.8.2 Evaluating the Quality of Explanations

The explanation can be quantitatively assessed in different ways. In this case, we present common solutions for semantic segmentation.

Plausibility This metric evaluates the degree to which an explanation aligns with human intuition or domain knowledge [47, 152, 81]. For segmentation tasks, it can be measured by comparing the generated saliency map with the ground-truth mask,

using standard vision metrics like Intersection-over-Union (IoU) to quantify their spatial agreement.

Faithfulness This metric evaluates how accurately an explanation reflects the model's internal reasoning process. Unlike plausibility, which compares the explanation to human knowledge, faithfulness assesses if the explanation truly represents the features the model found important. A common way to measure faithfulness is with the sufficiency.

Sufficiency The sufficiency assesses if the features highlighted by an explanation are sufficient for the model to make its prediction [81]. This is tested by using the explanation to mask the input image, keeping only the most important pixels, to re-evaluate the model's performance. A small drop in performance suggests that the explanation has captured the features that are truly sufficient for the model to make its prediction.

2.8.3 Model-Agnostic Performance Diagnostics

Beyond explaining individual predictions, XAI techniques can also be used to diagnose a model's systematic behavior, identifying consistent failure modes that may be hidden by aggregate performance metrics.

Subgroup Discovery This technique identifies interpretable subgroups of data where a model's performance significantly diverges from the average [122]. A subgroup is defined by a set of descriptive, human-understandable attributes (e.g., samples with "low soil moisture" and "high precipitation variability"). By finding subgroups with high performance divergence, one can uncover the specific conditions under which a model systematically fails or excels.

Feature Attribution with Shapley Values Once underperforming subgroups are identified, feature attribution methods can explain which features are most responsible for the performance deviation. Shapley values [162], a concept from cooperative game theory, provide a way to distribute the contribution of each feature to an outcome. Global Shapley values [122] extend this to quantify the overall importance of each feature attribute across all identified subgroups, revealing the key drivers behind the model's systematic performance variations.

These post-hoc methods allow us to better understand a trained model's reasoning. While valuable, this approach tries to explain an existing opaque system. This raises the question: can we instead design models that are more transparent from the beginning?

2.9 Kolmogorov-Arnold Networks

While post-hoc XAI methods provide valuable insights, a parallel line of research explores architectures that are inherently more interpretable. Kolmogorov-Arnold Networks (KANs) are a class of neural network architectures that have recently emerged as an alternative to Multi-Layer Perceptrons (MLPs) [102]. Their design is inspired by the Kolmogorov-Arnold representation theorem, which states that any continuous multivariate function can be represented as a finite composition of continuous univariate functions [87, 9].

The differences between KANs and MLPs are summarized in Figure 2.19. The main difference is the location and nature of their non-linear activation functions. In an MLP, learnable parameters are located on the network's edges, representing linear transformations, while fixed, non-linear activation functions (e.g., ReLU, Sigmoid) are applied at the nodes. In contrast, KANs place learnable, univariate activation functions on the edges of the network, while the nodes simply perform summation. By learning the activation functions themselves, KANs can adapt to the underlying structure of the data, potentially capturing more complex patterns than MLPs [102]. KAN has shown promising results in tasks like solving partial differential equations and fitting physical data.

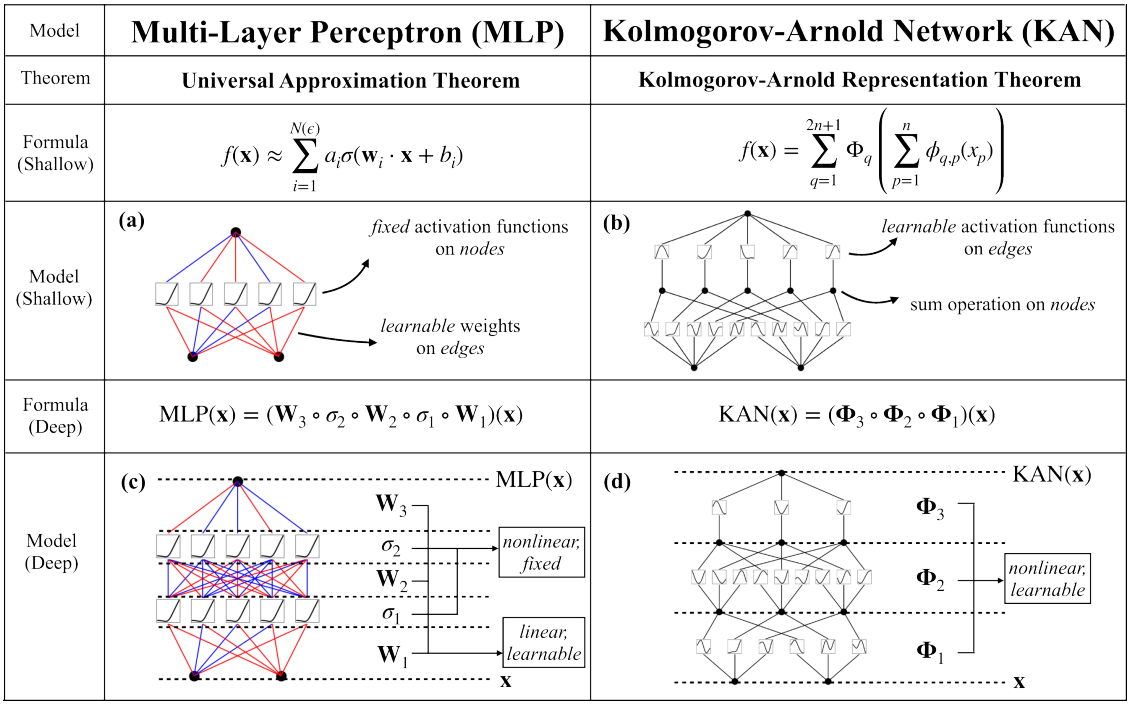


Figure 2.19: Comparison of a standard MLP layer and a KAN layer. Adapted from [102].

2.9.1 Architectural Principles and Components

A KAN is composed of stacked layers, much like an MLP. However, the computation within each layer is distinct. For a KAN layer with an input dimension of n_{in} and an output dimension of n_{out}, the relationship between the input vector $\mathbf{x} = (x_1, \ldots, x_{n_{in}})$ and the pre-activation output for the j-th neuron, z_j, is given by:

$$z_j = \sum_{i=1}^{n_{in}} \phi_{j,i}(x_i) \tag{2.1}$$

Here, each $\phi_{j,i}$ is a learnable univariate function applied to the i-th input component, x_i. This function is typically parameterized as a B-spline, which is a piecewise polynomial function defined by a set of learnable coefficients over a grid of control points [22]. During training, the network learns the optimal coefficients for each B-spline, effectively shaping each activation function $\phi_{j,i}$ to best fit the input-output relationship. An optional fixed activation function, g_j, can then be applied at the output node with summation.

A strength of this architecture is its inherent interpretability. Since each $\phi_{j,i}$ is a simple 1D function, it can be directly visualized by plotting its output against its input. It is possible to inspect exactly how the network is transforming each individual feature at every stage, offering a high level of transparency. Due to their structure, KAN layers can serve as a drop-in replacement for linear layers in various neural networks or can be used in combinations with other well-established components like convolutions.

By learning custom activation functions, KANs present a potential pathway towards models that are powerful but also transparent. This last section completes the foundational theory, discussing all aspects from data to trusted, actionable intelligence.

Chapter 3

Machine Learning in Ecology

In an era of strong environmental change, the preservation of Earth's natural ecosystems requires monitoring capabilities both broad in scale and granular in detail. Machine learning, powered by the great availability of remote sensing data, provides the essential toolkit for this task. This chapter presents the novel contributions designed to address some critical challenges in ecological monitoring. We will demonstrate how new data sources, innovative model architectures, and new learning paradigms provide more accurate, efficient, and interpretable insights into the planet's health status. The chapter is structured in four topics: **agricultural** monitoring, **forestry** analysis, **hydrological** forecasting, and multi-modal **land use land cover** retrieval.

3.1 Forestry

Accurate and scalable measurement of forest canopy height is fundamental to monitoring global ecosystem health [180]. Traditional methodologies, such as in-situ measurements or airborne laser scanning (ALS), provide high accuracy but are prohibitively expensive and slow for large-scale applications [8, 60, 170]. While machine learning models utilizing satellite data offer a more scalable alternative, they can be used to produce maps at a low spatial resolution (e.g, using GEDI or Sentinel) [12, 52, 89, 123]. However, higher resolution is needed to map fine-grained details like single trees.

Recent advancements have increased the resolutions of the models using the self-supervised pretraining of large Vision Transformer (ViT) models on millions of in-domain, high-resolution satellite images, which are then fine-tuned for the specific task of canopy height estimation [169]. While effective, it involves high computational costs and relies on massive datasets, creating a significant barrier to accessibility for a broader public. This highlights a critical research gap: the need for an efficient, low-cost, but highly accurate method for generating high-resolution canopy height maps.

Depth Any Canopy: Leveraging Depth Foundation Models for Canopy Height Estimation [136] addresses this gap by exploring the potential of cross-domain transfer learning from general-purpose foundation models. While the pretraining nature is different, we hypothesize that a monocular depth estimation model for natural images can be efficiently adapted for the task of canopy height estimation from single-view aerial imagery. To this end, we introduce Depth Any Canopy (DAC), a novel model created by fine-tuning the Depth Anything v2 foundation model. Our work demonstrates that with minimal resources, it is possible to achieve performance comparable to or superior to expensive baselines, while being more accessible. Additionally, it shows superior performance in capturing complex vegetation structures, highlighting its potential as a scalable solution for global forest monitoring.

3.1.1 Methodology

Our primary contribution is a new model derived from an existing foundational one. We detail its architecture and the specific training procedure used for adaptation.

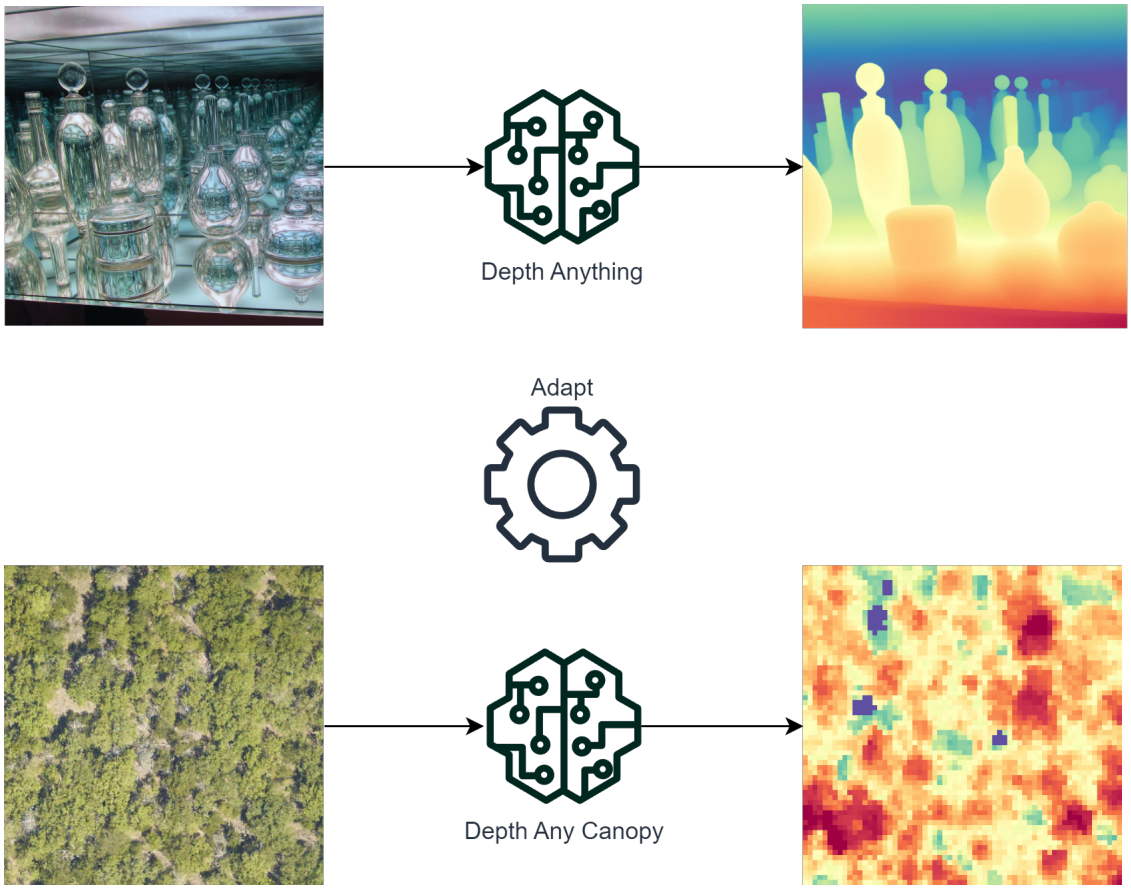


Figure 3.1: **From Depth Anything [195] to Depth Any Canopy**. Using natural imagery as training data, Depth Anything is a monocular depth estimate foundation model. In order to create Depth Any Canopy (DAC), we refine and modify Depth Anything v2 for the purpose of assessing tree canopy height in remote sensing data.

Task

The task of canopy height estimation from a single RGB image can be framed as pixel-wise regression. Given an input image I of shape $H \times W \times 3$, where H and W are the height and width, the model has to predict a canopy height map (CHM) M of shape $H \times W$. Each pixel $M_{i,j}$ in the output map represents the estimated

height of the tree canopy at the corresponding spatial location (i, j) in the input image. The ground level is considered to have a height of zero, so the task is to measure the vertical distance from the ground to the top of the canopy for every pixel. The task can be seen in Figure 3.1

Model Architecture

The core of Depth Any Canopy is the Depth Anything v2 model [195]. This model's architecture is based on the well-known encoder-decoder principle. The encoder is a Vision Transformer (ViT) [49] pre-trained with the DINOv2 self-supervised learning method [120]. DINOv2 is renowned for learning robust visual features that generalize well across various domains without supervision. The decoder follows the Dense Prediction Transformer (DPT) design [132], which is designed to reconstruct dense, pixel-wise outputs (like a depth map) from the transformer's encoding. The original model was trained in a student-teacher framework on over 62 million images, learning to predict relative depth from a vast corpus of unlabeled natural images. We fine-tune the model to adapt to remotely sensed data, which generally shows a top-to-bottom perspective, which is less common in natural imagery.

3.1.2 Experimental Setup

In this section, we detail the datasets, the baselines, and the metrics for the task.

Datasets

For training, we utilize the EarthView dataset [175], focusing on its NEON subset, which contains very high-resolution (0.1m) RGB aerial imagery paired with 1m resolution CHMs derived from ALS. We filtered this dataset with the Q-Align vision-language model [185] to score each image for artifacts like motion blur and warping as shown in Figure 3.2. We kept only samples with a quality score above 2.5, resulting in a clean dataset of 45,781 samples. For evaluation, we used the held-out test set from the filtered EarthView dataset and the High-Resolution Canopy Height Maps (HRCHM) dataset [169].

Baselines and Comparative Methods

Our primary baseline for comparison was the state-of-the-art model from Tolan et al. [169], which we will call SSL-H. This model is a ViT-Huge that was pre-trained using the DINOv2 method on a massive dataset of 18 million WorldView satellite images before being fine-tuned on the HRCHM dataset. We also compared our fine-tuned DAC models against the zero-shot performance of the original Depth

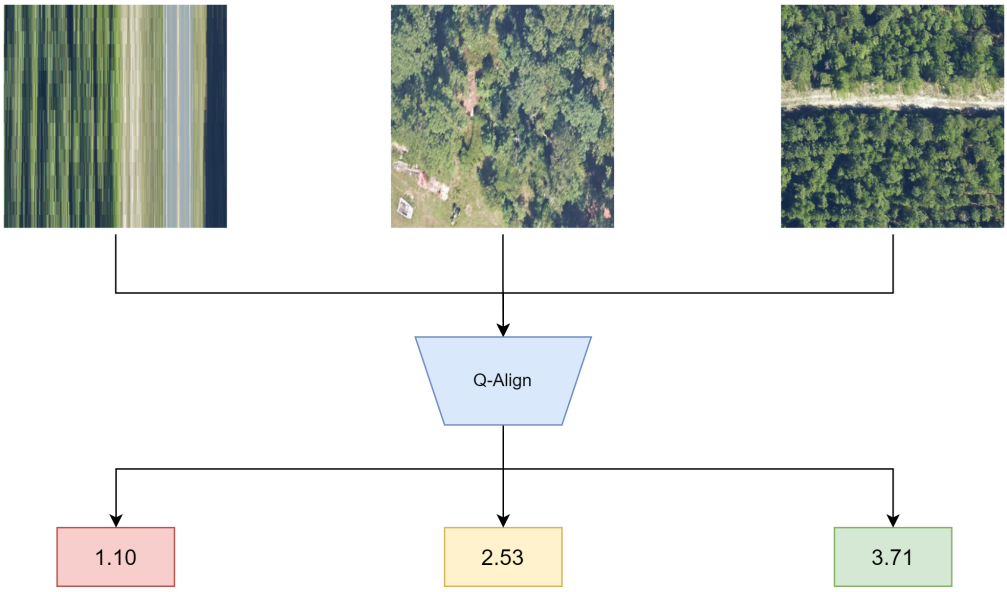


Figure 3.2: Examples of Q-Align quality scores on NEON RGB images. A noisy sample with a score of 1.10 on the left, a medium-quality sample with a score of 2.53, and a higher-quality sample received a score of 3.71 on the right. Samples affected by warping and motion blurs can be found using this technique.

Anything v2 small and base models (DA-S and DA-B) to quantify the improvement gained from our fine-tuning procedure.

Evaluation Metrics

We measured performance using three metrics for a more complete overview:

1. Mean Absolute Error (MAE): The L1 distance between the predicted and ground truth canopy heights, measuring the absolute accuracy of the regression.

2. Intersection-Over-Union (IoU): To evaluate the model's ability to identify tree canopy extent, both predicted and ground truth CHMs were converted to binary masks using a minimal height threshold (1×10^{-4}).

3. Pearson Correlation (PC): Calculated on the tree-covered areas (as defined by the ground truth mask) to assess whether the relative predicted heights are consistent with the ground truth, regardless of absolute error.

We also report the number of model parameters and GFLOPs to compare computational efficiency.

3.1.3 Results and Discussion

Quantitative Analysis

As seen in Table 3.1, Depth Any Canopy (DAC) is substantially more efficient and performs better or on par with the state-of-the-art SSL-H baseline across both assessment datasets. While the PC is often greater on HRCHM for SSL-H, the fine-tuning of Depth Anything v2 on EarthView yields good comparative results and consistent performance under both MAE and IoU criteria. When it comes to MAE and PC on HRCHM, SSL-H performs best, although its IoU is lower than DA's. In comparison to HRCHM, EarthView's performance is lower, especially on MAE, where it drops by a factor of 10. Additionally, we can use smaller models to achieve good performance.

The DAC-S model contains only 24.8M parameters and requires 115 GFLOPs, while SSL-H has 677M parameters and 414 GFLOPs. This represents a 27 times reduction in model size and a nearly 4 times reduction in computational cost. The zero-shot performance of Depth Anything was worse than the fine-tuned versions, confirming the necessity of the adaptation step.

Model	FT	# Params	GFLOPs	EarthView[175]			HRCHM[169]		
				MAE ↓	IoU ↑	PC ↑	MAE ↓	IoU ↑	PC ↑
SSL-H[169]	N	677M	414	0.2236	0.4164	0.1544	**0.0306**	0.485	**0.7441**
DA-S[195]	N	**24.8M**	**115**	0.4116	0.4164	0.2892	0.5960	**0.6474**	0.1791
DA-B[195]	N	97.5M	381	0.4607	0.4164	**0.361**	0.5972	**0.6474**	0.1692
DAC-S[195]	Y	**24.8M**	**115**	<u>0.1410</u>	<u>0.5323</u>	0.2740	<u>0.1025</u>	0.5672	0.6102
DAC-B[195]	Y	97.5M	381	**0.1304**	**0.5926**	<u>0.3483</u>	0.1203	0.5494	<u>0.6171</u>

Table 3.1: **Results on EarthView and HRCHM datasets.** The best for each metric is bolded, while the second-best is underlined. *FT* indicates if the model is fine-tuned on EarthView. *DA* and *DAC* refers to Depth Anything v2 and Depth Any Canopy, while *DAC-S* and *DAC-B* refers to the ViT-S and ViT-B variants, respectively.

Qualitative Analysis

Qualitative examples are provided in Figures 3.3 and 3.4. The SSL-H baseline, while effective on its native HRCHM dataset, often produces overly smooth predictions and fails to capture fine-grained details on the EarthView dataset. In contrast, DAC generates CHMs with a higher level of detail, capturing complex forest structures and correctly identifying smaller or sparser vegetation. This robustness could be attributed to the rich features learned by Depth Anything, which was trained on a diverse range of complex visual scenes.

