Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (36th cycle)

Semantic Image Segmentation in Remote Sensing Scenarios

Edoardo Arnaudo

* * * * *

Supervisors
Prof. B.Caputo, Supervisor
Dr. F.Dominici, Co-supervisor

Doctoral examination committee
Prof. Marco Grangetto, Referee, Università degli Studi di Torino
Prof. Minh-Tan Pham, Referee, Université Bretagne Sud
Prof. Paolo Garza, Politecnico di Torino
Dr. Dino Ienco, UMR Tetis - INRAE, Montpellier
Dr. Anastasios Dimou, CERTH, Thessaloniki

Politecnico di Torino
30th October, 2024

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.

. . . .

Edoardo Arnaudo

Turin, 30th October, 2024

Summary

Semantic segmentation is a fundamental task in computer vision that aims to assign a semantic label to each pixel in an image, parsing it into meaningful and coherent regions. Thanks to the recent progresses of deep learning in computer vision, semantic segmentation has gained significant attention due to its wide range of applications, including autonomous driving or medical image analysis. In the context of remote sensing, this task plays a crucial role in extracting valuable information from satellite and aerial imagery. Remote sensing data, such as