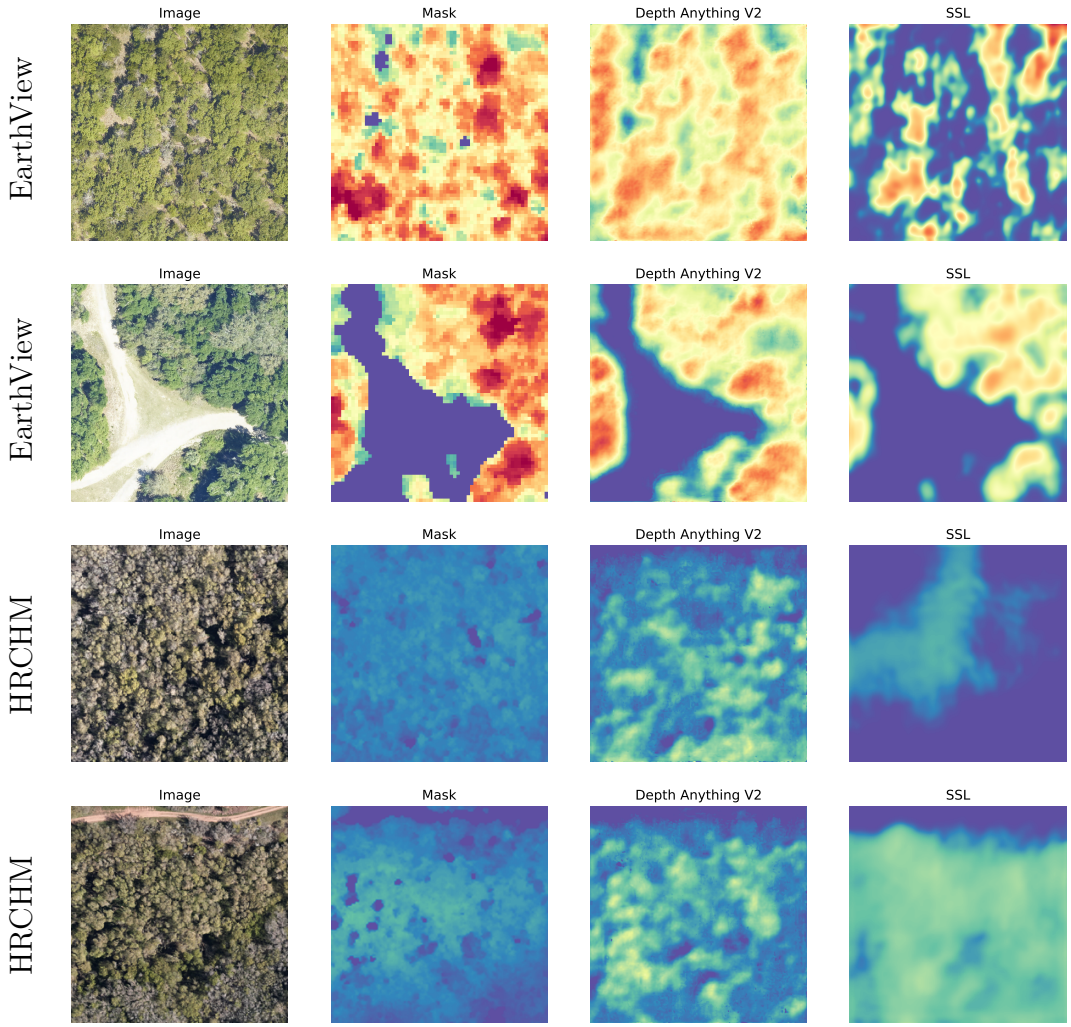
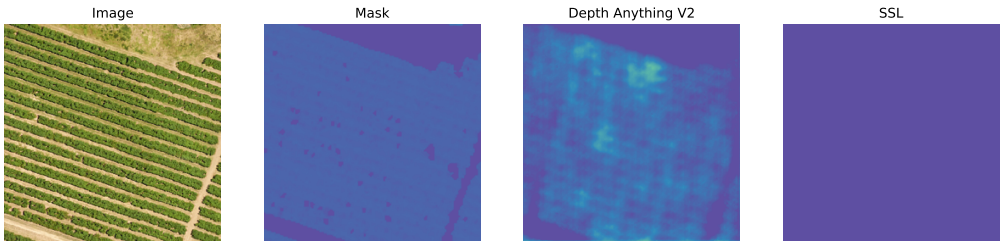


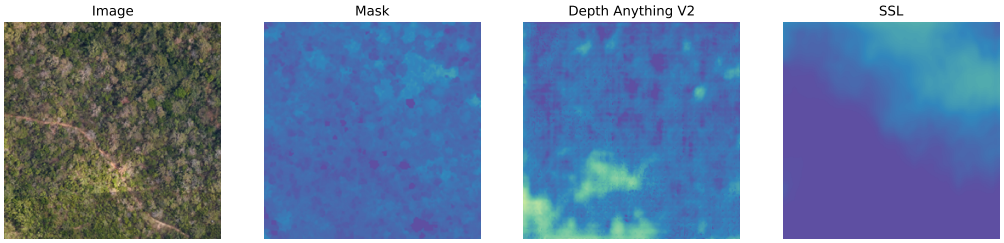
Figure 3.3: **Example predictions of Depth Any Canopy (DAC-S) and SSL-Huge model from Tolan et al. [169]** on NEON imagery from the EarthView and HRCHM datasets. Left to right: NEON RGB Image, Ground Truth Canopy Height Map, DAC-S predicted CHM and SSL-H predicted CHM.

Ablation Studies

The comparison between the zero-shot Depth Anything (DA) models and their counterparts Depth Any Canopy (DAC) serves to understand if the fine-tuning process is necessary. In any case, we can improve the base model with the finetuning: for example, on the HRCHM dataset, the MAE for the ViT-S model dropped from 0.5960 in the zero-shot case to 0.1025 after fine-tuning. This suggests that while the foundation model provides a powerful starting point, the domain-specific fine-tuning is essential for achieving state-of-the-art results in canopy height estimation.



(a) Example 1



(b) Example 2

Figure 3.4: Examples of SSL-H constraints that Depth Any Canopy addresses. The SSL-H model by Tolan et al. [169] tends to produce predictions that are either too smooth or that predict zero height for tiny vegetation, according to our analysis. Depth Any Canopy can recover vegetation heights for intricate sceneries and edge scenarios because it is optimized using the Depth Anything v2 weights.

Discussion

The results demonstrate that by fine-tuning a general-purpose monocular depth model, we can surpass a much larger, state-of-the-art model that was pre-trained on 18 million in-domain satellite images. This finding suggests how we can develop high-performance remote sensing models without large resources. This approach enables achieving competitive results without access to massive computational resources or proprietary large-scale datasets. The entire fine-tuning process for DAC-S required less than two hours on a single GPU, with an estimated cost of less than $1.30 and a carbon footprint of only 0.14 $kgCO_2$, which is orders of magnitude less than the pre-training cost.

3.1.4 Section Summary

In this section, we introduced Depth Any Canopy (DAC), a novel and highly efficient model for forest canopy height estimation. By fine-tuning the Depth Anything v2 foundation model, DAC achieves performance superior to or comparable to the state-of-the-art while requiring only a fraction of the computational resources. This work provides a powerful example of a data-efficient method that makes top-tier results accessible without massive, proprietary datasets. However, this success

relies on the availability of a suitable pre-trained model. For many ecological tasks, such as forecasting long-term water dynamics, no such foundation exists. The following section addresses this fundamental bottleneck by tackling the challenge of creating a new foundational benchmark and the corresponding model to solve the task.

3.2 Hydrology

The accurate forecasting of surface water dynamics is an increasingly critical task to face climate change and water scarcity, which greatly impact ecosystems, agriculture, and energy production [151, 176]. While advancements in satellite remote sensing have improved our ability to monitor water bodies, the current state-of-the-art presents significant limitations for predictive modeling.

Satellite missions like Landsat [186] and Sentinel [104] have enabled the creation of large-scale global surface water datasets, primarily using spectral indices and machine learning for water body delineation [124, 193]. However, they are designed for analyzing past events rather than forecasting future dynamics. They are not explicitly structured for developing predictive models and often lack the integration of auxiliary drivers such as climate variables and topography. Similarly, while deep learning has been applied to time-series forecasting, its use in predicting future water dynamics, particularly by incorporating exogenous climatic factors, remains a largely unexplored area [16]. The absence of a comprehensive, multi-modal benchmark dataset curated specifically for developing surface water forecasting models, along with standardized predictive tasks, remains a critical gap in the literature.

To address this gap, *HydroChronos: Forecasting Decades of Surface Water Change* [28] introduces a novel dataset and a corresponding model baseline. We present HydroChronos, the first large-scale, multi-modal dataset tailored for surface water dynamics forecasting, integrating over three decades of satellite imagery with corresponding climate and elevation data. We define three standardized forecasting tasks to create a unified benchmark for future research: binary change detection, direction of change classification, and magnitude of change regression. As a baseline, we propose the AquaClimaTempo UNet (ACTU), a spatiotemporal deep learning model with a dedicated branch for incorporating climatic drivers. The efficacy of this model is enhanced by the integration of a multi-component regression loss function. Additionally, its strengths and weaknesses are highlighted using an Explainable AI (XAI) analysis. This approach provides the necessary insights to improve future research.

3.2.1 Methodology

In this section, we detail the data acquisition from multiple sources and the description of a novel deep learning architecture designed for water dynamics forecasting.

Data Acquisition and Sources

The raw data for the HydroChronos dataset were sourced from three different modalities.

- Satellite Imagery: To capture both long-term historical changes and recent dynamics, we utilized imagery from two satellite missions. Landsat-5 [188] Thematic Mapper (TM) data covers the period from 1990 to 2010, providing a historical perspective. Sentinel-2 [104] MultiSpectral Instrument (MSI) data covers the period from 2015 to 2024, offering better quality. We selected Top-Of-Atmosphere (TOA) images with minimal cloud coverage, focusing on the May-August period in the Northern Hemisphere to ensure comparable conditions.

- Climate Data: Time series of monthly climate variables were sourced from the TERRACLIMATE [1] dataset. This global dataset provides 14 variables at a resolution of approximately 4.6 km, including temperature, precipitation, soil moisture, and evapotranspiration.

- Topographic Data: A static Digital Elevation Model (DEM) was sourced from the Copernicus GLO30 DEM [4] dataset, which provides global coverage at a spatial resolution of approximately 30 meters.

The geographical locations for the dataset cover regions in Europe, the United States, and Brazil. It includes a diverse set of lakes and rivers selected from the HydroLAKES [111] and HydroRIVERS [92] databases as shown in Figure 3.5.

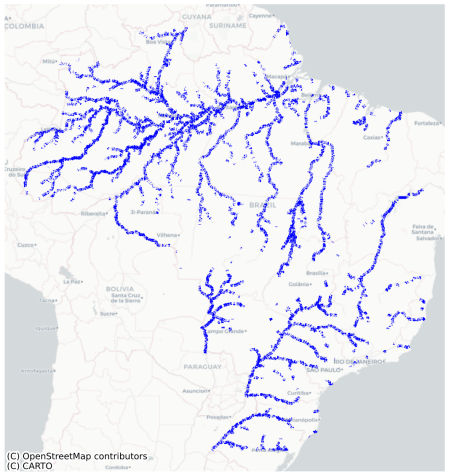
To ensure consistency between Landsat-5 and Sentinel-2 sensors, we selected six comparable spectral bands (Blue, Green, Red, NIR, and two SWIR bands) as shown in Table 3.2. All imagery was harmonized and projected to WGS84 at a uniform spatial resolution of 30m.

Annotation

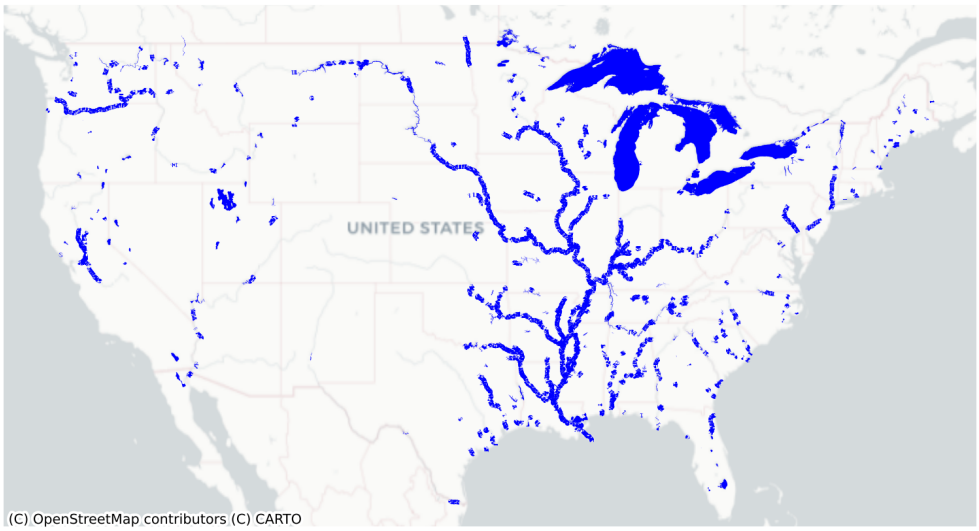
Given the difficulty of obtaining manually annotated ground truth for water dynamics at this scale, we adopted a semi-automated approach based on a well-established spectral index. The Modified Normalized Difference Water Index (MNDWI) [191] was calculated for each image as

$$MNDWI = (Green - SWIR)/(Green + SWIR)$$

(a) Europe



(b) Brazil



(c) United States

Figure 3.5: Distribution of lakes and rivers in HydroChronos

To create a robust target variable that represents a trend rather than a noisy state, we defined the target T as the pixel-wise median difference between a future time series of MNDWI images and a past time series: $T = \text{median}(P) - \text{median}(F)$. This approach smooths minor changes and sensor artifacts, focusing on more persistent hydrological shifts. To handle data corruption from clouds, we applied cloud masking to all images to exclude invalid areas from the analysis.

Landsat	Sentinel	Description	Central Wavelength (L/S)
B1	B2	Blue	485/492 nm
B2	B3	Green	560/560 nm
B3	B4	Red	660/665 nm
B4	B8	NIR	830/833 nm
B5	B11	SWIR	1650/1610 nm
B7	B12	SWIR	2220/2190 nm

Table 3.2: Landsat (L) and Sentinel (S) coupled bands included in the dataset. *NIR* is Near InfraRed and SWIR is *Short-Wave InfraRed*

Dataset Statistics

The final HydroChronos dataset consists of approximately 1900 time series for testing and 16,000 for training, for over 100,000 individual samples. The dataset is split temporally. The Landsat-5 data (1990-2010) is designated for pre-training (given the lower sensor quality), allowing the model to learn long-term dynamics. The more recent and higher-quality Sentinel-2 data (2015-2024) is used for fine-tuning and testing, with data from Brazil used for fine-tuning and data from Europe and the USA reserved for the final test set. Each sample contains the image time series (6 bands), the static DEM, and the corresponding monthly climate time series.

Tasks

The three tasks are formulated in the following way, and a visual example is shown in Figure 3.6:

1. Binary change detection: it can be framed as binary semantic segmentation. Given a timeseries P, a target T, and a threshold t to define what we consider a relevant change, we create a binary mask $M_c = |T| > t$. The task focuses on creating a model to predict M_c [28].

2. Direction classification: it can be framed as a multiclass semantic segmentation task with 3 classes: negative, positive, or no change. Given a timeseries P, a target T, and a threshold t, we create a mask M_d where a pixel m_d is assigned to the negative change class if $m_d < t$, to the positive change class if $m_d > t$, otherwise it is assigned to the no change class. The task focuses on creating a model to predict M_d [28].

3. Magnitude regression: the previous tasks assume the existence of a threshold t to define relevant changes. However, it can be of interest to model every "small" change in the area. Given a timeseries P and a target T, the task

focuses on creating a model to regress the values of $|T|$. This task can be framed as pixel-wise regression. Preliminary experiments also tried to address the regression of T, but with little success, so we reported only these settings as baseline, leaving this last task for future work [28].

(a) Regression (b) Change (c) Direction

Figure 3.6: Visual example of tasks for Lake Tahoe. In regression, the values range from 0 to 2 (blue to red). In change detection, labels are no-change (blue) and change (red). In direction classification, labels are negative change (blue), no-change (grey), and positive change (red).

AquaClimaTempo U-Net

The AquaClimaTempo UNet (ACTU) is a spatiotemporal forecasting architecture designed to predict surface water changes using satellite imagery, optionally integrating climate and elevation data. It extends the U-Net [145] with Convolutional LSTM (ConvLSTM) [164] layers to model temporal dynamics.

The model's workflow is as follows, and it is schematized in Figure 3.7:

- Image Feature Extraction: A pyramidal backbone (e.g., ConvNext [183]) processes a time series of satellite images to extract multi-scale spatial features for each timestep. Optionally, a static Digital Elevation Model (DEM) can be concatenated to the images as an additional input channel.
- Climate Data Integration: A parallel Climate Encoder processes time series of climate variables (e.g., temperature, precipitation). It generates multi-scale feature maps that are spatially aligned with the image features.
- Gated Fusion: A Gated Fusion mechanism dynamically balances the influence of the image and climate features. It computes a weighting mask to create a combined feature representation at each scale, allowing the model to adaptively prioritize information from each modality.
- Temporal Aggregation and Prediction: The fused, multi-scale features are processed by ConvLSTM layers, which aggregate the information across the

time dimension. Finally, a UNet decoder uses these features to generate the final prediction mask.

Regression Loss

To address the significant class imbalance in the task, we developed a composite regression loss function. The final loss, L_T, is a weighted sum of a multiscale loss (L_{MS}) and a wavelet loss (L_W). The Huber loss is used as the base regression loss for both components. The total loss is formulated as:

$$L_T = \alpha L_{MS} + (1 - \alpha) L_W$$

Multiscale Loss The multiscale loss, L_{MS}, computes the regression loss across several downscaled versions of the prediction and ground truth images. This penalizes the model for errors at different spatial scales. Given a base regression loss L, prediction P, ground truth T, and M different scale factors $S = \{s_0, ..., s_M\}$, the multiscale loss is defined as:

$$L_{MS}(P, T) = \frac{1}{M}(L(P, T) + \sum_{l=1}^{M} L(D_{s_l}(P), D_{s_l}(T)))$$

where D_x represents the downscaling operation with factor x.

Wavelet Loss The wavelet loss, L_W, leverages the Discrete Wavelet Transform (DWT) to decompose the prediction and target images into different frequency subbands. This allows the model to learn from both coarse (approximation coefficients Y_L) and fine-detailed (detail coefficients Y_H) structures in the data. The DWT is computationally efficient and captures errors across different scales and orientations. Given a regression loss L, the wavelet loss is defined as:

$$L_W(Y_L^P, Y_L^t, Y_H^P, Y_H^t) = \alpha L_L(Y_L^P, Y_L^t) + \sum_{i=1}^{N} w_i \cdot L_{H,i}(Y_{H,i}^P, Y_{H,i}^t)$$

where L_L is the loss on the low-frequency components and $L_{H,i}$ is the loss on the high-frequency components for N levels, weighted by α and w_i respectively.

3.2.2 Experimental Setup

In our experiments, the input of the models consists of a time series of 5 past images.

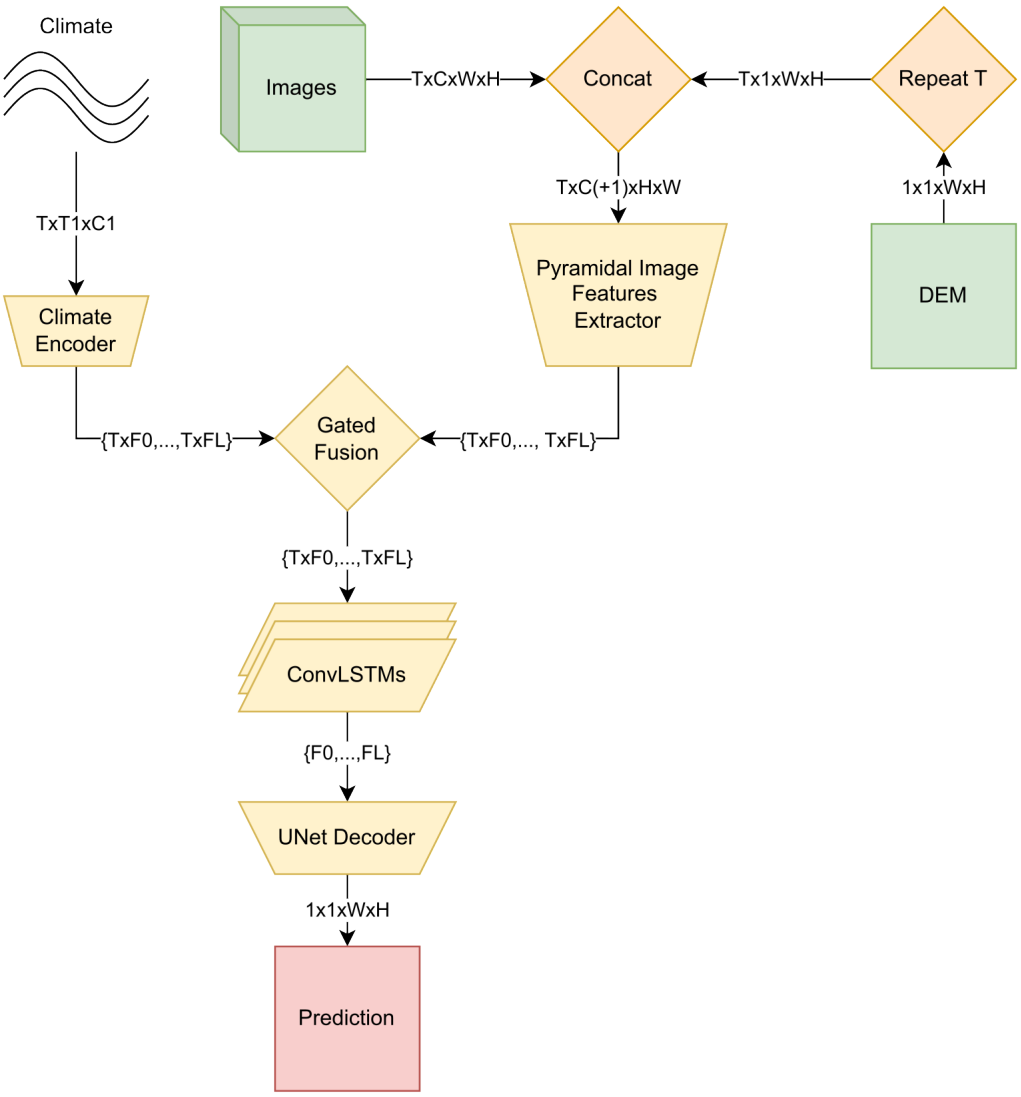


Figure 3.7: The architecture of AquaClimaTempo UNet (ACTU). In the case that DEM is supplied, it is concatenated along the channel axis and repeated once for each sample in the image timeseries. Multiscale embeddings are provided by the *Pyramidal Image Feature Extractor*. The *climate encoder* generates multiscale embeddings that are *gate fused* with the image embeddings if a climate timeseries is supplied. The *UNet decoder* uses the multiscale embeddings that *ConvLSTMs* give for the timeseries to produce the final prediction.

Baselines and Comparative Methods

We compared our proposed model, AquaClimaTempo UNet (ACTU), against two standard baselines:

- Constant Prediction: A naive baseline that predicts no change will occur between the past and future.
- Persistence Model: A more robust baseline that predicts the future change will be the same as the most recently observed change, calculated as the difference between the last known timestep and the median of the previous timesteps.

Evaluation Metrics

Performance was measured using a suite of metrics tailored to each task.

- Classification Tasks (Change Detection and Direction): We used Precision (P), Recall (R), and F1-score (F) for each class.
- Regression Task (Magnitude): We used Mean Absolute Error (MAE) and Pearson Correlation (PC). To account for the imbalance in the target values, we also computed MAE on the top-10% (MAE@10) and top-20% (MAE@20) of pixels with the highest magnitude changes. Furthermore, to evaluate the regression model's ability to perform as a classifier, we thresholded its output at $t = 0.1$ and $t = 0.2$ and computed precision (P@t), recall (R@t), and F1-score (F@t).

3.2.3 Results and Discussion

Our experiments demonstrate that the proposed ACTU model significantly outperforms the baselines across all tasks.

Quantitative Analysis

Across all three tasks—change detection, direction classification, and magnitude regression—all configurations of the AquaClimaTempo UNet (ACTU) significantly and substantially outperform the Constant and Persistence baselines. To quantify the contribution of each auxiliary input modality, we also perform an input-modality ablation study. In particular, while the satellite image time series is always provided, we toggle the availability of the climate time series (C) and the DEM (D). This directly measures the contribution of climate and topography information to each task under otherwise identical training and evaluation settings.

For change detection (Table 3.3), all ACTU models achieve F1-scores for the "change" class that are higher than the baselines. The inclusion of Digital Elevation Model (DEM) data provides the highest F1-score for detecting changes, while using both DEM and climate data provides the best recall.

In the more challenging direction classification task (Table 3.4), ACTU models excel at identifying "no change" but struggle to classify the direction of change

(positive or negative). Incorporating climate data improves the detection of positive changes.

For magnitude regression (Table 3.5), ACTU variants again demonstrate large improvements over the baselines across all metrics. While the base ACTU model excels at magnitude prediction, the larger ACTU-L model is the top-performing configuration for this task.

			No Change (NoCHG)			Change (CHG)		
Model	D	C	P	R	F	P	R	F
Constant	N	N	81.54	**100**	**89.25**	0	0	0
Persistence	N	N	88.73	41.64	54.07	23.86	**81.77**	34.98
ACTU	N	N	90.51*	82.78*	85.66*	44.87*	60.65*	48.79*
ACTU	N	Y	88.92*	85.86*	86.6*	45.45*	53.1*	45.83*
ACTU	Y	N	**90.57**	82.89*	85.75*	45.19*	61.01*	**49.38***
ACTU	Y	Y	90.53*	81.71*	85.08*	43.68*	62.33*	48.67*
ACTU-L	N	N	90.03°	84.3°	86.43°	**45.46°**	57.34°	47.94°
ACTU-L	Y	Y	89.2°	84.95°	86.37°	45.03°	54.39°	46.49°

Table 3.3: Results of change detection for models with optional use of climatic variables (C) and DEM (D), i.e., an input-modality ablation over auxiliary modalities. * denotes a statistically significant difference ($p < 0.01$) in persistence as determined by the t-test. When comparing ACTU-L with the same ACTU configuration, the statistical difference is shown by °.

			Negative Change (NEG)			No Change (NONE)			Positive Change (POS)		
Model	D	C	P	R	F	P	R	F	P	R	F
Constant	N	N	0	0	0	81.54	**100**	**89.25**	0	0	0
Persistence	N	N	14.94	47.37	17.48	88.73	41.64	54.07	9.25	**31.18**	11.61
ACTU	N	N	27.5*	**49.38***	**30.27***	**90.23***	84.58*	86.67*	27.73*	19.84*	19.47*
ACTU	N	Y	**29.84***	36.53*	27.75*	88.44	87.5*	87.42*	26.16*	24.12*	**21.38***
ACTU	Y	N	28.18*	34.43*	25.81*	87.73*	88.65*	87.54*	27.39*	20.25*	19.09*
ACTU	Y	Y	28.36*	37.35*	27.28*	88.18*	88.17*	87.6*	27.98*	22.06*	20.77*
ACTU-L	N	N	28.42°	38.21°	27.62°	88.5°	88.01°	87.76°	**28.25°**	22.04°	20.88°
ACTU-L	Y	Y	27.52°	34.84°	25.31°	88.15	88.09	87.55	27.13°	21.28°	19.44°

Table 3.4: Results of direction classification for models that optionally use climate variables (C) and DEM (D), i.e., an input-modality ablation over auxiliary modalities. * denotes a t-test-determined statistically significant difference ($p < 0.05$) in persistence. The statistical difference between ACTU-L and the identical ACTU configuration is shown by the symbol °.

Model	D	C	MAE	MAE@10	MAE@20	PC
Constant	N	N	.0351	.142	.1038	-
Persistence	N	N	.1281	.1892	.171	31.81
ACTU	N	N	**.0261***	.0873*	.0611*	46.45*
ACTU	N	Y	.0266*	.0911*	.0639*	44.4*
ACTU	Y	N	.0297*	.0886*	.0643*	44.46*
ACTU	Y	Y	.0315*	.088*	.0634*	41.41*
ACTU-L	N	N	.0275°	**.0843°**	**.0589°**	**46.62**
ACTU-L	Y	Y	.0282°	.0923°	.066°	43.45°

Model	D	C	P@0.1	R@0.1	F@0.1	P@0.2	R@0.2	F@0.2
Constant	N	N	0	0	0	0	0	0
Persistence	N	N	23.86	**81.77**	34.98	13.38	**75.72**	20.25
ACTU	N	N	**51.97***	41.61*	43.04*	45.27*	24.48*	26.85*
ACTU	N	Y	51.32*	41.59*	42.77*	42.34*	18.12*	21.02*
ACTU	Y	N	49.36*	45.09*	43.64*	40.67*	28.26*	27.8*
ACTU	Y	Y	46.92*	45.18*	42.67*	39.57*	27.91*	27.3*
ACTU-L	N	N	50°	**47.19°**	**45.61°**	44.45°	27.55°	**28.65°**
ACTU-L	Y	Y	51.63°	38.48°	40.65°	43.24°	21.41°	23.41°

Table 3.5: Results of magnitude regression for models that optionally use climate variables (C) and DEM (D), i.e., an input-modality ablation over auxiliary modalities. * denotes a t-test-determined statistically significant difference ($p < 0.05$) in persistence. The statistical difference between ACTU-L and the identical ACTU configuration is shown by the symbol °.

Loss-function Ablation

The ablation on our custom regression loss function in Table 3.6 showed its superiority over a standard Huber loss and its individual components (multi-scale loss, wavelet loss). The combined loss function provided a balanced contribution from both the spatial and frequency domains, leading to better overall regression performance and higher precision when thresholding the output for classification.

Discussion

The success of the ACTU model demonstrates the feasibility of using deep learning for long-term, large-scale hydrological forecasting by integrating satellite, climate, and topographic data. The improvements over baselines highlight the

Table 3.6: Comparison between the wavelet loss (L_W), multiscale loss (L_{MS}), their combination L_T, and standard application of the regression loss (L). * indicates statistically significant difference ($p < 0.01$) with respect to L_T according to the t-test.

Model	MAE	MAE@10	MAE@20	PC
L_T	**.0261**	.0873	.0611	**46.45**
L	.029*	.0861*	.06*	42.9*
L_{MS}	.0291*	**.0839***	**.0585***	43.82*
L_W	.0285*	.1047*	.0765*	42.1*

Model	P@0.1	R@0.1	F@0.1	P@0.2	R@0.2	F@0.2
L_T	**51.97**	41.61	43.04	**45.27**	24.48	**26.85**
L	46.14*	45.65*	42.25	39.89*	24.25	24.61*
L_{MS}	45.85*	**49.84***	**44.83**	39.38*	**27.2**	26.75*
L_W	51.24	25.11*	29.88*	38.51*	16.24*	18.42*

predictive power contained in historical data for forecasting future water dynamics. Our XAI analysis enriches these findings, highlighting the potential areas for improvement in future research. The climate subgroup analysis identified specific areas (e.g., the Great Salt Lake) and climatic conditions (e.g., high variability in evapotranspiration) where the model systematically underperforms. Additionally, the saliency analysis of the importance of spectral bands revealed the strong impact of Red, NIR, and one of the SWIR channels.

3.2.4 Section Summary

In this section, we introduced HydroChronos, the first large-scale, multi-modal benchmark for surface water forecasting, alongside the AquaClimaTempo UNet, a baseline model that effectively synthesizes satellite, climate, and elevation data. By creating this foundational dataset, we have enabled a new avenue of research into a critical ecological problem. Having addressed the challenges of data scarcity, we now follow the path of our last analysis in providing models that are not just accurate but also trustworthy for real-world adoption. The next section investigates this crucial aspect of model explainability, exploring how architectural choices can provide domain experts with more interpretable insights.

3.3 Agriculture

The application of deep learning to remote sensing data for modern precision agriculture provides essential tools for sustainable farming. Accurate segmentation of crop fields from satellite imagery enables precise monitoring of crop health, optimization of resource allocation like irrigation and fertilization, and informs large-scale economic and policy decisions [14, 13, 24]. However, the underlying models should be highly accurate and transparent, because practitioners and policymakers need to trust a model's outputs to adhere to regulations, which is only possible if its decision-making process is understandable and verifiable [173].

The current state-of-the-art for semantic segmentation tasks in remote sensing heavily relies on robust architectures like the U-Net, which has proven highly effective for identifying crop boundaries [11, 199, 41]. However, the inherent opacity of deep neural networks presents a significant challenge. Their "black-box" nature can be a barrier to adoption in applications where accountability is crucial [3, 115, 10]. To mitigate this, the field of Explainable Artificial Intelligence (XAI) offers methods to interpret model behavior. While some approaches build interpretability directly into the model's design, they often do so at the cost of predictive power [88]. A common alternative is the use of post-hoc techniques, which generate explanations for a pre-trained model without altering its performance. Among these, saliency maps produced by methods like Grad-CAM [161] are frequently used to visualize which parts of an input image were most influential in a model's prediction, a technique already validated in earth observation contexts [81, 61].

A research gap emerges at the intersection of architectural innovation and explainability. The recent development of Kolmogorov-Arnold Networks (KANs) introduces a new paradigm that promises greater intrinsic interpretability [102]. An adaptation of this, the U-KAN, has demonstrated superior performance and efficiency in medical imaging [94], but it remains completely unexplored for agricultural remote sensing and crop field segmentation. Additionally, while KANs are theoretically more transparent, no study has evaluated the quality or nature of their post-hoc explanations. It is unknown if they generate more faithful and plausible saliency maps compared to a traditional U-Net. *KAN you see it? KANs and Sentinel for Effective and Explainable Crop Field Segmentation* [139] addresses this gap by presenting the first application and comparative analysis of the U-KAN architecture for crop field segmentation. We evaluate its performance against the U-Net, and we investigate the explainability of both models through a systematic analysis of their generated saliency maps to determine which architecture provides more reliable and useful insights for agricultural applications.

3.3.1 Methodology

In this section, we present the task and the architectures analyzed.

Task

The task is crop field segmentation from satellite imagery, which can be framed as binary semantic segmentation as shown in Figure 3.8. Given a satellite image I with size $W \times H \times D$, the objective is to generate a corresponding binary segmentation mask M. This output mask M, of size $W \times H$, should assign a label to each pixel. A pixel value of 1 in the mask M signifies that the corresponding pixel in the input image I belongs to a cultivated area. In contrast, a value of 0 indicates it belongs to a non-cultivated area. The goal is to train a model to generate M given I.

Figure 3.8: Crop field segmentation task. From left to right: a Sentinel-1 image (VV polarization), the corresponding Sentinel-2 RGB image, and the ground truth binary mask where cultivated areas are highlighted in white.

Model Architecture

The investigation involves two architectures: U-Net and U-KAN. The U-Net [145] is a well-established fully convolutional network known for its efficacy in semantic segmentation. Its architecture comprises a symmetric encoder-decoder structure. The encoder, or contracting path, progressively downsamples the input image using a series of convolutional and max-pooling layers. The decoder, or expanding path, symmetrically upsamples the feature maps and concatenates them with corresponding high-resolution features from the encoder path via skip connections. This design allows the network to combine deep, semantic information with precise localization details. A schematic is provided in Figure 3.9

The U-KAN [94] architecture modifies the standard U-Net by replacing the deepest layers of the encoder and decoder with KAN-based blocks. Instead of using fixed activation functions (like ReLU) on nodes, KANs [102] feature learnable activation functions on the network edges, parameterized as splines. This allows the network to learn custom, potentially more complex relationships within the

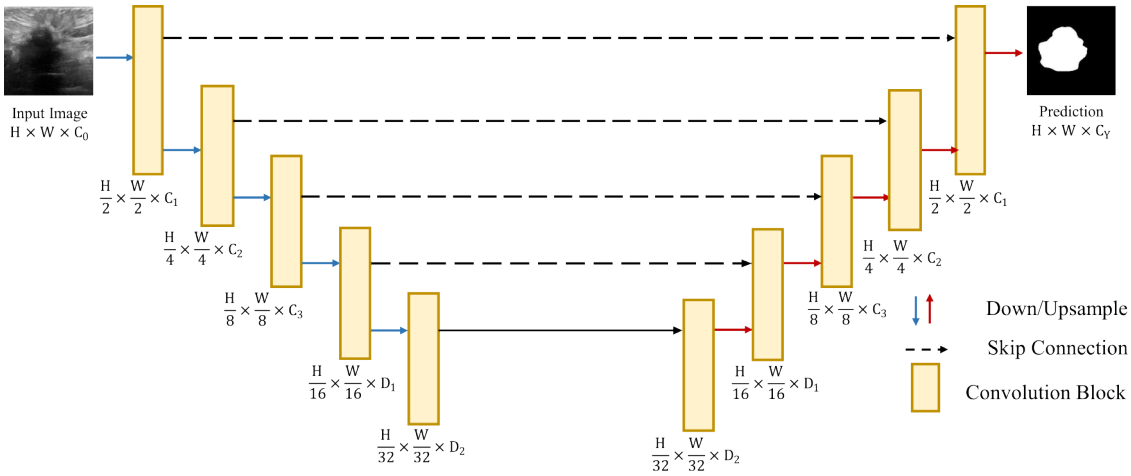


Figure 3.9: UNet architecture [145]. The encoder path (left) progressively downsamples the input, while the decoder path (right) upsamples the feature maps to recover spatial resolution. Skip connections (dashed lines) pass high-resolution features from the encoder to the decoder.

data. In the U-KAN, the bottleneck of the U-Net is composed of blocks containing a tokenization layer, a KAN layer, a downsampling layer, and a normalization layer, while preserving the overall U-shaped structure and skip connections. This hybrid approach aims to leverage the spatial feature extraction power of convolutions in the shallower layers with the adaptive learning capability of KANs in the deeper feature space. A schematic is provided in Figure 3.10.

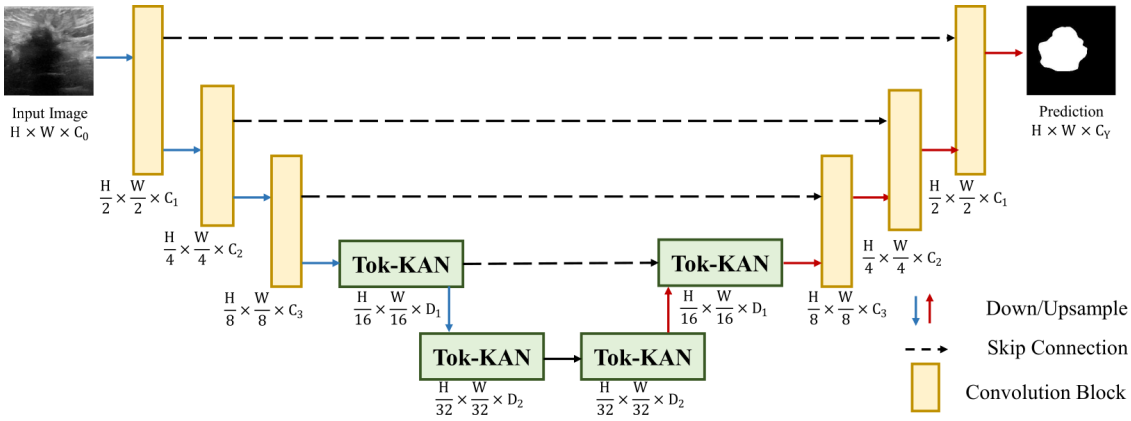


Figure 3.10: UKAN architecture [94]. This model replaces the convolutional blocks in the deepest layers with Tok-KAN blocks. These blocks leverage Kolmogorov-Arnold Networks (KANs) to learn adaptive activation functions.

3.3.2 Experimental Setup

In this section, we detail the experimental setup.

Datasets

We employ the South Africa Crop Type dataset [181], which provides Sentinel-1 (S1) and Sentinel-2 (S2) imagery over a region in South Africa characterized by small and irregularly shaped crop fields. The images are of size 256×256 pixels. For S1, we used the VV and VH polarization bands (2 channels). For S2, we used 12 multispectral bands. We annotate S2 data with cloud masking using the s2cloudless algorithm [165]; images with a cloud coverage greater than 70% over the annotated crop areas were excluded. As no official splits were provided, we randomly divided the data into training (2019 samples), validation (267 samples), and test (364 samples) sets. We ensure similar class distributions across splits according to the chi-square test ($p > 0.9$).

Evaluation Metrics

We employ Intersection-over-Union (IoU), F1-Score, Precision, and Recall to measure the segmentation performance for the positive (cultivated) class. To assess computational efficiency, we also measured Giga Floating Point Operations Per Second (GFLOPs).

For the explainability analysis, we used Grad-CAM [161] to generate saliency maps and evaluated their quality using three metrics:

- Plausibility: The alignment between the saliency map and the ground truth mask, measured with IoU and F1-score.
- Sufficiency: The model's performance on an image where only the most salient pixels (as determined by the explanation) are retained. A smaller performance drop indicates a more sufficient explanation.
- Per-channel Relevance: The impact of occluding individual input channels on the resulting saliency map, measured by the IoU between the original and occluded-channel saliency maps. A lower IoU indicates higher channel importance.

KAN Initialization

KAN parameters are initialized as follows (i.e, as suggested in official implementations [102, 94]): (i) the base linear weights are initialized with Kaiming-uniform initialization; (ii) spline coefficients are not sampled directly, but obtained by sampling small random noise values on the interior knot grid and converting these samples to B-spline coefficients via a least-squares fit.

3.3.3 Results and Discussion

In this section, we present the results for both the segmentation task and the explainability analysis, followed by a qualitative analysis.

Quantitative Analysis

The experimental results in Table 3.7 demonstrate the effectiveness and efficiency of the U-KAN architecture. On the 12-channel Sentinel-2 data, U-KAN achieved a higher IoU (74.82%) compared to U-Net (72.95%). On the 2-channel Sentinel-1 data, their IoU scores were comparable (65.36% for U-KAN vs. 65.59% for U-Net). However, U-KAN achieved this performance with approximately half the computational cost, requiring only 45.65 GFLOPs compared to U-Net's 80.65 GFLOPs on S2 data.

	Model	GFLOPs ↓	IoU ↑	F1 ↑	Prec ↑	Rec ↑
S1	U-Net	79.89	**65.59**	**79.21**	74.56	**85.56**
	U-KAN	**44.89**	65.36	79.03	**77.40**	77.50
S2	U-Net	80.65	72.95	84.35	72.57	**94.33**
	U-KAN	**45.65**	**74.82**	**85.59**	**75.24**	93.31

Table 3.7: Results of U-Net and U-KAN using Sentinel-1 (S1) and Sentinel-2 (S2) data.

The U-KAN was also preferred by the explainability metrics. U-KAN's saliency maps obtained a higher IoU with the ground truth (73.52% vs. 68.19%) in the plausibility assessment (Table 3.8), suggesting that its explanations are more spatially linked with the actual agricultural fields. An interesting finding about sufficiency (Table 3.9) is the fluctuation in the Precision metric. Precision has improved for both U-KAN and U-Net, while other measures have decreased, which is consistent with eliminating fewer important pixels. This improvement in accuracy is significant since both networks show improved capacity to distinguish pixels that belong to the crop class by eliminating less significant pixels.

Model	IoU	F1	Prec	Rec
U-KAN	**73.52**	80.77	**84.49**	77.74
U-Net	68.19	**84.50**	82.23	**87.13**

Table 3.8: Plausibility of Explanations for Sentinel-2.

Model	IoU	F1	Prec	Rec
U-KAN	67.94 (-6.88)	80.59 (-5.0)	84.46 (+9.22)	77.45 (-15.86)
U-Net	68.85 (-4.1)	81.26 (-3.09)	84.10 (+11.53)	78.95 (-15.38)

Table 3.9: Sufficiency of Explanations for Sentinel-2.

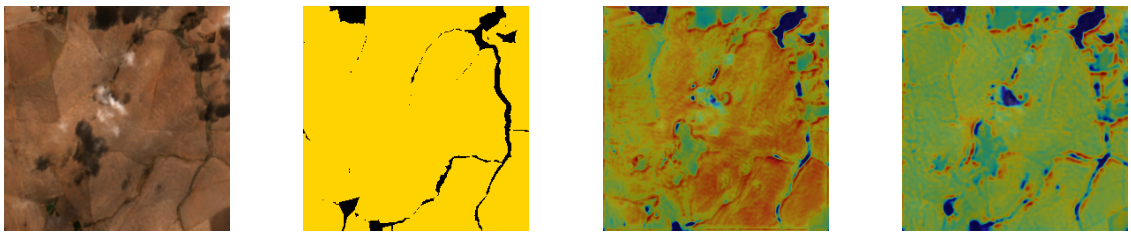
The per-channel relevance analysis in Table 3.10 revealed that both models identified the Red Edge (B05), Narrow Near-Infrared (B8A), and Shortwave Infrared (B11) bands as the most relevant, which aligns with their known sensitivity to vegetation biomass and moisture content.

	B01	B02	B03	B04	B05	B06	B07	B08	B8A	B09	B11	B12
U-KAN	72.92	74.62	46.90	45.90	**0.11**	47.49	38.51	15.61	**0.00**	42.48	**0.19**	47.03
U-Net	68.08	68.90	46.24	46.26	**0.13**	47.19	36.91	15.54	**0.00**	42.55	**0.18**	48.30

Table 3.10: Per-channel relevance (lower is better) based on IoU for Sentinel-2.

Qualitative Analysis

Visual analysis of the Grad-CAM saliency maps, as shown in Figure 3.11, highlights the difference in the models' decision process. The U-Net model tends to focus on the interior of the cultivated areas, while the U-KAN focuses on the boundaries of the crop fields. This suggests that U-KAN prioritizes the precise delineation of edges. This characteristic could be helpful for applications requiring precise boundary mapping.

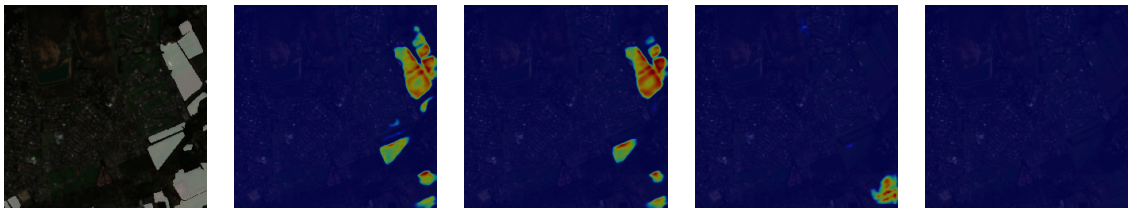


(a) RGB Image (b) Ground Truth (c) U-Net (d) U-KAN

Figure 3.11: (a) displays the image from Sentinel-2 in RGB, and (b) shows the corresponding ground truth, with crop field areas for segmentation highlighted in yellow. (c) and (d) present the saliency maps generated by U-Net and U-KAN, respectively, where red pixels indicate the areas of highest network focus.

A visual example of the per-channel relevance analysis is shown in Figure 3.12. When all channels are present, the saliency map highlights the target crop areas.

However, when an important channel like B11 was occluded, the model failed to focus on any specific area. This provides qualitative evidence that the model relies heavily on this specific band to identify crop fields.



(a) RGB and GT (b) 12 Channels (c) No B01 (d) No B06 (e) No B11

Figure 3.12: Per-channel relevance examples of U-KAN. The figure shows the ground truth over the original RGB image (a) and the saliency map of all 12 channels (b). Images (c), (d), and (e) display saliency maps generated by obscuring channels corresponding to B01, B06, and B11, respectively.

Discussion

The results show some advantages of integrating KAN layers into segmentation architectures for agricultural remote sensing. U-KAN is more computationally efficient, since it can achieve better or comparable accuracy with half the GFLOPs. The most insightful finding is the qualitative difference in their decision process. U-KAN's tendency to focus on boundaries suggests a potentially robust segmentation strategy. By working on the edges, the model may be less distracted by intra-field variations (e.g., slight changes in crop health or density). This makes U-KAN particularly promising for tasks where precise boundary delineation is the primary objective.

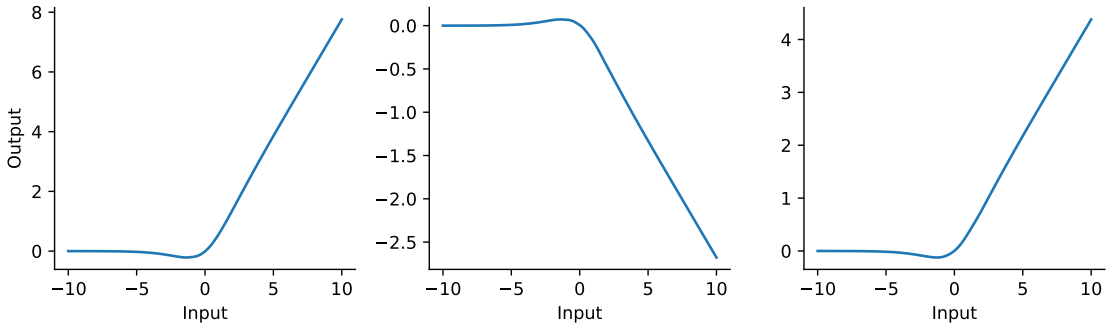


Figure 3.13: Activation functions for the first element of the embeddings of the KAN decoder layer.

Additionally, the analysis of the KAN layers learned activation functions shows the model has learned complex functions tailored to the data, as shown in Figure 3.13. The identification of relevant spectral bands via per-channel relevance analysis provides practical suggestions for model optimization.

3.3.4 Section Summary

In this section, we presented the first application of the U-KAN architecture to crop field segmentation. We demonstrated that this novel architecture is not only more computationally efficient than a standard U-Net but also achieves the results with a different decisional process, focusing on crop boundaries. This suggests that specialized models are essential for building trust with domain experts, a critical step for operational deployment. So far, we have explored solutions for data efficiency and model interpretability on specific, well-defined tasks. The final section of this chapter will now tackle the ultimate challenge: creating a single, unified framework that can synthesize multiple, disparate sensor modalities and natural language for a broad semantic retrieval task.

3.4 Land Use Land Cover Classification

Accurate and timely Land Use and Land Cover (LULC) classification derived from satellite remote sensing is fundamental to addressing critical global challenges, from managing sustainable agricultural practices and monitoring environmental change to enabling rapid disaster response [187, 172, 51, 63, 149]. As the volume and diversity of Earth observation data grow, it is necessary to effectively search relevant imagery from vast, multimodal archives. Natural language can be used to construct a query in a more meaningful and precise way from the human perspective. However, this approach needs to bridge the semantic gap between textual queries and the complex visual content of satellite data.

Recent advances in Text-To-Remote-Sensing-Image Retrieval (T2RSIR) have been largely driven by the adaptation of contrastive learning frameworks [130, 78], which have proven effective at aligning text with remote sensing imagery. However, the main focus has been on high-resolution aerial or RGB satellite imagery [80, 198, 97, 179, 129]. This reliance on the visible spectrum limits the employment of crucial information from other sensor modalities. For instance, Synthetic Aperture Radar (SAR) offers all-weather, day-and-night imaging capabilities essential for infrastructure monitoring and change detection [172], while multispectral imagery (MSI) provides spectral bands effective to analyze vegetation health and soil properties [59, 51, 76, 168]. While the alignment between multispectral data was done only with geographic coordinates [85], rich semantic text was not included, leaving the full potential of a unified, multi-sensor retrieval system unexplored.

This context reveals two different research gaps. First, existing corpora are inadequate for training and validating multi-sensor models, as they are often limited to RGB-only imagery and are annotated with ambiguous free-form text. Second, no existing architecture is capable of jointly creating a shared semantic space that unifies textual descriptions with both multispectral optical and SAR imagery. To address these issues, *Text-to-Remote-Sensing-Image Retrieval beyond RGB Sources* [29] introduces the CrisisLandMark corpus, a new, large-scale dataset of over 647,000 Sentinel-1 (SAR) and Sentinel-2 (optical) images with structured LULC and crisis event annotations. To leverage this data, a novel multimodal architecture, CLOSP (Contrastive Language Optical SAR Pretraining), is also proposed. This framework is the first to align text, optical MSI, and SAR data (and optionally locations) into a unified embedding space, enabling cross-modal knowledge transfer and advancing the state-of-the-art in semantic retrieval for Earth observation.

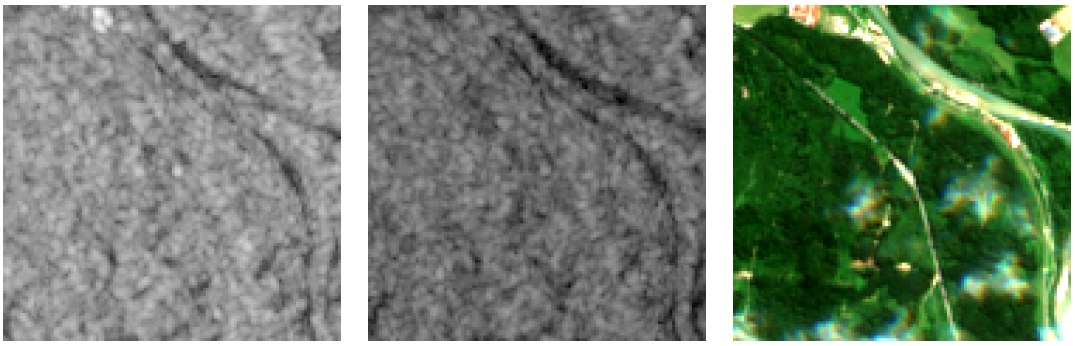
3.4.1 Methodology

Data Acquisition and Sources

The CrisisLandMark corpus was constructed by aggregating and processing data from five distinct public datasets as shown in Table 3.11: re-BEN [39], CaBuAr [25], MMFlood [116], Sen12Flood [131], and QuakeSet [137]. These sources provide a mix of general land cover scenes and data from specific crisis events (wildfires, floods, earthquakes). The imagery consists of Sentinel-1 Ground Range Detected (GRD) products, which provide dual-polarization (VV and VH) SAR data, and Sentinel-2 Level-2A (L2A) products, which offer 12-band multispectral optical data. This combination provides diverse sensor modalities and thematic content. A sample from the dataset is shown in Figure 3.14.

Dataset	Task	S1 (#)	S2 (#)	Crisis event
re-BEN	Classification	286159	286214	N
CaBuAr	Segmentation	N	3272	wildfire
QuakeSet	Classification	21430	N	earthquake
MMFlood	Segmentation	27880	N	flooding
Sen12Flood	Classification	2873	18975	flooding
		338342	308461	

Table 3.11: Composition of the corpus. We separately report the Sentinel-1 (S1) and Sentinel-2 (S2) data sources, crisis event data, and sample size from each dataset.



(a) Sentinel-1 VV (b) Sentinel-1 VH (c) Sentinel-2 RGB

Figure 3.14: Sample images from Sentinel-1 VV, and Sentinel-1 VH, and Sentinel-2 RGB of the same area. The scale of Sentinel-1 image values for each channel is the same.

Data Processing and Annotation

We create a harmonized and analysis-ready corpus: all source images were resampled to a uniform spatial resolution of 10 meters and partitioned into non-overlapping 120x120 pixel patches. We couple each image with structured, multi-label textual annotations. Instead of relying on ambiguous free-form text, we harmonized two authoritative land cover classification systems. For European regions, we used the detailed 43-class CORINE Land Cover (CLC) system. For crisis-specific datasets (which cover areas outside Europe), we queried the Dynamic World (DW) near-real-time 9-class system. We created a mapping from the fine-grained CLC classes to the 9 high-level DW classes to establish a unified label space. The original crisis tags (e.g., "wildfire," "earthquake damage") from the specialized datasets were retained. This approach provides unambiguous, machine-readable semantic information, enabling rigorous model training and evaluation. The classes are detailed in Table 3.12.

Dataset Structure and Statistics

The final CrisisLandMark corpus contains 647,603 image-text pairs, split into a training set (20%) and a retrieval/evaluation set (80%) using stratified multi-label sampling [159] to ensure consistent class distribution. The harmonized label set consists of 9 land cover classes and 3 crisis event classes. From the retrieval set, we generated a query set of 2,047 unique multi-label queries by taking all co-occurring label combinations. We define a graded relevance rel based on the IoU between the query labels L_q and the image labels L_i to provide a more fine-grained evaluation:

$$rel = round\left(10 \cdot \frac{|L_q \cap L_i|}{|L_q \cup L_i|}\right) \tag{3.1}$$

Label	Description	Images (%)
Dynamic World Land Cover Classes		
Trees	Any significant cover of trees.	69.00
Crops	Land cultivated for agriculture.	57.28
Shrub and Scrub	Areas dominated by shrubs or low, woody vegetation.	35.80
Water	Open and permanent water bodies.	28.92
Grass	Land covered predominantly by grasses and other non-woody vegetation.	27.18
Built	Artificial, man-made surfaces and structures.	18.65
Flooded Vegetation	Areas where vegetation (e.g., forests, crops) is temporarily inundated with water.	7.62
Bare	Land with little to no vegetation cover.	6.95
Snow and Ice	Surfaces permanently or seasonally covered by snow or ice.	2.38
Crisis Event Classes		
Flooded Area	Land temporarily submerged by water due to a flood event.	7.69
Earthquake Damage	Visible structural damage to built areas or significant land deformation caused by seismic activity.	3.31
Burned Area	Land showing evidence of recent fire, characterized by burn scars and the destruction of vegetation.	0.54

Table 3.12: The CrisisLandMark corpus's harmonized classes, their descriptions, and the proportion of samples that belong to each class.

Model Architecture

We propose CLOSP, a contrastive learning architecture designed to align text, MSI, and SAR imagery. The model employs three separate encoders: a text encoder (based on Sentence Transformers [141]), an MSI image encoder (pre-trained on SSL4EO [178]), and a SAR image encoder (also pre-trained on SSL4EO [178]). Since paired optical and SAR images of the same location at the same time are rare, CLOSP uses the shared textual annotations as a bridge. It aligns each visual modality with the text independently, projecting them into a shared latent space without enforcing a direct alignment between the two images. We also developed

GeoCLOSP, which also incorporates geographic coordinates. It adds a location encoder from SatCLIP [85] that learns embeddings from the latitude and longitude coordinates of each image. This location embedding is then aligned with the corresponding image embedding in parallel with the primary image-text alignment.

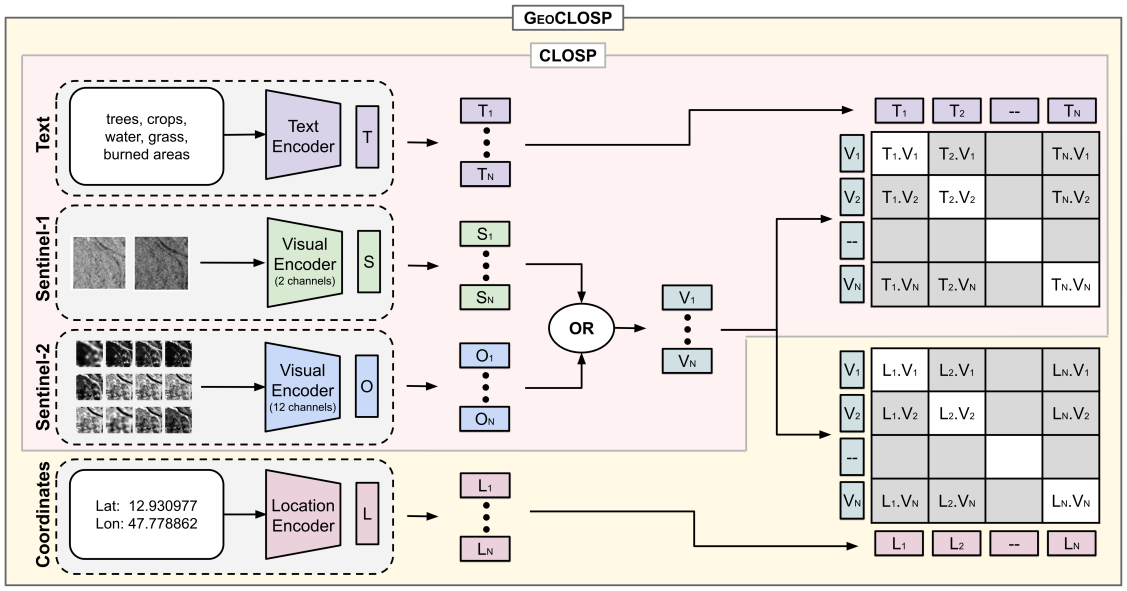


Figure 3.15: Textual descriptions and SAR and MSI satellite images are aligned by the CLOSP model. One modality, either SAR or optical, is chosen for each element in a batch of N items (M SAR and M MSI), and the respective image embeddings are linked with their corresponding textual embeddings. In order to ensure that negative pairs—which are created by combining textual and image embeddings from various items within the batch—are successfully separated, the model is trained to maximize alignment for these positive pairs, which are represented by white cells in the matrix. By adding a location encoder that parallels the image-text alignment and aligns an item's geographical coordinates with the accompanying satellite picture, GeoCLOSP expands on the CLOSP architecture.

Training Procedure

The CLOSP model is trained using a symmetric cross-entropy loss function, following the approach of CLIP [130]. For a given batch, the model maximizes the cosine similarity between the embeddings of corresponding image-text pairs (positive pairs) while minimizing the similarity for all other non-corresponding pairs in the batch (negative pairs). The total loss is the average of the image-to-text and text-to-image losses. The GeoCLOSP model extends this objective, averaging four components: the standard image-to-text and text-to-image losses, plus an

image-to-location and location-to-image loss, which aligns the visual and geographic embeddings.

3.4.2 Experimental Setup

We used ChromaDB as a vector database for the retrieval with dot product as similarity metric.

Baselines and Comparative Methods

We compared CLOSP and GeoCLOSP against several state-of-the-art T2RSIR models, including CLIP [130], SkyCLIP [179], RemoteCLIP [97], and SenCLIP [80]. Since these baselines are designed for RGB inputs, we created 3-channel inputs by extracting the RGB bands from Sentinel-2 and creating false-color composites for Sentinel-1 [5]. We evaluated both the original models and versions fine-tuned on our training set (denoted with a "-T" suffix). We also compared against a model with two specialized separate encoders (Text-SAR, Text-MSI) trained independently, which we call BiCLIP. All the tested configurations and their respective resource consumption in GFLOPs are expressed in Table 3.13.

	Vision Backbone	Textual Backbone	GFLOPs
CLIP	ResNet50	CLIP-Transformer	3
SkyCLIP	ViT-L	CLIP-Transformer	103
RemoteCLIP	ResNet-50	CLIP-Transformer	3
SenCLIP	ResNet-50	CLIP-Transformer	3
CLOSP-RN	ResNet-50	MiniLM	3
CLOSP-VS	ViT-S	MiniLM	9
CLOSP-VL	ViT-L	MiniLM	120
GeoCLOSP	ResNet-50	MiniLM	3

Table 3.13: Models along with the corresponding visual and textual frameworks. Every CLOSP suffix identifies the vision backbone that is being used. At indexing time, the GFLOPs are computed using 12-channel optical images, which is the worst-case scenario.

Evaluation Metrics

For the primary text-to-image retrieval task, we measured performance using Recall@k, Precision@k, and normalized Discounted Cumulative Gain (nDCG)@k for k $\in \{10, 50, 100, 1000\}$. For the zero-shot classification task, we used macro-averaged Precision, Recall, and F1-score across all classes, which is appropriate for the imbalanced nature of the dataset.

3.4.3 Results and Discussion

Quantitative Analysis

The text-to-image retrieval results, detailed in Table 3.14, demonstrate the superiority of the CLOSP family. Specifically, the GeoCLOSP model achieves the highest performance across the majority of metrics, with an nDCG@1000 of 57.76%. This represents a significant improvement of nearly 20 absolute percentage points over the strongest fine-tuned baseline, SkyCLIP-T, which scored 37.88%.

As illustrated in Figure 3.16, both the CLOSP-RN and GeoCLOSP models consistently outperform the best-performing baseline, SkyCLIP-T, across all cutoff levels. The performance curves for nDCG and Precision show a substantial margin between our models and the existing state-of-the-art, confirming the significant advantage of the unified training. All CLOSP models also surpass all non-fine-tuned baselines, the dummy predictor, and the specialized BiCLIP.

Model	nDCG@10	nDCG@1000	P@1000	R@1000
Dummy	28.03	33.64	2.50	0.15
CLIP	18.64	20.70	4.57	0.13
RemoteCLIP	19.27	21.68	4.08	0.09
SkyCLIP	24.88	28.21	7.10	0.21
SenCLIP	34.01	37.60	18.15	0.52
RemoteCLIP-T	29.00	26.57	8.70	0.27
SkyCLIP-T	33.46	37.88	17.57	0.58
SenCLIP-T	20.59	23.76	5.41	0.13
BiCLIP	38.33	44.90	27.20	1.31
CLOSP-RN	50.50	56.23	40.66	2.05
CLOSP-VS	49.18	54.51	40.22	2.14
CLOSP-VL	47.82	55.91	42.14	**2.32**
GeoCLOSP	**51.14**	**57.76**	**42.98**	2.10

Table 3.14: Mean performance (%) for each model in terms of nDCG, Precision (P), and Recall (R) at given cutoffs. **Bold** values indicate the best result for each metric.

Ablation Studies

To understand the benefits of our unified training strategy, we evaluated the models on the Sentinel-1 and Sentinel-2 portions of the corpus separately. The results in Table 3.15 show a large improvement for Sentinel-1 (SAR) retrieval,

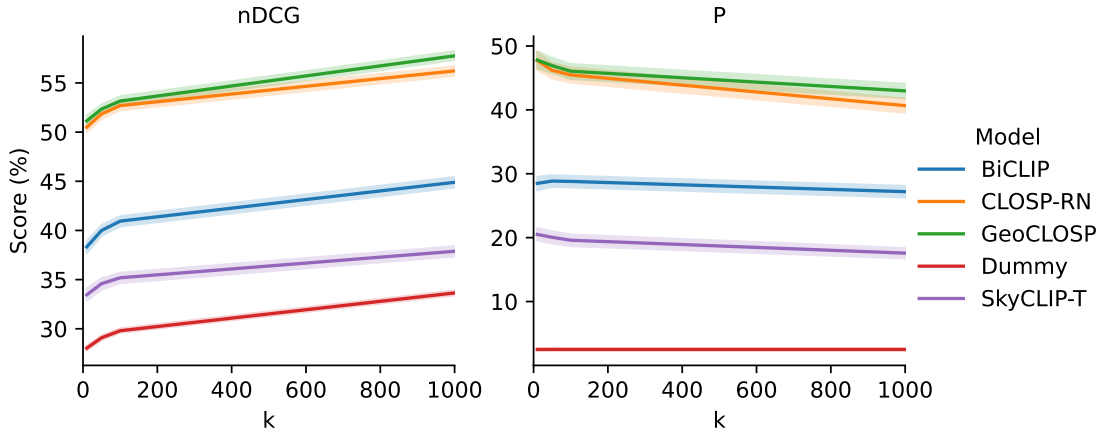


Figure 3.16: Mean nDCG (left) and Precision (right) performance at different cutoff levels. The bands represent the 95% confidence intervals.

with CLOSP-VL achieving an nDCG@1000 of 55.65% compared to the specialized BiCLIP's 37.35%. This provides strong evidence of cross-modal knowledge transfer, where knowledge from the easier-to-interpret optical data is used to disambiguate the complex SAR imagery. This gain is achieved with only a negligible performance drop on the Sentinel-2 data.

Model	nDCG@10		nDCG@1000		P@1000		R@1000	
	S1	S2	S1	S2	S1	S2	S1	S2
BiCLIP	31.99	**49.56**	37.35	**55.72**	18.51	**39.65**	0.78	2.24
CLOSP-RN	45.93	49.33	53.60	54.80	37.51	38.46	1.76	1.92
CLOSP-VS	45.96	42.93	51.79	49.96	37.43	34.86	1.75	2.12
CLOSP-VL	**46.59**	45.97	**55.65**	53.85	**41.83**	38.04	**2.12**	**2.26**

Table 3.15: Mean performance (%) for each model by modality in terms of nDCG, Precision (P), and Recall (R) at given cutoffs. **Bold** values indicate the best result for each metric. *S1* is Sentinel-1 SAR data and *S2* is Sentinel-2 multispectral data.

We also evaluated the models on a zero-shot multi-label classification task as proposed in Radford et al. [130] and Wang et al. [179]. The results are reported in Table 3.16. Again, the CLOSP models significantly outperformed all baselines, with CLOSP-VS achieving the highest F1-score of 41.82%. Interestingly, GeoCLOSP did not show an improvement in this purely semantic task, further highlighting the trade-off between learning semantic versus spatial features.

Model	F	P	R
Dummy	12.87	10.52	16.67
CLIP	25.12	21.99	52.02
RemoteCLIP	25.61	21.15	48.11
SkyCLIP	29.10	22.58	58.96
SenCLIP	17.73	20.33	49.61
RemoteCLIP-T	15.81	11.32	50.00
SkyCLIP-T	32.81	26.30	68.18
SenCLIP-T	5.41	4.48	33.34
BiCLIP	34.98	30.89	69.83
CLOSP-RN	41.56	35.59	**82.25**
CLOSP-VS	**41.82**	45.22	55.68
CLOSP-VL	37.31	**62.48**	37.40
GeoCLOSP	41.24	40.40	68.14

Table 3.16: Zero-shot classification performance (%) for each model in terms of F1-Score (F), Precision (P), and Recall (R). **Bold** values indicate the best result for each metric.

Discussion

Our work leads to three principal findings. First, integrating multispectral and SAR data compared to RGB-only improves the T2RSIR performance. Second, our unified training approach successfully bridges the gap between MSI and SAR modalities, enabling powerful knowledge transfer to difficult-to-interpret SAR data. Third, the comparison in by-class zero-shot classification between CLOSP and GeoCLOSP in Table 3.17 uncovers a fundamental trade-off between semantic and geographic representations. GeoCLOSP excels at retrieving images for classes where location is a key defining characteristic (e.g., "earthquake damage," "snow and ice"). Conversely, the purely semantic CLOSP model performs better on geographically widespread classes like "crops" and "trees." This suggests that the optimal retrieval strategy is application-dependent, with GeoCLOSP acting as a specialized tool for location-specific analysis and CLOSP as a superior general-purpose semantic retriever.

3.4.4 Section Summary

In this section, we introduced CrisisLandMark, a large-scale, multi-sensor corpus, and CLOSP, a novel architecture that unifies text, multispectral, and SAR imagery into a shared semantic space. Our framework significantly outperforms

	CLOSP-RN		GeoCLOSP		SkyCLIP-T	
Class	F	nDCG	F	nDCG	F	nDCG
bare	16.11	**25.47**	**18.08**	22.81	11.90	1.99
built	**46.15**	**72.08**	42.92	58.39	30.93	27.59
burned area	2.34	60.07	**2.84**	**61.76**	1.39	0.00
crops	**76.76**	**90.53**	68.82	67.84	67.00	79.21
earthquake damage	16.33	0.00	**21.45**	**65.50**	14.65	23.29
flooded area	38.87	32.09	**45.94**	15.91	20.39	**32.40**
flooded vegetation	22.06	**43.54**	**25.60**	25.69	14.93	2.17
grass	**52.61**	**74.40**	52.51	42.13	40.75	63.25
shrub and scrub	58.58	**50.87**	**60.60**	30.81	53.65	40.05
snow and ice	9.46	18.47	**11.15**	**60.90**	6.40	0.15
trees	**84.84**	**80.58**	78.95	59.23	80.45	58.59
water	**74.61**	**99.77**	65.96	99.30	51.30	99.67

Table 3.17: Performance by class in terms of F1-score (F) for zero-shot classification and nDCG@1000 (nDCG) for retrieval.

existing methods and, through knowledge transfer, enhances the interpretation of challenging SAR data. This work represents the culmination of this chapter's argument: it is a data-efficient, specialized architecture for multi-modal synthesis that is both a powerful tool for long-term ecological analysis and an essential capability for rapid crisis response. These findings provide the foundation for the next generation of retrieval systems and serve as a direct bridge to the topics explored in the next chapter.

3.5 Chapter Summary

The contributions detailed in this chapter prove the thesis that tackling planetary-scale challenges requires a portfolio of targeted, data-efficient, and interpretable models rather than a single, monolithic approach. This portfolio was constructed by addressing the core bottlenecks in ecological data science. First, we addressed the fundamental challenge of data scarcity with two complementary solutions. In forestry, we demonstrated a paradigm for resource-efficient analysis by adapting a general-purpose foundation model, achieving state-of-the-art results with minimal cost. Then, for hydrology, where no such foundation existed, we solved the problem by creating HydroChronos, the first large-scale benchmark for spatio-temporal water forecasting and the corresponding model.

After providing solutions for the data problem, we then demonstrated the necessity for model specialization. In agriculture, we showed specialized architectures

to build trust using explainable insights for practitioners. Finally, the CrisisLandMark corpus and the CLOSP model synthesize natural language with complex, multi-sensor archives to create a coherent, unified framework.

This final contribution, which moves beyond the limitations of RGB-only systems, serves as a direct bridge from the long-term monitoring essential for understanding ecological systems to the acute needs of disaster response. The same technologies and data sources are equally critical when the temporal scale shifts from years to hours. We now transition from the domain of ecological monitoring to Machine Learning in Crisis Management, where we will explore how these techniques can be adapted to provide the actionable intelligence needed when the primary challenge is not just to understand, but to act.

Chapter 4

Machine Learning in Crisis Management

While the previous chapter focused on the long-term monitoring of ecological systems, we discuss the applications of machine learning in crisis management. The temporal scale shifts from years to hours, and the primary goal of machine learning becomes the rapid conversion of chaotic, multi-modal data into clear, actionable intelligence for first responders and decision-makers. The challenges are distinct: data is often incomplete, noisy, and arrives in overwhelming volumes, demanding models that are not only accurate but also robust and timely.

This chapter presents some contributions across different phases and types of disasters. We first address the task of post-disaster damage assessment, introducing novel datasets and architectures for delineating **wildfire burn scars** and for monitoring the impact of **earthquakes** using all-weather satellite imagery. We move to the "information crisis" domain, presenting a novel reinforcement learning framework for summarizing the **evolution of a crisis** from heterogeneous text streams in near-real-time.

4.1 Burned Area Delination

The increasing frequency and intensity of forest wildfires, exacerbated by climate change, pose a significant threat to global ecosystems, economies, and societies [53, 150, 67]. Effective disaster management and post-event recovery planning rely on the ability to quickly and accurately map the extent of the damage. Earth Observation (EO) data, acquired from satellite missions such as Sentinel-2 [104] and Landsat [186], offer the opportunity to monitor these events at a continental scale. This data can be used with modern machine learning to provide timely and precise assessments of burned areas, supporting authorities and first responders. This section details a multi-faceted contribution to this challenge, addressing critical gaps in both data availability and modeling methodology.

4.1.1 State-of-the-Art in Burned Area Delineation

The task of burned area delineation—a binary semantic segmentation to classify each pixel of a satellite image as "burned" or "unburned"—has been approached with different methods over the years.

Traiditional Index-based Methods

Historically, the field has been dominated by methods leveraging spectral indices. These indices, such as the Normalized Burn Ratio (NBR) [148], Burned Area Index (BAI) [106], and BAIS2 [57], are calculated by combining different spectral bands from multispectral satellite imagery that are sensitive to vegetation health and moisture content. The resulting single-channel image highlights regions affected by fire. To produce a final binary mask, these index maps are typically

paired with a thresholding technique, such as the Otsu method or manually calibrated values [18, 19]. A drawback of these methods is the difficulty in identifying a universal threshold that performs consistently across different regions, vegetation types, and atmospheric conditions, as they often assume a linear separability between burned and unburned classes [30, 155].

Deep Learning Approaches

The advent of deep learning opens the possibility to learn complex, non-linear feature representations directly from data. Convolutional Neural Networks (CNNs), particularly encoder-decoder architectures like U-Net [145] and DeepLab [32], revolutionized the field by demonstrating state-of-the-art performance in numerous remote sensing applications, including burned area delineation [133, 134, 62, 86, 55, 166, 69]. These models surpassed the performance of traditional index-based techniques by learning from vast amounts of labeled data.

More recently, inspired by successes in natural language processing, the computer vision community has adopted Transformer-based architectures. Models like the Vision Transformer (ViT) [48] and SegFormer [190] have set a new state-of-the-art in various segmentation tasks. Their application to burned area delineation has also proven effective, highlighting the potential for processing complex satellite imagery [27].

4.1.2 Data Scarcity and Diminishing Model Returns

Despite the progress in model architecture, the field faces a significant bottleneck. The performance of deep learning models, especially data-hungry Transformers, is limited by the availability of large-scale, high-quality labeled datasets. The main issues for the EO community are the lack of large-scale datasets for burned area delineation, with existing resources covering limited geographic areas or time spans [40, 126]. This scarcity limits the development of robust and generalizable models but also makes it difficult to benchmark new architectures.

Additionally, our investigations revealed a clear trend: simply increasing the size and complexity of standard architectures (e.g., using larger variants) did not provide performance gains and could lead to overfitting. This suggests that existing datasets are not large enough to effectively train billion-parameter state-of-the-art models. So, the challenge is not just about creating better models, but about addressing the problem of limited data and the inefficient use of that data by large architectures.

To address these gaps, *CaBuAr: California Burned Areas dataset for delineation* [25] tackles the problem of data scarcity. We developed CaBuAr, a new, large-scale, publicly available dataset of burned areas in California, comprising thousands of square kilometers of pre- and post-fire Sentinel-2 imagery. This resource was created

to facilitate the training and evaluation of large deep learning models. Second, to address the issue of inefficient data usage, *Magnifier: A Multi-grained Neural Network-based Architecture for Burned Area Delineation*[26] introduces Magnifier, a novel multi-grained neural network architecture. Magnifier is designed to better exploit the available labeled data by combining features computed at different levels of contextual detail. This approach enables smaller, more efficient models to achieve or even surpass the performance of significantly larger state-of-the-art networks without requiring additional labeled data. Together, these contributions advance the field by tackling the problem from both a data-centric and a model-centric perspective

4.1.3 Methodology

Data Acquisition and Sources

The raw data for the CaBuAr dataset were acquired from two primary sources. The multispectral imagery consists of Level-2A products from the Sentinel-2 [104] satellite mission. The ground truth annotations were derived from vector data publicly released by California's Department of Forestry and Fire Protection (CAL FIRE).

Data Processing and Annotation

For each wildfire event, a post-fire Sentinel-2 image was collected within one month of the fire's containment to minimize the effects of vegetation regrowth. To enable change detection studies, a corresponding pre-fire image was also collected for each event. It is collected within a four-week temporal window centered exactly one year prior to the post-fire acquisition date to ensure similar seasonal characteristics. Finally, each pre- and post-fire image pair was manually inspected to discard invalid samples (e.g., due to excessive cloud cover or data artifacts).

Dataset Structure and Statistics

The final CaBuAr dataset is the largest of its kind, covering 340 wildfires and a total burned surface of approximately 11,000 km^2. It provides a larger number of samples and a greater temporal span (2015-2022) than previously available public datasets for this task, as shown in Table 4.1. Each pre- and post-fire image contains 12 channels. The high class imbalance, with a low percentage of burned pixels relative to the total area, makes it a challenging benchmark.

	CaBuAr [25]	Europe [40]	Indonesia [127]
Resolution	20m	10m	30m
Channels	12	12	8
Forest Fires	340	73	81
Start-End date	01/2015 - 12/2022	07/2017 - 07/2019	01/2019 - 12/2021
Burned surface	$\sim$ 11000 km^2	$\sim$ 2000 km^2	$\sim$7000 km^2
Number of images	688	449	227

Table 4.1: Characteristics of burned area segmentation datasets.

Problem Statement

The task of burned area delineation can be framed as binary semantic segmentation. Given a single multispectral image I_{post} or a bi-temporal pair of co-registered pre-fire and post-fire images (I_{pre}, I_{post}) and the corresponding binary mask M, where the value 1 indicates a pixel classified as burned, otherwise is 0. The task is to develop a machine learning model to automatically generate M, given the images.

Model Architecture

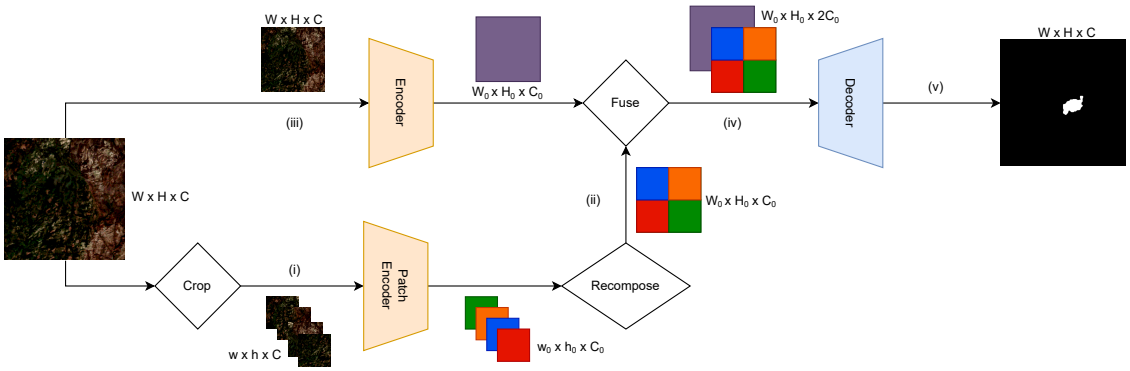


Figure 4.1: Magnifier architecture. In the lower branch, (i) the image is cropped in smaller patches (as shown in Figure 4.2a), giving each patch to an encoder. (ii) The encodings are concatenated by putting each one in the original position in the image (as shown in Figure 4.2b). In the upper branch (iii), the entire image is given to an encoder. (iv) The two encodings are concatenated along the channel axis, and (v) they are given to the decoder to get the final prediction.

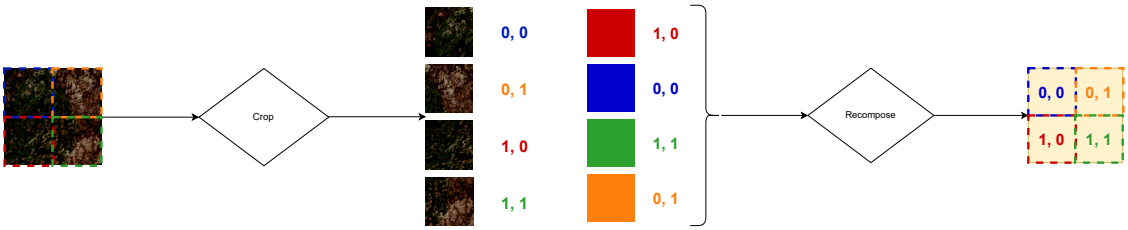
To establish strong baselines and evaluate our proposed architecture, we utilized several models for semantic segmentation:

1. U-Net [145]: A well-established CNN architecture known for its effectiveness in satellite image segmentation. Its encoder-decoder structure with skip connections allows it to capture both high-level semantic information and fine-grained spatial details.

2. SegFormer [190]: A Vision Transformer-based model designed for efficiency and high performance. Its architecture features a hierarchical encoder that generates multiscale features. The decoder consists of a simple All-MLP structure that aggregates information from the different scales to produce the final segmentation mask.

3. Magnifier [26]: Our proposed dual-encoder architecture, which we applied on top of base models like U-Net [145] and DeepLabV3+ [33] to evaluate its impact.

The core of the Magnifier architecture is its dual-encoder structure, consisting of a "global" path and a "local" path as shown in Figure 4.1. In the global path, the entire input image is fed into a standard encoder to learn coarse-grained, contextual features. In parallel, the local path crops the same image into a grid of smaller, non-overlapping patches as shown in Figure 4.2a. Each patch is processed by a second, independent encoder to capture fine-grained details. The feature maps from the local path are then reassembled (Figure 4.2b), and a fusion step concatenates them with the global feature map. This representation is passed to a single decoder to generate the final segmentation mask.



(a) **Cropping procedure.** The image is cropped in patches, and each of them keeps the original position associated.

(b) **Recompose procedure.** The encodings of an image are merged into a single embedding matrix using the position information.

Figure 4.2: Crop and Recompose operations used by Magnifier

4.1.4 Experimental Setup

Datasets

The research presented in this section utilized two key datasets. The initial investigation of Vision Transformers was conducted on a publicly available European

burned area dataset. After the creation of the CaBuAr, it is used as the primary dataset for evaluation.

Baselines and Comparative Methods

The performance of the deep learning models was compared against traditional methods based on calculating indices such as NBR [148], NBR2 [147], BAI [106], and BAIS2 [57], followed by automatic thresholding using Otsu's method [121, 19]. We also compare with BurntNet [86], a deep learning based ad-hoc solution.

Evaluation Metrics

Model performance was quantified using standard metrics for semantic segmentation:

- F1 Score: The harmonic mean of precision and recall, providing a balanced measure of a model's performance on the positive (burned) class.
- Intersection over Union (IoU): A measure of the overlap between the predicted segmentation mask and the ground truth mask.

4.1.5 Results and Discussion

Quantitative Analysis

The experimental results from all demonstrated the superiority of deep learning models over traditional spectral index-based methods [25, 26, 27]. Table 4.2 highlights our best model, Magnifier (DeepLabV3+ w/ ResNet18), surpasses all index-based methods and BurntNet on two out of the three datasets (in one of the three fails to converge) while being more computationally efficient (theoretically) with nearly one-third of the GFLOPs. It is also important to note BAIS is not applicable to Landsat-8 Indonesian dataset, because it works with Sentinel-2 bands.

Qualitative Analysis

The predictions from spectral indices were noisy, with numerous false positives scattered across unburned areas. While all deep models tend to provide better predictions, our architecture improves even more when both large and small version fails, as shown in Figure 4.3.

Table 4.2: Summary comparison of our best proposed model, **Magnifier (DeepLabV3+ w/ ResNet18)**, against all traditional spectral index methods and a state-of-the-art competitor (BurntNet [86]). Our model demonstrates a significant performance leap over all index-based methods and achieves the best overall results across the three diverse datasets.

Model	**↓GFLOPs**	**↓Params**	**California**		**Europe**		**Indonesia**	
			F1	**IoU**	**F1**	**IoU**	**F1**	**IoU**
NBR	-	-	15.0	10.3	44.0	29.7	18.8	10.4
NBR2	-	-	22.6	15.9	49.2	34.4	30.1	17.9
BAIS	-	-	4.0	2.6	17.6	10.3	8.8	4.6
BAIS2	-	-	19.4	14.8	28.5	17.5	-	-
BurntNet (SotA)	219.0	35M	71.6	62.2	**84.4**	**73.7**	-	-
Magnifier (Ours)	**76.9**	24M	**77.8**	**64.0**	83.7	72.7	**84.7**	**73.5**

(a) Ground Truth (b) Small (c) Magnifier (d) Large

Figure 4.3: Example RGB images and corresponding ground truth with predictions from a small, large, and magnifier model.

Ablation Studies

Table 4.3 details that these performance gains come from the architecture itself. Applying the Magnifier methodology to a small base model consistently boosts its performance, enabling it to outperform its corresponding larger version in 4 out of 5 configurations based on Mean Rank (MR). This demonstrates that the multi-grained approach allows for more effective feature extraction and data utilization than simply increasing model size.

Discussion

The results provided several insights. First, our initial study established Vision Transformers as a highly promising architecture for burned area delineation; however, their applications are limited by the availability of data. The study's

	Backbone	Version	M	↓GFLOPs	↓Params (M)	California F1	California IoU	Europe F1	Europe IoU	Indonesia F1	Indonesia IoU	MR
DeepLabV3+	MobNet	Small		6.0	1M	64.8	48.4	72.6	59.1	73.3	58.1	2.7
		Small	✓	8.6	2M	69.7	54.1	79.7	67.5	80.2	69.6	**1.0**
		Large		9.4	3M	60.5	44.4	74.0	60.3	75.5	60.9	2.3
	ResNet	18		40.2	11M	73.8	59.4	83.6	72.7	83.7	72.0	2.2
		18	✓	76.9	22M	77.8	64.0	83.7	72.7	84.7	73.5	**1.0**
		101		115.6	42M	76.0	62.3	81.9	70.3	82.4	70.2	2.7
U-Net	MobNet	Small		20.7	1M	62.6	47.0	75.9	62.3	74.6	59.7	2.3
		Small	✓	25.5	2M	67.3	51.6	79.1	66.4	82.4	70.2	**1.0**
		Large		24.7	3M	66.3	49.9	73.5	60.8	70.6	55.1	2.7
	ResNet	18		47.0	11M	73.2	58.0	82.1	70.4	82.0	69.6	2.0
		18	✓	78.1	22M	74.4	59.6	81.0	69.2	82.9	71.0	**1.5**
		101		127.6	42M	73.5	60.0	80.8	68.6	81.8	69.3	2.5
SF	MiT	B0		16.0	3M	71.7	56.5	81.8	70.4	82.3	70.0	2.2
		B0	✓	21.3	6M	71.5	56.9	82.5	71.0	82.2	69.9	2.0
		B1		31.4	13M	69.0	53.7	82.4	71.4	83.0	71.0	**1.8**

Table 4.3: Detailed performance analysis of deep learning architectures. For each architecture family, we show how applying our Magnifier methodology (✓) to a small base model compares against the base model itself and a larger variant. MobileNetV3 is abbreviated as MobNet, and SegFormer as SF. Best results for each metric within a backbone group are in **bold**.

limitations directly motivated the creation of CaBuAr, which now provides a robust resource for advancing research in this domain. The Magnifier architecture provides a different solution other than collecting more data samples for the task. The performance differences observed between the California and Europe datasets also highlight the persistent challenge of domain generalization, indicating that models still struggle to adapt to new geographical regions without fine-tuning [26]. While we report GFLOPs and parameter counts as proxies for computational cost, no on-device latency/throughput measurements were conducted in this thesis due to the unavailability of Jetson-Nano-class embedded hardware during the revision period. In practical deployments, end-to-end latency depends on the full inference stack (runtime, operator support, memory transfers, and pre/post-processing), and therefore theoretical complexity does not necessarily translate linearly into wall-clock performance. As future work, the models will be exported and benchmarked on representative edge devices, and their robustness will be tested under reduced-precision inference (e.g., INT8) via post-training quantization and, if needed, quantization-aware training.

4.1.6 Section Summary

In this section, we addressed the critical challenge of post-wildfire assessment by tackling its primary bottleneck: the scarcity of large-scale, labeled data. We first demonstrated the potential of Vision Transformer architectures on existing public data. Recognizing that the progress of the field was constrained by the quality of available data. By creating CaBuAr, we provided a foundational public benchmark, and with Magnifier, we introduced a novel architecture designed for superior data efficiency. This supports one of the thesis arguments: targeted, data-aware models can outperform larger, generalist architectures on specialized tasks.

We now turn to another facet of robust crisis monitoring: computational efficiency and all-weather operational capability. The next section will explore earthquake monitoring using SAR all-weather capability and the need for on-board processing demands highly resource-aware models. This necessitates a shift in both sensor modality, from optical to SAR, and in architectural design.

4.2 Earthquake Monitoring

Effective earthquake monitoring is a critical component of disaster management, traditionally relying on seismometer networks to detect and characterize seismic events. While indispensable, these ground-based systems cannot provide full and accurate global coverage, leaving significant gaps, particularly in remote or inaccessible regions. Satellite remote sensing data offers a powerful complementary view that can easily reach every part of the globe [137]. The combination of remote sensing with modern machine learning has revolutionized Earth observation, thanks to the availability of large-scale datasets [69, 166] from missions like Sentinel [171, 104] and successfully applied to various emergency management tasks, such as flood detection [20] and burned area delineation [25].

However, the application of machine learning to earthquake analysis from satellite imagery remains underdeveloped. Current state-of-the-art machine learning in seismology focuses predominantly on analyzing seismic wave data from seismograms to perform tasks like event detection and phase picking [117, 202, 31, 98]. While some research has explored image analysis for disaster response, it often relies on RGB imagery from social media, which is limited by the visible spectrum, on human presence and functioning communication infrastructure [7, 118]. The potential for using Sentinel-1 data for earthquake analysis has been demonstrated through manual analysis [58]. However, manual analysis is not scalable and struggles with complex, non-linear data relationships. Additionally, for practical real-time applications like on-board satellite processing, it is necessary to balance accuracy with computational demands [73, 153, 72, 128].

The current literature lacks any publicly available, large-scale datasets of satellite imagery specifically curated for automated earthquake analysis models. The

absence of such a resource has hindered the transition from manual analysis and seismic wave-based methods to more scalable, data-driven remote sensing solutions. To address this gap, *QuakeSet: A Dataset and Low-Resource Models to Monitor Earthquakes through Sentinel-1* [137] introduces the QuakeSet dataset, a new, publicly available collection of over 155 global earthquakes captured in tri-temporal Sentinel-1 imagery. This work proposes a series of new machine learning tasks, including earthquake detection and magnitude regression, and provides a comprehensive benchmark of both traditional and deep learning models. We focus on evaluating low-resource architectures, and we demonstrate the performance gains of using bitemporal data. This research provides the foundation to advance the field of automated earthquake monitoring from space.

4.2.1 Methodology

Data Acquisition and Sources

The raw data for QuakeSet were acquired from two primary sources: the Sentinel-1 mission and the International Seismological Centre (ISC) Bulletin.

We utilized Level-1 Ground Range Detected (GRD) products from the Sentinel-1 mission [171], acquired in the Interferometric Wide swath (IW) mode. This mode provides dual-polarization data (VV and VH) at a high resolution of 10×10 meters, which is well-suited for monitoring land changes. The C-band SAR instrument on Sentinel-1 has the advantage of being able to penetrate clouds.

From the ISC Bulletin [79], we collected ground truth annotations for earthquakes that occurred between 2018 and 2021. It provided essential metadata for each event, including the hypocenter coordinates (latitude, longitude, depth) [21], the timestamp, and its magnitude [46]. We used the Reviewed ISC Bulletin, which is manually verified by analysts to ensure data quality and coherence.

Data Processing and Annotation

First, we collected from the ISC Bulletin all known and reviewed seismic events with a magnitude greater than 4.0, focusing on events that produce detectable surface changes. For each earthquake, we defined a $20km \times 20km$ area of interest centered on its epicenter.

We collected a tri-temporal series of Sentinel-1 images for each event as shown in Figure 4.4. This series includes:

- A "post-event" image, captured within a 1 to 13-day window after the earthquake.
- A "pre-event" image, captured within a 1 to 13-day window before the earthquake.

- A "neutral" image, captured between 13 and 25 days before the event to serve as a negative control for change detection tasks.

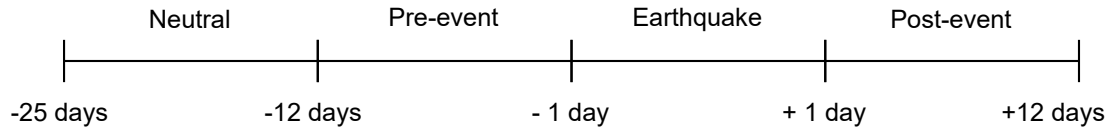


Figure 4.4: Temporal windows of collected samples

Events for which a complete tri-temporal series could not be formed were excluded. This process resulted in a collection of 155 distinct earthquake events as shown in Figure 4.5.

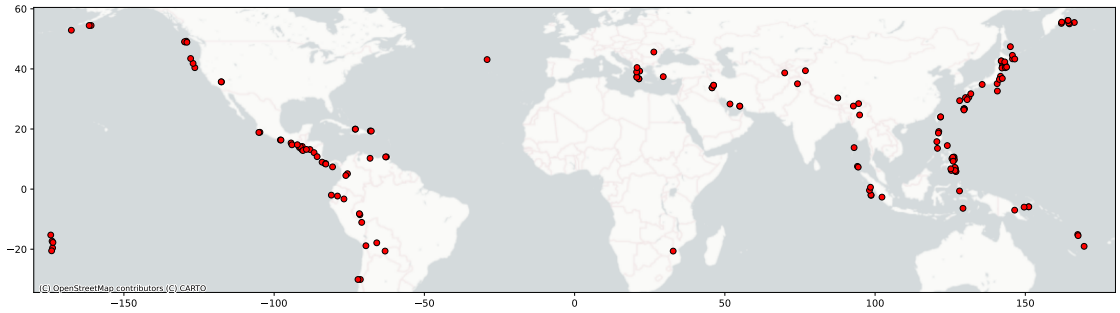


Figure 4.5: Earthquakes epicenters around the globe

To make the data more manageable for machine learning models, the $20km \times 20km$ areas were patched from the center to create images of size 512×512 pixels with two channels (VV and VH), resulting in a total of 1,906 patches. In Table 4.4, you can see a high-level summary of the statistics of the dataset.

Problem Statement

In the benchmark, we evaluate the following tasks:

Earthquake Detection We have a set of time-series $\mathcal{T}$, where each time-series $T_s \in \mathcal{T}$ is composed of N images, with $N \leq 2$, of size $W \times H \times C$ related to the same spatial area S at timesteps $\{T_1, \ldots, T_N\}$ and a ground truth value $G_t \in \{0, 1\}$ (where 1 indicate S was affected by an earthquake, 0 otherwise). Given a training dataset D_{tr}, composed of a set of pairs (T_s, G_t), we train a machine learning model M. Given a test dataset D_{ts}, composed of a set of T_s, we can predict G_t for each sample T_s using M. This can be framed as a supervised classification task.

Earthquakes	155
Temporal window	2018-2021
Image channels	2 (VV and VH)
Area size	$20km \times 20km$
Temporal difference	1-13 days
Time series length	3
Patch size	512×512
Magnitudes	> 4 mb

Table 4.4: QuakeSet Statistics

Magnitude Regression We have a set of time-series $\mathcal{T}$, where each time-series $T_s \in \mathcal{T}$ is composed of N images, with $N \leq 2$, of size $W \times H \times C$ related to the same spatial area S at different timesteps $\{T_1, \ldots, T_N\}$ and a ground truth value $G_t \in \{0...M_m\}$ (where M_m is the maximum value for the given magnitude scale). Given a training dataset D_{tr}, composed of a set of pairs (T_s, G_t), we train a machine learning model M. Given a test dataset D_{ts}, composed of a set of T_s, we can regress G_t for each sample T_s using M. This regression task is a supervised one.

We evaluated the tasks under two distinct input configurations:

1. Single Image ($N = 1$): Models were trained on individual post-event (positive class) and pre-event (negative class) images.

2. Bi-temporal Time Series ($N = 2$): Models were trained on pairs of images. Positive samples consisted of a pre-event and post-event pair, while negative samples consisted of a neutral and pre-event pair.

4.2.2 Experimental Setup

Baselines and Comparative Methods

The classical machine learning baselines included Support Vector Machines (SVM) with polynomial and RBF kernels, and Random Forest (RF). We applied Principal Component Analysis (PCA) to reduce the feature dimensionality of the flattened images to approximately 2,000 components before training shallow models. For deep learning, we focused on models designed for efficiency and low-resource environments: MobileNetV2 [71], ConvNextV2-Atto [184], MiT-B0 [190], and MobileViTV2 [110].

Evaluation Metrics

For the binary earthquake detection task, we used Accuracy, as the classes are balanced. For the magnitude regression task, we used Mean Absolute Error

(MAE). We also measured resource consumption, reporting the number of model parameters, inference time (in seconds), and Mega Floating Point Operations Per Second (MFLOPs).

4.2.3 Results and Discussion

Quantitative Analysis

Model	Params	Accuracy ↑	Time (s) ↓	MFLOPs ↓
SVC (RBF kernel)	-	0.6341	0.3640	1.0486
SVC (Poly kernel)	-	0.5440	0.3138	1.0486
RFC	-	0.7241	0.3130	1.0508
MobileNetV2 (CNN)	2.2M	0.9472	0.1109	207.5949
MiT-B0 (ViT)	3.4M	0.8865	0.0909	430.7062
ConvNextV2 (CNN)	3.7M	0.9374	0.1003	373.7727
MobileViTV2 (ViT)	4.9M	0.6536	0.1746	964.5956

Table 4.5: Performance of models for earthquake detection with bitemporal time-series

Model	Params	MAE ↓	Accuracy ↑	Time (s) ↓	MFLOPs ↓
SVR (RBF kernel)	-	2.2368	0.5519	0.3640	1.0486
SVR (Poly kernel)	-	2.6015	0.5440	0.3138	1.0486
RFR	-	1.9930	0.5440	0.3130	1.0508
MobileNetV2 (CNN)	2.2M	0.5456	0.9374	0.1109	207.5949
MiT-B0 (ViT)	3.4M	0.8496	0.9276	0.0909	430.7062
ConvNextV2 (CNN)	3.7M	0.6494	0.9315	0.1003	373.7727
MobileViTV2 (ViT)	4.9M	1.7612	0.7378	0.1746	964.5956

Table 4.6: Performance of models for magnitude regression with bitemporal time-series

In the bi-temporal setting (Tables 4.5 and 4.6), deep learning models outperformed shallow methods. MobileNetV2, a lightweight CNN, achieved the highest accuracy (94.72%) in detection and the lowest MAE (0.5456) in regression. An interesting finding was that smaller deep learning models (MobileNetV2, MiT-B0) performed better than their larger counterparts (ConvNextV2, MobileViTV2), suggesting that for this specific SAR data, increased model capacity does not necessarily lead to better performance.

Ablation Studies

The experiment with a single post-image serves as an ablation study on the importance of temporal information. The results unequivocally show that providing change information is crucial, as performance drops dramatically (e.g., over 20% in accuracy for MobileNetV2, comparable to a Random Forest on bi-temporal data.) when models are restricted to a single image. For both earthquake detection and magnitude regression, the bi-temporal approach provides better results, highlighting the importance of temporal change information.

Furthermore, a preliminary experiment was conducted to assess the benefit of integrating external domain knowledge. A global seismic hazard map was added as an extra input channel to the bi-temporal images. In this test, we do not get any performance improvements, suggesting that either the hazard map information was too coarse or that a more sophisticated method of feature fusion is required.

Discussion

This section advanced the frontier of automated earthquake monitoring by creating QuakeSet, the first large-scale benchmark for this task using SAR imagery. Our experiments confirmed that lightweight, efficient deep learning models can achieve high performance, validating our focus on computational efficiency for practical, real-world deployment.

The superior performance of smaller models over larger ones suggests that for detecting fine-grained changes rather than complex object semantics, a more constrained architecture may be more effective at capturing the relevant signal without overfitting to noise. This means in practice that it is feasible to deploy such models on resource-constrained systems. The integration of geophysical data and the exploration of longer temporal sequences are avenues for future research.

4.2.4 Section Summary

In this section, we introduced QuakeSet, a novel dataset for earthquake monitoring from satellite. We detailed its creation and defined a set of benchmark tasks. The experiments demonstrate that low-resource deep learning models can achieve high performance on earthquake detection and magnitude regression tasks with bi-temporal data. Our findings establish a strong baseline for future research in this domain and confirm the feasibility of using automated SAR image analysis for disaster management.

Together, our work on wildfires and earthquakes establishes a robust methodology for assessing a disaster's physical footprint. However, this is only one dimension of actionable intelligence. A truly holistic knowledge requires understanding the evolving human and societal impact. Therefore, we now move from the spatio-temporal analysis of the physical world to the analysis of human-generated data

streams. The next section tackles the challenge of algorithmic efficiency, introducing a novel reinforcement learning framework to synthesize chaotic textual data into a coherent summary for decision-makers in near real-time.

4.3 Crisis Evolution

Automating the extraction of actionable intelligence from vast, heterogeneous data streams offers critical support to disaster-response personnel [103, 154]. The generation of coherent event timelines from noisy and time-evolving data, such as social media and news feeds, has been formalized in benchmarks like the CrisisFACTS task [109]. The challenges are the high degree of redundancy across multiple data sources, the complexity of satisfying numerous information requests from responders simultaneously, and the general lack of human-annotated data to guide supervised models.

The state-of-the-art methodologies for this task mainly rely on two main paradigms: end-to-end dense retrieval models like ColBERT or a multi-stage Retrieve & Rerank pipeline [84, 160, 125]. This last approach first retrieves a broad set of documents with efficient algorithms (e.g., BM25) and then refines the selection with neural re-rankers [65, 84, 42]. While effective, they share fundamental limitations. They are inefficient since their processing time scales linearly with the number of input queries. Additionally, they are not designed for online processing of continuous data streams and must manage content overlap in a separate filtering stage, which leads to the retrieval of much redundant information.

This landscape highlights the need for a solution that can concurrently process multiple queries in an online setting while addressing redundancy at an early stage. To address this gap, *DQNC2S: DQN-based Cross-stream Crisis event Summarizer* [135] introduces a novel framework, DQNC2S, which frames the task of multi-stream timeline generation using Deep Reinforcement Learning (DRL). By leveraging a Deep Q-Network [114], this approach retrieves relevant information on-the-fly without relying on pre-built data indexes, making it suitable for real-time applications. Additionally, the system is designed to have an inference time independent of the number of queries. It also integrates a redundancy filter directly into the agent's reward function, addressing by design the need for non-redundant data. This model achieves a new state-of-the-art performance in a more scalable way on the CrisisFACTS benchmark.

4.3.1 Methodology

The DQNC2S framework is composed of a three-step process: (1) an initial weak annotation stage to do the training process, (2) the DQN-based agent for online text retrieval, and (3) a final topic modeling and abstraction phase to refine

the final summary.

Problem Statement

For a given event e from a set of crises E, we consider a collection of textual data streams S (e.g., news, social media) over a time period T. The data from these streams consists of timestamped documents te_s^t. The summarization process is guided by a set of queries, Q_e, which represent the specific informational needs of emergency response personnel. The objective is to generate a daily crisis timeline for event e. Each daily entry in this timeline presents a ranked list of key "facts" relevant to the queries in Q_e. These facts are supported by one or more source documents te_s^t from the original streams, ordered by importance.

Weak Annotation for Training

To guide the training of the DQN agent without requiring manual labels, we employ a weak annotation strategy. This process uses established extractive Question Answering (QA) models (i.e, Electra [38] and LongFormer [15]) to generate an initial importance score for each piece of text. For a given crisis event e, each text is paired with every event-specific query $q \in Q_e$. A confidence score (CF) is generated by the QA models for each pair. A query-text tuple is considered valid only if both models provide an answer with a confidence level exceeding 80%. Each text is then assigned a final score, Sc, equal to the number of unique queries it successfully answers, providing a proxy for its relevance.

Model Architecture: DQN-based Text Retrieval

The core is a text retrieval agent based on a Deep Q-Network (DQN). The agent interacts with the stream of incoming texts in an online manner, making a decision for each one. The action space is binary: $A = 0$ (keep the text) or $A = 1$ (discard the text). The observation space (state) is a 770-dimensional vector comprising the 768-dimensional BERT embedding of the current text, the percentage of the text selection budget remaining (P_t), and the maximum cosine similarity (Si_m) between the current text and all previously kept texts.

The agent's learning is guided by a reward function R designed to promote the selection of relevant yet non-redundant content. The function is defined as:

$$R = \begin{cases} -5 & \text{if } Sc = 0 \wedge A = 0 \\ 1 & \text{if } Sc = 0 \wedge A = 1 \\ N_{Sc}(1 - Si_m) & \text{if } Sc > 0 \wedge A = 0 \\ -N_{Sc}(1 - Si_m) & \text{if } Sc > 0 \wedge A = 1 \end{cases} \tag{4.1}$$

where N_{Sc} is the text score Sc normalized by the total number of queries for the event, $|Q_e|$. This structure penalizes the agent for keeping an irrelevant text ($Sc = 0$) and rewards it for discarding it. Conversely, for a relevant text ($Sc > 0$), the reward is proportional to its relevance (N_{Sc}) but discounted by its similarity to already selected content ($1 - Si_m$), explicitly discouraging redundancy. An example process is shown in Figure 4.6.

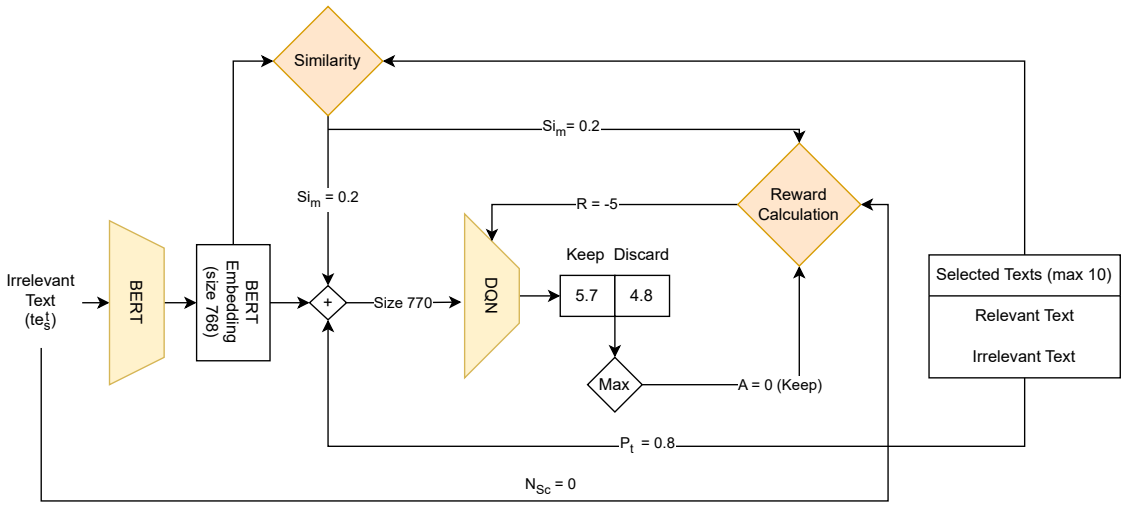


Figure 4.6: Application of DQN agent to an irrelevant text.

Training and Timeline Generation

The DQN agent is trained using the weakly annotated data. The Q-Network itself consists of a pre-trained mpnet-base-v2 SentenceBERT [140] model followed by three linear layers. During testing, the trained agent processes texts for a given day. The selected texts can then be passed through optional post-processing. We explore using BERTopic [64] for clustering the retrieved content into coherent facts and BART-CNN [93] for abstractive summarization to rephrase the content as shown in Figure 4.7. The final importance score for a text within a fact is calculated as the difference between the Q-value for keeping it and the Q-value for discarding it, reflecting the agent's confidence in its decision.

4.3.2 Experimental Setup

Datasets

We evaluate our method on the official CrisisFACTS 2022 [108] dataset. This benchmark consists of eight crisis events, each described by multiple data streams

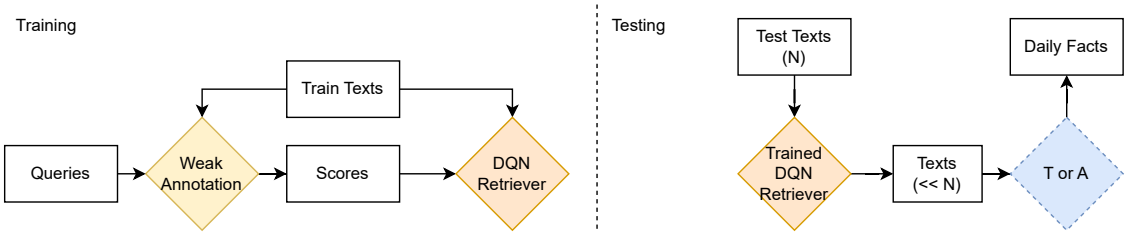


Figure 4.7: Framework during training on the left and testing on the right. During testing, topic modeling (T) or abstraction (A) are optional.

(Twitter, Reddit, News) over several days as detailed in Table 4.7. The dataset provides ground truth summaries for evaluation, which are extracted from Wikipedia, the ICS-209 All-Hazards Dataset, and annotations from NIST assessors.

Event ID	Name	Queries	Texts	Days
001	Lilac Wildfire 2017	52	45578	9
002	Cranston Wildfire 2018	52	25172	6
003	Holy Wildfire 2018	52	25482	6
004	Hurricane Florence 2018	51	180286	15
005	Maryland Flood 2018	48	37598	4
006	Saddleridge Wildfire 2019	52	34480	4
007	Hurricane Laura 2020	51	52561	2
008	Hurricane Sally 2020	51	67632	8

Table 4.7: CrisisFACTS 2022 events

Baselines and Comparative Methods

Our approach is compared against the top-performing systems from the CrisisFACTS 2022 challenge, specifically *unicamp* and *ohmkiz*, both of which are based on the Retrieve & Re-rank paradigm. We also include a ColBERT-based end-to-end model and the official baselines provided by the challenge organizers for a comprehensive comparison.

Evaluation Metrics

As done in the official challenge, we evaluate performance using Rouge-2 F1-Score and BERT-Score between the generated summaries and the ground truth. To assess the efficiency of our approach, we also measure and compare the inference execution time.

4.3.3 Results and Discussion

Quantitative Analysis

The results in Table 4.8 demonstrate the effectiveness of our approach. The DQNC2S model and its variants achieve the highest BERT-Score across all summary types, indicating superior semantic coherence with the reference. The combined model, DQNC2S-T+A (with topic modeling and abstraction), emerges as the best-performing method on average, yielding a mean ROUGE-2 score of 0.0796 and a BERT-Score of 0.5325. The mean rank also shows the full extractive solution DQNC2S is more capable than the strongest baseline.

DQNC2S demonstrates a significant advantage in efficiency, since the agent makes a decision in approximately 0.03 seconds per text, regardless of the number of queries. In contrast, the best competitors require at least $N \cdot 0.0752$ (unicamp) and $N \cdot 0.0293$ (ohmkiz).

	ICS		NIST		Wikipedia		Mean
Method	R2	BS	R2	BS	R2	BS	Rank
baseline.run1	0.0418	0.4432	0.1326	0.5565	0.0281	0.5296	2.4167
baseline.run2	0.0428	0.4427	0.1308	0.5565	0.0281	0.5274	2.0000
ohm_kiz.ColBERT	0.0497	0.4500	0.1386	0.5460	0.0307	0.5423	4.6667
ohm_kiz.QACrisis	0.0464	0.4432	0.1471	0.5642	0.0337	0.5448	6.0833
ohm_kiz.QAasnq	0.0507	0.4477	0.1468	0.5628	0.0362	0.5646	7.1667
unicamp.NM2	**0.0581**	0.4591	0.1338	0.5573	0.0281	0.5321	5.6667
unicamp.NM1	**0.0581**	0.4591	0.1338	0.5573	0.0281	0.5321	5.6667
DQNC2S	0.0406	0.4554	**0.1540**	**0.5715**	0.0402	0.5516	7.6667
DQNC2S-T	0.0513	0.4579	0.1450	0.5667	0.0317	0.5538	7.5833
DQNC2S-A	0.0412	**0.4596**	0.1538	0.5706	0.0394	0.5538	8.2500
DQNC2S-T+A	0.0452	0.4560	0.1515	0.5709	**0.0420**	**0.5707**	**8.8333**

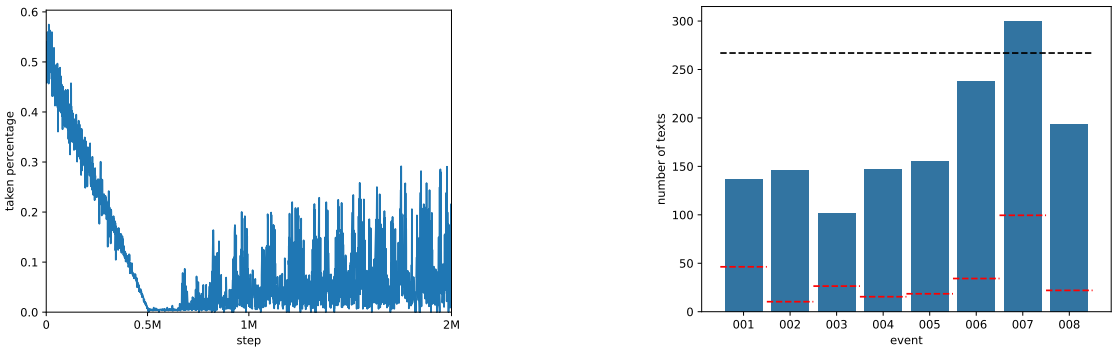
Table 4.8: Comparison of mean Rouge-2 (R2) and BERT-Score (BS)

Model Behavior Analysis

The analysis of the training process (Figure 4.8a) shows that the agent initially explores by taking a high percentage of texts, then quickly learns to be more selective, stabilizing after approximately 500,000 steps.

At inference time, the model adapts the number of selected texts to the event's need, rarely using the whole budget of 300 texts. The number of texts retrieved by the agent shows a positive correlation (0.74) with the number of facts in the ground

truth, indicating that the model has effectively learned to identify the appropriate volume of information for each day.



(a) Mean percentage of taken

(b) Mean number of selected text

Figure 4.8: Mean percentage of "take" action per episode during training (a) and the mean number of retrieved text per event (b). The red and black lines are the mean value and the maximum number of daily NIST facts, respectively.

Ablation Studies

The different variants of our model serve as an ablation study on the post-processing components. The base DQNC2S model already outperforms most baselines, highlighting the strength of the retrieval mechanism. The introduction of topic modeling (DQNC2S-T) and abstraction (DQNC2S-A) individually provide further improvements. The best results are achieved by DQNC2S-T+A, which combines both, demonstrating that clustering retrieved texts into facts and then summarizing them is the most effective pipeline for generating crisis timeline summaries.

Discussion

The results confirm that our framework is highly effective for cross-stream crisis summarization. By framing the task as a sequential decision-making problem, the DQN agent learns a robust policy to simultaneously optimize for relevance and non-redundancy in an online setting. The superior performance on BERT-Score suggests that the model excels at capturing the semantic meaning. The independence of inference time from the number of queries is a practical advantage over traditional multi-stage pipelines, particularly in real-time scenarios.

4.3.4 Section Summary

In this section, we introduced DQNC2S, a novel DRL-based framework that frames crisis timeline generation as a sequential decision problem. Our experiments on the CrisisFACTS benchmark demonstrated that this approach achieves state-of-the-art summary quality while uniquely providing constant-time inference regardless of the number of active queries. This contribution addresses the critical challenge of algorithmic efficiency, a necessary component for any truly scalable, real-time crisis intelligence system. By doing so, it complements our previous solutions for data efficiency (Magnifier) and computational efficiency (QuakeSet models), completing our portfolio of specialized tools for crisis management.

4.4 Chapter Summary

In this chapter, we presented contributions in the domain of crisis management. We addressed data gaps in post-disaster assessment by creating two foundational benchmark datasets: CaBuAr for wildfire analysis and QuakeSet for all-weather earthquake monitoring. These resources enable more robust and reproducible research. Additionally, we tackled the challenge of information overload by developing DQNC2S, a novel and highly scalable reinforcement learning framework for summarizing evolving crisis events from multi-stream textual data.

A unifying thread connects these diverse contributions: the persistent challenge of efficiency, which we have shown is multifaceted and not a monolithic problem. We began by tackling data efficiency, designing an architecture to achieve superior performance on limited labeled data. We then addressed computational efficiency with lightweight models suitable for resource-constrained environments. Finally, we worked on algorithmic efficiency, creating a near real-time agent.

This portfolio of solutions, tailored to the acute, short-term shocks of crisis management, provides a robust validation of our central thesis. With these empirical results in hand, we are now positioned to synthesize the lessons learned from both long-term ecological monitoring and short-term crisis response. The concluding chapter will discuss the broader implications of this specialized, efficiency-driven approach for the future of planetary-scale machine learning and outline a path forward for building more resilient, sustainable, and data-informed systems.

Chapter 5

Conclusion

This dissertation has confronted a fundamental challenge at the intersection of computer engineering, remote sensing, and data science: the effective and efficient transformation of vast, heterogeneous spatio-temporal data into actionable intelligence for ecological monitoring and crisis management. This chapter synthesizes the contributions of this research, discusses their impact, and delineates pathways for future research.

5.1 The Core Problem

The proliferation of Earth Observation (EO) satellites and real-time textual data streams has created unprecedented opportunities for monitoring our planet and responding to crises. However, the volume, velocity, and variety of these data sources overwhelm traditional analytical methods and general-purpose machine learning models. The computational cost, data requirements, and lack of domain specificity in existing models create a significant bottleneck, preventing the conversion of raw data into the insights needed by ecologists, first responders, and policymakers.

This thesis suggests that the development of specialized, computationally efficient spatio-temporal machine learning architectures is essential to bridge these gaps. By designing models tailored to the unique properties of environmental and disaster-related data, it is possible to unlock the full potential of the data itself, enabling robust monitoring and rapid, evidence-based decision-making.

5.2 Key Findings

The body of work is a framework built upon two pillars: the creation of foundational, large-scale datasets and the design of novel, efficient model architectures.

First, this research addressed data scarcity by establishing benchmarks for modeling spatio-temporal phenomena across different timescales. For analyzing gradual, long-term change, the HydroChronos dataset provided a new standard for surface water forecasting. To enable the assessment of sudden changes, the CaBuAr and QuakeSet datasets delivered the first comprehensive resources for wildfire and earthquake damage assessment. Finally, bridging both domains, the CrisisLandMark dataset established a new benchmark for query-based retrieval, unifying the human language with the complex nature of satellite data. These datasets have not only enabled the research presented in this thesis but also serve as resources for the scientific community, fostering reproducible and comparative research.

Building upon this empirical foundation, this thesis introduced a suite of specialized and computationally efficient architectures, each engineered to solve a distinct problem in spatio-temporal analysis. The U-KAN architecture demonstrated a significant leap in efficiency for crop field segmentation by integrating Kolmogorov-Arnold Networks into a U-Net structure. The Depth Any Canopy model provided a novel, low-cost solution for estimating forest canopy height. To bridge the gap between satellite data and textual queries, the CLOSP framework was developed for cross-modal, text-based retrieval of multi-sensor satellite imagery. Finally, the AquaClimaTempo UNet (ACTU), developed with HydroChronos, set a new state-of-the-art in surface water forecasting. The Magnifier architecture was proposed as a data-efficient solution for burnt area mapping. Lastly, to process high-velocity textual information, the DQNC2S framework utilized deep reinforcement learning to perform real-time, query-focused summarization of crisis events.

All together, they demonstrate a systematic approach to transform raw spatio-temporal data (from satellite pixels to text streams) into actionable intelligence. Whether segmenting agricultural land, forecasting water levels, mapping burn scars, or summarizing evolving events, each component highlights that specialization and efficiency are necessary for building impactful real-world systems.

5.3 Technical Contributions

The primary technical and scientific contributions of this dissertation are threefold:

- Architectural Innovation in Spatio-Temporal Learning: we introduced several novel deep learning models (U-KAN, Depth Any Canopy, ACTU, Magnifier) that are not only performant but are also designed with computational efficiency as a core principle. These models demonstrate how domain-specific architectural modifications can provide large performance gains over general-purpose models.

- Creation of Foundational Benchmark Datasets: This thesis enriched the research landscape by contributing four large-scale, publicly available datasets (HydroChronos, CrisisLandMark, CaBuAr, QuakeSet). These resources address critical gaps in ecological and crisis informatics, providing the high-quality, curated data necessary to train and benchmark future generations of models.

- Advancement of Multi-Modal and Real-Time Frameworks: This research pushed the boundaries of data fusion and processing with the CLOSP framework for text-to-image satellite retrieval and the DQNC2S framework for real-time text summarization. These contributions propose new methods for combining different data types and for processing information streams under latency constraints.

5.4 Practical Implications

The methodologies developed in this thesis provide a comprehensive toolkit for decision-makers managing different planetary systems in dynamic equilibrium ecosystems or a rapidly evolving crisis. For stakeholders focused on long-term sustainability, models like U-KAN and ACTU offer scalable tools for precision agriculture and water resource management. For those responding to immediate events, systems trained on QuakeSet and CaBuAr can rapidly map damage to guide resource allocation. Additionally, frameworks like CLOSP can be useful to both groups, allowing any analyst to instantly retrieve relevant satellite imagery with simple text queries, breaking down the barriers between complex EO data and operational needs.

5.5 Limitations and Scope

To currently frame the contributions, it is essential to acknowledge the thesis's limitations and define its scope.

First, the models developed were validated using datasets from specific geographic regions and sensor types. Although the datasets are large and diverse, the generalization performance of the models to new, unseen geographies or to data from different satellite constellations (e.g., commercial high-resolution satellites) requires additional investigation. The Depth Any Canopy model, while effective with limited data, still relies on the availability of a specific data modality.

Second, while computational efficiency was a core design principle, this research focused on algorithmic efficiency at the model level. The engineering challenges associated with deploying these models in a global-scale operational system (including

data ingestion pipelines, continuous model monitoring, and inference infrastructure) were outside the scope of this work.

Finally, the datasets, while meticulously curated, represent a sanitized version of the real world. Operational systems must deal with a higher degree of noise, including sensor artifacts, atmospheric interference, and the ambiguity inherent in unstructured text, which may require additional pre-processing techniques.

5.6 Future Works

This research opens several promising avenues for future research. The following directions are proposed:

- Integrated Multi-Modal Systems: A fusion of the text-retrieval and image-analysis models. An integrated system could leverage the CLOSP framework to allow a user to submit a natural language query. The system will first retrieve the relevant satellite imagery and then automatically apply a segmentation model to delineate the areas of interest.
- Transition to Operational Deployment: Future work should focus on the engineering challenges of moving these models from research prototypes to operational tools. This involves developing user-friendly interfaces for users and integrating these models into existing Geographic Information System (GIS) platforms.
- Enhancement of Algorithmic Capabilities: The DQNC2S framework for text summarization could be extended to a multi-modal context, incorporating both text and image streams from social media to generate richer, more comprehensive situational summaries during disasters. Furthermore, the forecasting capabilities of the ACTU model could be enhanced by exploring more sophisticated temporal architectures, potentially enabling longer-term and more accurate predictions of surface water dynamics.

5.7 Concluding Statement

This dissertation has demonstrated that the path to unlocking the value of spatio-temporal data lies not in the application of monolithic, general-purpose algorithms, but in the design of specialized, efficient, and domain-aware machine learning systems. By engineering novel architectures to effectively analyze planetary systems through both the slow evolution of ecosystems and the abrupt chaos of crises, this research provides a unified framework and a robust set of tools for converting complex data into actionable intelligence. As our planet faces increasing pressures, the development of such integrated systems is a critical necessity for

building a more resilient and sustainable future. As our planet faces additional environmental pressures and the frequency of natural disasters is increasing, the development of new systems is a critical necessity for building a more resilient and sustainable future.

Bibliography

[1] John T. Abatzoglou et al. "TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015". In: *Scientific Data* 5.1 (Jan. 2018). ISSN: 2052-4463. DOI: 10.1038/sdata.2017.191. URL: http://dx.doi.org/10.1038/sdata.2017.191.

[2] Turgut Acikara et al. "Contribution of Social Media Analytics to Disaster Response Effectiveness: A Systematic Review of the Literature". In: *Sustainability* 15.11 (May 2023), p. 8860. ISSN: 2071-1050. DOI: 10.3390/su15118860. URL: http://dx.doi.org/10.3390/su15118860.

[3] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160.

[4] European Space Agency. *Copernicus DEM*. 2022. DOI: 10.5270/esa-c5d3d65. URL: http://dx.doi.org/10.5270/ESA-c5d3d65.

[5] European Space Agency. *S1 Products*. https://sentiwiki.copernicus.eu/web/s1-products. [Accessed 19-01-2025]. 2024.

[6] Emre Aksan et al. *A Spatio-temporal Transformer for 3D Human Motion Prediction*. 2021. arXiv: 2004.08692 [cs.CV]. URL: https://arxiv.org/abs/2004.08692.

[7] Firoj Alam, Muhammad Imran, and Ferda Ofli. "Image4act: Online social media image processing for disaster response". In: *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*. 2017, pp. 601–604.

[8] Hans-Erik Andersen, Stephen E Reutebuch, and Gerard F Schreuder. "Automated individual tree measurement through morphological analysis of a LIDAR-based canopy surface model". In: *Proc. of the 1st International Precision Forestry Symposium*. 2001, pp. 11–21.

[9] Vladimir Arnold. "On functions of three variables". In: *Proceedings of the USSR Academy of Sciences*. 1957, pp. 678–681.

[10] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information fusion* 58 (2020), pp. 82–115.

[11] Ayushi and Preetpal Kaur Buttar. “Satellite Imagery Analysis for Crop Type Segmentation Using U-Net Architecture”. In: *Procedia Computer Science* 235 (2024), pp. 3418–3427. ISSN: 1877-0509. DOI: 10.1016/j.procs.2024.04.322. URL: http://dx.doi.org/10.1016/j.procs.2024.04.322.

[12] Alexander Becker et al. “Country-wide retrieval of forest structure from optical and SAR satellite imagery with deep ensembles”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 195 (2023), pp. 269–286.

[13] Jan Behmann et al. “A review of advanced machine learning methods for the detection of biotic stress in precision crop protection”. In: *Precision Agriculture* 16.3 (Aug. 2014), pp. 239–260. ISSN: 1573-1618. DOI: 10.1007/s11119-014-9372-7. URL: http://dx.doi.org/10.1007/s11119-014-9372-7.

[14] Abdelaziz A. Belal et al. “Precision Farming Technologies to Increase Soil and Crop Productivity”. In: *Springer Water*. Springer International Publishing, 2021, pp. 117–154. ISBN: 9783030785741. DOI: 10.1007/978-3-030-78574-1_6. URL: http://dx.doi.org/10.1007/978-3-030-78574-1_6.

[15] Iz Beltagy, Matthew E. Peters, and Arman Cohan. “Longformer: The Long-Document Transformer”. In: *arXiv:2004.05150* (2020).

[16] Vitus Benson et al. “Multi-modal Learning for Geospatial Vegetation Forecasting”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 27788–27799.

[17] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. “LocalBins: Improving Depth Estimation by Learning Local Distributions”. In: *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022, pp. 480–496. ISBN: 9783031197697. DOI: 10.1007/978-3-031-19769-7_28. URL: http://dx.doi.org/10.1007/978-3-031-19769-7_28.

[18] Wu Bin et al. “A Method of Automatically Extracting Forest Fire Burned Areas Using Gf-1 Remote Sensing Images”. In: *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. 2019, pp. 9953–9955. DOI: 10.1109/IGARSS.2019.8900399.

[19] Wu Bin et al. “A Method of Automatically Extracting Forest Fire Burned Areas Using Gf-1 Remote Sensing Images”. In: *IGARSS 2019*. 2019, pp. 9953–9955.

[20] Derrick Bonafilia et al. "Sen1Floods11: A Georeferenced Dataset to Train and Test Deep Learning Flood Algorithms for Sentinel-1". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2020.

[21] István Bondár and Dmitry Storchak. "Improved location procedures at the International Seismological Centre". In: *Geophysical Journal International* 186.3 (2011), pp. 1220–1244.

[22] Carl de Boor. *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer, 1978, pp. 1–314. ISBN: 978-1-4612-6333-3.

[23] Katarzyna Borys et al. "Explainable ai in medical imaging: An overview for clinical practitioners–saliency-based xai approaches". In: *European journal of radiology* 162 (2023), p. 110787.

[24] Preetpal Kaur Buttar and Manoj Kumar Sachan. "Semantic segmentation of satellite images for crop type identification in smallholder farms". In: *The Journal of Supercomputing* 80.2 (July 2023), pp. 1367–1395. ISSN: 1573-0484. DOI: 10.1007/s11227-023-05533-4. URL: http://dx.doi.org/10.1007/s11227-023-05533-4.

[25] Daniele Rege Cambrin, Luca Colomba, and Paolo Garza. "CaBuAr: California burned areas dataset for delineation [Software and Data Sets]". In: *IEEE Geoscience and Remote Sensing Magazine* 11.3 (2023), pp. 106–113. DOI: 10.1109/MGRS.2023.3292467.

[26] Daniele Rege Cambrin, Luca Colomba, and Paolo Garza. "Magnifier: A Multigrained Neural Network-Based Architecture for Burned Area Delineation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18 (2025), pp. 12263–12277. DOI: 10.1109/JSTARS.2025.3565819.

[27] Daniele Rege Cambrin, Luca Colomba, and Paolo Garza. "Vision Transformers for Burned Area Delineation". In: *Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. 2022.

[28] Daniele Rege Cambrin et al. *HydroChronos: Forecasting Decades of Surface Water Change*. 2025. arXiv: 2506.14362 [cs.CV]. URL: https://arxiv.org/abs/2506.14362.

[29] Daniele Rege Cambrin et al. *Text-to-Remote-Sensing-Image Retrieval beyond RGB Sources*. 2025. arXiv: 2507.10403 [cs.CV]. URL: https://arxiv.org/abs/2507.10403.

[30] C Alina Cansler and Donald McKenzie. "How robust are burn severity indices when applied in a new region? Evaluation of alternate field-based and remote-sensing methods". In: *Remote sensing* 4.2 (2012), pp. 456–483.

[31] Chanthujan Chandrakumar et al. “Algorithms for Detecting P-Waves and Earthquake Magnitude Estimation: Initial Literature Review Findings”. In: (2023).

[32] Liang-Chieh Chen et al. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (2018), pp. 834–848. DOI: 10.1109/TPAMI.2017.2699184.

[33] Liang-Chieh Chen et al. “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”. en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 833–851. ISBN: 978-3-030-01234-2. DOI: 10.1007/978-3-030-01234-2_49.

[34] Liang-Chieh Chen et al. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017. DOI: 10.48550/ARXIV.1706.05587. URL: https://arxiv.org/abs/1706.05587.

[35] Ting Chen et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html.

[36] Kyunghyun Cho et al. “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches”. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Ed. by Dekai Wu et al. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. DOI: 10.3115/v1/W14-4012. URL: https://aclanthology.org/W14-4012/.

[37] Özgün Çiçek et al. “3D U-Net: learning dense volumetric segmentation from sparse annotation”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 424–432.

[38] Kevin Clark et al. “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators”. In: *ICLR*. 2020. URL: https://openreview.net/pdf?id=r1xMH1BtvB.

[39] Kai Norman Clasen et al. “reBEN: Refined BigEarthNet Dataset for Remote Sensing Image Analysis”. In: (2024). arXiv: 2407.03653 [cs.CV]. URL: https://arxiv.org/abs/2407.03653.

[40] Luca Colomba et al. “A Dataset for Burned Area Delineation and Severity Estimation from Satellite Imagery”. In: *CIKM2022*. CIKM ’22. Atlanta, GA, USA: ACM, 2022, pp. 3893–3897. ISBN: 9781450392365. DOI: 10.1145/3511808.3557528.

[41] Isaac Corley, Caleb Robinson, and Anthony Ortiz. "A Change Detection Reality Check". In: *arXiv preprint arXiv:2402.06994* (2024).

[42] Zhuyun Dai et al. "Convolutional Neural Networks for Soft-Matching N-Grams in Ad-Hoc Search". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, pp. 126–134. ISBN: 9781450355810. DOI: 10.1145/3159652.3159659. URL: https://doi.org/10.1145/3159652.3159659.

[43] Feiyue Deng et al. "A Novel Combination Neural Network Based on ConvLSTM-Transformer for Bearing Remaining Useful Life Prediction". In: *Machines* 10.12 (Dec. 2022), p. 1226. ISSN: 2075-1702. DOI: 10.3390/machines10121226. URL: http://dx.doi.org/10.3390/machines10121226.

[44] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[45] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[46] Domenico Di Giacomo and Dmitry A Storchak. "A scheme to set preferred magnitudes in the ISC Bulletin". In: *Journal of Seismology* 20 (2016), pp. 555–567.

[47] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[48] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2020. DOI: 10.48550/ARXIV.2010.11929. URL: https://arxiv.org/abs/2010.11929.

[49] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[50] Elias Dritsas and Maria Trigka. "Remote Sensing and Geospatial Analysis in the Big Data Era: A Survey". In: *Remote Sensing* 17.3 (Feb. 2025), p. 550. ISSN: 2072-4292. DOI: 10.3390/rs17030550. URL: http://dx.doi.org/10.3390/rs17030550.

[51] Matthias Drusch et al. "Sentinel-2: ESA's optical high-resolution mission for GMES operational services". In: *Remote sensing of Environment* 120 (2012), pp. 25–36.

[52] Ralph Dubayah et al. “GEDI launches a new era of biomass inference from space”. In: *Environmental Research Letters* 17.9 (2022), p. 095001.

[53] Jean-luc Dupuy et al. “Climate change impact on future wildfire danger and activity in southern Europe: a review”. In: *Annals of Forest Science* 77.2 (2020), pp. 1–24. DOI: 10.1007/s13595-020-00933-5.

[54] David Eigen, Christian Puhrsch, and Rob Fergus. “Depth map prediction from a single image using a multi-scale deep network”. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’14. Montreal, Canada: MIT Press, 2014, pp. 2366–2374.

[55] Alessandro Farasin et al. “Supervised Burned Areas delineation by means of Sentinel-2 imagery and Convolutional Neural Networks”. In: *Proceedings of the 17th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2020)*. 2020, pp. 24–27.

[56] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. “AdaBins: Depth Estimation Using Adaptive Bins”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 4008–4017. DOI: 10.1109/CVPR46437.2021.00400.

[57] Federico Filipponi. “BAIS2: Burned Area Index for Sentinel-2”. In: *Proceedings* 2.7 (2018). ISSN: 2504-3900. DOI: 10.3390/ecrs-2-05177. URL: https://www.mdpi.com/2504-3900/2/7/364.

[58] Gareth J Funning and Astrid Garcia. “A systematic study of earthquake detectability using Sentinel-1 Interferometric Wide-Swath data”. In: *Geophysical Journal International* 216.1 (2019), pp. 332–349.

[59] Bo-Cai Gao. “NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space”. In: *Remote sensing of environment* 58.3 (1996), pp. 257–266.

[60] David LA Gaveau and Ross A Hill. “Quantifying canopy height underestimation by laser pulse penetration in small-footprint airborne laser scanning data”. In: *Canadian Journal of Remote Sensing* 29.5 (2003), pp. 650–657.

[61] Caroline M Gevaert. “Explainable AI for earth observation: A review including societal and regulatory perspectives”. In: *International Journal of Applied Earth Observation and Geoinformation* 112 (2022), p. 102869.

[62] Rafik Ghali et al. “Wildfire Segmentation Using Deep Vision Transformers”. In: *Remote Sensing* 13.17 (2021), p. 3527. ISSN: 2072-4292.

[63] Samuel N Goward et al. “The Landsat 7 mission: Terrestrial research and applications for the 21st century”. In: *Remote Sensing of Environment* 78.1-2 (2001), pp. 3–12.

[64] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022).

[65] Jiafeng Guo et al. "A Deep Relevance Matching Model for Ad-Hoc Retrieval". In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM '16. Indianapolis, Indiana, USA: Association for Computing Machinery, 2016, pp. 55–64. ISBN: 9781450340731. DOI: 10.1145/2983323.2983769. URL: https://doi.org/10.1145/2983323.2983769.

[66] R. Hadsell, S. Chopra, and Y. LeCun. "Dimensionality Reduction by Learning an Invariant Mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.

[67] Jessica E Halofsky, David L Peterson, and Brian J Harvey. "Changing wildfire, changing forests: the effects of climate change on fire regimes and vegetation in the Pacific Northwest, USA". In: *Fire Ecology* 16.1 (2020), pp. 1–26. DOI: 10.1186/s42408-019-0062-8. URL: https://doi.org/10.1186/s42408-019-0062-8.

[68] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 9726–9735. DOI: 10.1109/CVPR42600.2020.00975.

[69] Patrick Helber et al. "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.7 (2019), pp. 2217–2226.

[70] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[71] Andrew Howard et al. "Searching for MobileNetV3". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2019, pp. 1314–1324. DOI: 10.1109/ICCV.2019.00140.

[72] Andrew Howard et al. *Searching for MobileNetV3*. 2019. arXiv: 1905.02244 [cs.CV]. URL: https://arxiv.org/abs/1905.02244.

[73] Andrew G Howard et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861* (2017).

[74] Huimin Huang et al. "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation". In: *ICASSP 2020*. 2020, pp. 1055–1059.

[75] Zhentao Huang et al. "DSCNN-LSTMs: A Lightweight and Efficient Model for Epilepsy Recognition". In: *Brain Sciences* 12.12 (Dec. 2022), p. 1672. ISSN: 2076-3425. DOI: 10.3390/brainsci12121672. URL: http://dx.doi.org/10.3390/brainsci12121672.

[76] Alfredo R Huete. "A soil-adjusted vegetation index (SAVI)". In: *Remote sensing of environment* 25.3 (1988), pp. 295–309.

[77] Andrei Iantsen, Dimitris Visvikis, and Mathieu Hatt. "Squeeze-and-Excitation Normalization for Automated Delineation of Head and Neck Primary Tumors in Combined PET and CT Images". In: *Lecture Notes in Computer Science*. Springer International Publishing, 2021, pp. 37–43. ISBN: 9783030671945. DOI: 10.1007/978-3-030-67194-5_4. URL: http://dx.doi.org/10.1007/978-3-030-67194-5_4.

[78] Gabriel Ilharco et al. *OpenCLIP*. Version 0.1. If you use this software, please cite it as below. July 2021. DOI: 10.5281/zenodo.5143773. URL: https://doi.org/10.5281/zenodo.5143773.

[79] *International Seismological Centre (2023), On-line Bulletin*. 2023. DOI: https://doi.org/10.31905/D808B830.

[80] Pallavi Jain et al. *SenCLIP: Enhancing zero-shot land-use mapping for Sentinel-2 with ground-level prompting*. 2024. arXiv: 2412.08536 [cs.CV]. URL: https://arxiv.org/abs/2412.08536.

[81] Ioannis Kakogeorgiou and Konstantinos Karantzalos. "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing". In: *International Journal of Applied Earth Observation and Geoinformation* 103 (2021), p. 102520.

[82] Vladimir Karpukhin et al. "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. URL: https://aclanthology.org/2020.emnlp-main.550.

[83] Bingxin Ke et al. "Marigold: Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025), pp. 1–18. DOI: 10.1109/TPAMI.2025.3591076.

[84] Omar Khattab and Matei Zaharia. "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT". In: SIGIR '20. Virtual Event, China: Association for Computing Machinery, 2020, pp. 39–48. ISBN: 9781450380164. DOI: 10.1145/3397271.3401075. URL: https://doi.org/10.1145/3397271.3401075.

[85] Konstantin Klemmer et al. *SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery*. 2024. arXiv: 2311.17179 [cs.CV]. URL: https://arxiv.org/abs/2311.17179.

[86] Lisa Knopp et al. "A Deep Learning Approach for Burned Area Segmentation with Sentinel-2 Data". In: *Remote Sensing* 12.15 (2020). ISSN: 2072-4292. DOI: 10.3390/rs12152422. URL: https://www.mdpi.com/2072-4292/12/15/2422.

[87] Andrej Nikolaevich Kolmogorov. "On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables". In: *Proceedings of the USSR Academy of Sciences*. 1956.

[88] Sven Kruschel et al. "Challenging the Performance-Interpretability Trade-Off: An Evaluation of Interpretable Machine Learning Models". In: *Business & Information Systems Engineering* (Feb. 2025). ISSN: 1867-0202. DOI: 10.1007/s12599-024-00922-2. URL: http://dx.doi.org/10.1007/s12599-024-00922-2.

[89] Nico Lang et al. "A high-resolution canopy height model of the Earth". In: *Nature Ecology & Evolution* 7.11 (2023), pp. 1778–1789.

[90] Katrin Lasinger et al. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer". In: *arXiv preprint arXiv:1907.01341* (2019).

[91] David Lazer et al. "Computational Social Science". In: *Science* 323.5915 (Feb. 2009), pp. 721–723. ISSN: 1095-9203. DOI: 10.1126/science.1167742. URL: http://dx.doi.org/10.1126/science.1167742.

[92] Bernhard Lehner and Günther Grill. "Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems". In: *Hydrological Processes* 27.15 (Apr. 2013), pp. 2171–2186. ISSN: 1099-1085. DOI: 10.1002/hyp.9740. URL: http://dx.doi.org/10.1002/hyp.9740.

[93] Mike Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *CoRR* abs/1910.13461 (2019). arXiv: 1910.13461. URL: http://arxiv.org/abs/1910.13461.

[94] Chenxin Li et al. "U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation". In: *arXiv preprint arXiv:2406.02918* (2024).

[95] Mengya Li et al. "A Deep Learning Method of Water Body Extraction From High Resolution Remote Sensing Images With Multisensors". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), pp. 3120–3132.

[96] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[97] Fan Liu et al. “Remoteclip: A vision language foundation model for remote sensing”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[98] Heyi Liu, Shanyou Li, and Jindong Song. “Discrimination between earthquake p waves and microtremors via a generative adversarial network”. In: *Bulletin of the Seismological Society of America* 112.2 (2022), pp. 669–679.

[99] Ze Liu et al. “Swin Transformer V2: Scaling Up Capacity and Resolution”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 12009–12019.

[100] Ze Liu et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 10012–10022.

[101] Zhuang Liu et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986.

[102] Ziming Liu et al. *KAN: Kolmogorov-Arnold Networks*. 2024. arXiv: 2404.19756 [cs.LG]. URL: https://arxiv.org/abs/2404.19756.

[103] Valerio Lorini et al. “Social media for emergency management: Opportunities and challenges at the intersection of research and practice”. In: *18th international conference on information systems for crisis response and management*. 2021, pp. 772–777.

[104] Magdalena Main-Knorn et al. “Sen2Cor for sentinel-2”. In: *Image and Signal Processing for Remote Sensing XXIII*. Vol. 10427. SPIE. 2017, pp. 37–48.

[105] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008. ISBN: 9780511809071. DOI: 10.1017/cbo9780511809071. URL: http://dx.doi.org/10.1017/CBO9780511809071.

[106] M Pilar Martín, Israel Gómez, and Emilio Chuvieco. “Burnt Area Index (BAIM) for burned area discrimination at regional scale using MODIS data”. In: *Forest Ecology and Management* 234 (2006), S221.

[107] Maxar. *Maxar’s Open Data Program*. License: BY-NC-4.0.

[108] Richard McCreadie and Cody Buntain. “CrisisFACTS: Buidling and Evaluating Crisis Timelines”. In: (2023), pp. 320–339.

[109] Richard McCreadie and Cody Buntain. "CrisisFacts: building and evaluating crisis timelines". In: *Proceedings of the 20th International Conference on Information Systems for Crisis Response and Management*. University of Nebraska at Omaha (USA), May 2023. DOI: 10.59297/jvqz9405. URL: http://dx.doi.org/10.59297/JVQZ9405.

[110] Sachin Mehta and Mohammad Rastegari. "Separable self-attention for mobile vision transformers". In: *arXiv preprint arXiv:2206.02680* (2022).

[111] Mathis Loïc Messager et al. "Estimating the volume and age of water stored in global lakes using a geo-statistical approach". In: *Nature Communications* 7.1 (Dec. 2016). ISSN: 2041-1723. DOI: 10.1038/ncomms13603. URL: http://dx.doi.org/10.1038/ncomms13603.

[112] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *2016 Fourth International Conference on 3D Vision (3DV)*. 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

[113] Yue Ming et al. "Deep learning for monocular depth estimation: A review". In: *Neurocomputing* 438 (May 2021), pp. 14–33. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.12.089. URL: http://dx.doi.org/10.1016/j.neucom.2020.12.089.

[114] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *nature* 518.7540 (2015), pp. 529–533.

[115] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[116] Fabio Montello, Edoardo Arnaudo, and Claudio Rossi. "MMFlood: A Multimodal Dataset for Flood Delineation From Satellite Imagery". In: *IEEE Access* 10 (2022), pp. 96774–96787. DOI: 10.1109/ACCESS.2022.3205419.

[117] S Mostafa Mousavi et al. "Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking". In: *Nature communications* 11.1 (2020), p. 3952.

[118] Dat T Nguyen et al. "Damage assessment from social media imagery data during disasters". In: *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*. 2017, pp. 569–576.

[119] Rodrigo Nogueira et al. *Multi-Stage Document Ranking with BERT*. 2019. arXiv: 1910.14424 [cs.IR]. URL: https://arxiv.org/abs/1910.14424.

[120] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV]. URL: https://arxiv.org/abs/2304.07193.

[121] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.

[122] Eliana Pastor et al. "How divergent is your data?" In: *Proceedings of the VLDB Endowment* 14.12 (2021), pp. 2835–2838.

[123] Jan Pauls et al. "Estimating Canopy Height at Scale". In: *arXiv preprint arXiv:2406.01076* (2024).

[124] Jean-François Pekel et al. "High-resolution mapping of global surface water and its long-term changes". In: *Nature* 540.7633 (Dec. 2016), pp. 418–422. ISSN: 1476-4687. DOI: 10.1038/nature20584. URL: http://dx.doi.org/10.1038/nature20584.

[125] Jayr Pereira et al. "Using Neural Reranking and GPT-3 for Social Media Disaster Content Summarization". In: ().

[126] Yudhi Prabowo et al. "Deep Learning Dataset for Estimating Burned Areas: Case Study, Indonesia". In: *Data* 7.6 (2022). ISSN: 2306-5729. DOI: 10.3390/data7060078.

[127] Yudhi Prabowo et al. "Deep learning dataset for estimating burned areas: Case study, Indonesia". In: *Data* 7.6 (2022), p. 78.

[128] Danfeng Qin et al. "MobileNetV4: Universal Models for the Mobile Ecosystem". In: *Computer Vision – ECCV 2024*. Springer Nature Switzerland, Nov. 2024, pp. 78–96. ISBN: 9783031736612. DOI: 10.1007/978-3-031-73661-2_5. URL: http://dx.doi.org/10.1007/978-3-031-73661-2_5.

[129] Bo Qu et al. "Deep semantic understanding of high resolution remote sensing image". In: *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*. 2016, pp. 1–5. DOI: 10.1109/CITS.2016.7546397.

[130] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *ICML*. 2021.

[131] Clément Rambour et al. "Flood detection in time series of optical and sar images". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 43.B2 (2020), pp. 1343–1346.

[132] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. "Vision transformers for dense prediction". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 12179–12188.

[133] Dmitry Rashkovetsky et al. "Wildfire Detection From Multisensor Satellite Imagery Using Deep Semantic Segmentation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), pp. 7001–7016. DOI: 10.1109/JSTARS.2021.3093625.

[134] Dmitry Rashkovetsky et al. “Wildfire Detection From Multisensor Satellite Imagery Using Deep Semantic Segmentation”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), pp. 7001–7016. DOI: 10.1109/JSTARS.2021.3093625.

[135] Daniele Rege Cambrin, Luca Cagliero, and Paolo Garza. “DQNC2S: DQN-Based Cross-Stream Crisis Event Summarizer”. In: *Advances in Information Retrieval*. Springer Nature Switzerland, 2024, pp. 422–430. ISBN: 9783031560637. DOI: 10.1007/978-3-031-56063-7_34. URL: http://dx.doi.org/10.1007/978-3-031-56063-7_34.

[136] Daniele Rege Cambrin, Isaac Corley, and Paolo Garza. “Depth Any Canopy: Leveraging Depth Foundation Models for Canopy Height Estimation”. In: *Computer Vision – ECCV 2024 Workshops*. Springer Nature Switzerland, 2025, pp. 71–86. ISBN: 9783031923876. DOI: 10.1007/978-3-031-92387-6_5. URL: http://dx.doi.org/10.1007/978-3-031-92387-6_5.

[137] Daniele Rege Cambrin and Paolo Garza. “QuakeSet: A Dataset and Low-Resource Models to Monitor Earthquakes through Sentinel-1”. In: *Proceedings of the International ISCRAM Conference* (May 2024). ISSN: 2411-3387. DOI: 10.59297/n89yc374. URL: http://dx.doi.org/10.59297/n89yc374.

[138] Daniele Rege Cambrin and Paolo Garza. “QuakeSet: A Dataset and Low-Resource Models to Monitor Earthquakes through Sentinel-1”. In: *Proceedings of the International ISCRAM Conference* (May 2024). ISSN: 2411-3387. DOI: 10.59297/n89yc374. URL: http://dx.doi.org/10.59297/n89yc374.

[139] Daniele Rege Cambrin et al. “KAN You See It? KANs and Sentinel for Effective and Explainable Crop Field Segmentation”. In: *European Conference on Computer Vision*. Springer. 2025, pp. 115–131.

[140] Nils Reimers and Iryna Gurevych. “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2020. URL: https://arxiv.org/abs/2004.09813.

[141] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410/.

[142] Stephen Robertson and Hugo Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends® in Information Retrieval* 3.4 (2009), pp. 333–389. ISSN: 1554-0677. DOI: 10.1561/1500000019. URL: http://dx.doi.org/10.1561/1500000019.

[143] Stephen E Robertson. "The probability ranking principle in IR". In: *Journal of documentation* 33.4 (1977), pp. 294–304.

[144] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241. ISBN: 9783319245744. DOI: 10.1007/978-3-319-24574-4_28. URL: http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[145] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Lecture Notes in Computer Science*. Springer International Publishing, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28. URL: https://doi.org/10.1007/978-3-319-24574-4_28.

[146] Ribana Roscher et al. "Explainable Machine Learning for Scientific Insights and Discoveries". In: *IEEE Access* 8 (2020), pp. 42200–42216. DOI: 10.1109/ACCESS.2020.2976199.

[147] E. Roteta et al. "Development of a Sentinel-2 burned area algorithm: Generation of a small fire database for sub-Saharan Africa". In: *Remote Sensing of Environment* 222 (2019), pp. 1–17. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2018.12.011. URL: https://www.sciencedirect.com/science/article/pii/S0034425718305649.

[148] D.P. Roy, L. Boschetti, and S.N. Trigg. "Remote sensing of fire severity: assessing the performance of the normalized burn ratio". In: *IEEE Geoscience and Remote Sensing Letters* 3.1 (2006), pp. 112–116. DOI: 10.1109/LGRS.2005.858485.

[149] David P Roy et al. "Landsat-8: Science and product vision for terrestrial global change research". In: *Remote sensing of Environment* 145 (2014), pp. 154–172.

[150] Julien Ruffault et al. "Increased likelihood of heat-induced large wildfires in the Mediterranean Basin". In: *Scientific reports* 10.1 (2020), pp. 1–9. DOI: 10.1038/s41598-020-70069-z.

[151] Sukanya S and Sabu Joseph. "Climate change impacts on water resources: An overview". In: *Visualization Techniques for Climate Change with Machine Learning and Artificial Intelligence*. Elsevier, 2023, pp. 55–76. ISBN: 9780323997140. DOI: 10.1016/b978-0-323-99714-0.00008-x. URL: http://dx.doi.org/10.1016/B978-0-323-99714-0.00008-X.

[152] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models". In: *arXiv preprint arXiv:1708.08296* (2017).

[153] Mark Sandler et al. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[154] Anita Saroj and Sukomal Pal. "Use of social media in crisis management: A survey". In: *International journal of disaster risk reduction* 48 (2020), p. 101584. URL: https://api.semanticscholar.org/CorpusID:218780990.

[155] Luigi Saulino et al. "Detecting Burn Severity across Mediterranean Forest Types by Coupling Medium-Spatial Resolution Satellite Imagery and Field Data". In: *Remote Sensing* 12.4 (2020). ISSN: 2072-4292. DOI: 10.3390/rs12040741. URL: https://www.mdpi.com/2072-4292/12/4/741.

[156] Franco Scarselli et al. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605.

[157] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

[158] Roy Schwartz et al. "Green AI". In: *Commun. ACM* 63.12 (Nov. 2020), pp. 54–63. ISSN: 0001-0782. DOI: 10.1145/3381831. URL: https://doi.org/10.1145/3381831.

[159] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. "On the stratification of multi-label data". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*. Springer. 2011, pp. 145–158.

[160] Philipp Seeberger and Korbinian Riedhammer. "Combining Deep Neural Reranking and Unsupervised Extraction for Multi-Query Focused Summarization". In: *arXiv preprint arXiv:2302.01148* (2023).

[161] Ramprasaath R Selvaraju et al. "Grad-CAM: Why did you say that?" In: *arXiv preprint arXiv:1611.07450* (2016).

[162] Lloyd S Shapley. "A Value for n-Person Games". In: *Contributions to the Theory of Games II*. Ed. by Harold W. Kuhn and Albert W. Tucker. Princeton: Princeton University Press, 1953, pp. 307–317.

[163] Shashi Shekhar et al. "Spatiotemporal Data Mining: A Computational Perspective". In: *ISPRS International Journal of Geo-Information* 4.4 (Oct. 2015), pp. 2306–2338. ISSN: 2220-9964. DOI: 10.3390/ijgi4042306. URL: http://dx.doi.org/10.3390/ijgi4042306.

[164] Xingjian Shi et al. *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting*. 2015. arXiv: 1506.04214 [cs.CV]. URL: https://arxiv.org/abs/1506.04214.

[165] Sergii Skakun et al. "Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2". In: *Remote Sensing of Environment* 274 (June 2022), p. 112990. ISSN: 0034-4257. DOI: 10.1016/j.rse.2022.112990. URL: http://dx.doi.org/10.1016/j.rse.2022.112990.

[166] Gencer Sumbul et al. "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding". In: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2019, pp. 5901–5904.

[167] Saeid Asgari Taghanaki et al. "Combo loss: Handling input and output imbalance in multi-organ segmentation". In: *Computerized Medical Imaging and Graphics* 75 (2019), pp. 24–33.

[168] Prasad S Thenkabail et al. "Accuracy assessments of hyperspectral waveband performance for vegetation analysis applications". In: *Remote sensing of environment* 91.3-4 (2004), pp. 354–376.

[169] Jamie Tolan et al. "Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar". In: *Remote Sensing of Environment* 300 (2024), p. 113888.

[170] Erkki Tomppo et al. "National forest inventories". In: *Pathways for Common Reporting. European Science Foundation* 1 (2010), pp. 541–553.

[171] Ramon Torres et al. "GMES Sentinel-1 mission". In: *Remote sensing of environment* 120 (2012), pp. 9–24.

[172] Ramón Torres et al. "The Sentinel-1 mission and its application capabilities". In: *2012 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2012, pp. 1703–1706.

[173] C UNESCO. *Recommendation on the ethics of artificial intelligence*. 2021.

[174] Ashish Vaswani et al. "Attention is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964.

[175] Diego Velazquez et al. "EarthView: A Large Scale Remote Sensing Dataset for Self-Supervision". In: *Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops*. Feb. 2025, pp. 1228–1237.

[176] Xander Wang and Lirong Liu. "The Impacts of Climate Change on the Hydrological Cycle and Water Resource Management". In: *Water* 15.13 (June 2023), p. 2342. ISSN: 2073-4441. DOI: 10.3390/w15132342. URL: http://dx.doi.org/10.3390/w15132342.

[177] Yi Wang et al. "SSL4EO-S12: A Large-Scale Multi-Modal, Multi-Temporal Dataset for Self-Supervised Learning in Earth Observation". In: *arXiv preprint arXiv:2211.07044* (2022).

[178] Yi Wang et al. "SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets]". In: *IEEE Geoscience and Remote Sensing Magazine* 11.3 (2023), pp. 98–106. DOI: 10.1109/MGRS.2023.3281651.

[179] Zhecheng Wang et al. "Skyscript: A large and semantically diverse vision-language dataset for remote sensing". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 6. 2024, pp. 5805–5813.

[180] Global Forest Watch. "Global forest watch". In: *World Resources Institute, Washington, DC Available from http://www. globalforestwatch. org (accessed March 2002)* (2002).

[181] Western Cape Department of Agriculture and Radiant Earth Foundation. *Crop type classification dataset for western cape, South Africa*. 2021.

[182] Christopher K. Wikle. "Modern perspectives on statistics for spatio-temporal data". In: *WIREs Computational Statistics* 7.1 (Nov. 2014), pp. 86–98. ISSN: 1939-0068. DOI: 10.1002/wics.1341. URL: http://dx.doi.org/10.1002/wics.1341.

[183] Sanghyun Woo et al. "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 16133–16142. DOI: 10.1109/CVPR52729.2023.01548.

[184] Sanghyun Woo et al. "Convnext v2: Co-designing and scaling convnets with masked autoencoders". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 16133–16142.

[185] Haoning Wu et al. *Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels*. 2023. arXiv: 2312.17090 [cs.CV]. URL: https://arxiv.org/abs/2312.17090.

[186] Michael A Wulder et al. "Current status of Landsat program, science, and applications". In: *Remote sensing of environment* 225 (2019), pp. 127–147.

[187] Michael A Wulder et al. "Fifty years of Landsat science and impacts". In: *Remote Sensing of Environment* 280 (2022), p. 113195.

[188] Michael A. Wulder et al. "Fifty years of Landsat science and impacts". In: *Remote Sensing of Environment* 280 (Oct. 2022), p. 113195. ISSN: 0034-4257. DOI: 10.1016/j.rse.2022.113195. URL: http://dx.doi.org/10.1016/j.rse.2022.113195.

[189] Ruan Xiaogang et al. "Monocular Depth Estimation Based on Deep Learning:A Survey". In: *2020 Chinese Automation Congress (CAC)*. 2020, pp. 2436–2440. DOI: 10.1109/CAC51589.2020.9327548.

[190] Enze Xie et al. "SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12077–12090.

[191] Hanqiu Xu. "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery". In: *International Journal of Remote Sensing* 27.14 (July 2006), pp. 3025–3033. ISSN: 1366-5901. DOI: 10.1080/01431160600589179. URL: http://dx.doi.org/10.1080/01431160600589179.

[192] Mingxing Xu et al. "Spatial-temporal transformer networks for traffic flow forecasting". In: *arXiv preprint arXiv:2001.02908* (2020).

[193] Dai Yamazaki, Mark A. Trigg, and Daiki Ikeshima. "Development of a global 90m water body map using multi-temporal Landsat images". In: *Remote Sensing of Environment* 171 (Dec. 2015), pp. 337–351. ISSN: 0034-4257. DOI: 10.1016/j.rse.2015.10.014. URL: http://dx.doi.org/10.1016/j.rse.2015.10.014.

[194] Sijie Yan, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[195] Lihe Yang et al. *Depth Anything V2*. 2024. arXiv: 2406.09414 [cs.CV]. URL: https://arxiv.org/abs/2406.09414.

[196] Michael Yeung et al. "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation". In: *Computerized Medical Imaging and Graphics* 95 (2022), p. 102026.

[197] Manzhu Yu, Qunying Huang, and Zhenlong Li. "Deep learning for spatiotemporal forecasting in Earth system science: a review". In: *International Journal of Digital Earth* 17.1 (Aug. 2024). ISSN: 1753-8955. DOI: 10.1080/17538947.2024.2391952. URL: http://dx.doi.org/10.1080/17538947.2024.2391952.

[198] Zhiqiang Yuan et al. "Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–19. DOI: 10.1109/TGRS.2021.3078451.

[199] Liheng Zhong, Lina Hu, and Hang Zhou. "Deep learning based multi-temporal crop classification". In: *Remote Sensing of Environment* 221 (Feb. 2019), pp. 430–443. ISSN: 0034-4257. DOI: 10.1016/j.rse.2018.11.032. URL: http://dx.doi.org/10.1016/j.rse.2018.11.032.

[200] Bolei Zhou et al. "Scene Parsing through ADE20K Dataset". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5122–5130.

[201] Yujia Zhou et al. *Ultron: An Ultimate Retriever on Corpus with a Model-based Indexer*. 2022. arXiv: 2208.09257 [cs.IR]. URL: https://arxiv.org/abs/2208.09257.

[202] Weiqiang Zhu et al. "An end-to-end earthquake detection method for joint phase picking and association using deep learning". In: *Journal of Geophysical Research: Solid Earth* 127.3 (2022), e2021JB023283.

[203] Xiao Xiang Zhu et al. "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources". In: *IEEE Geoscience and Remote Sensing Magazine* 5.4 (2017), pp. 8–36. DOI: 10.1109/MGRS.2017.2762307.

This Ph.D. thesis has been typeset by means of the TEX-system facilities. The typesetting engine was pdfLATEX. The document class was `toptesi`, by Claudio Beccari, with option `tipotesi=scudo`. This class is available in every up-to-date and complete TEX-system installation.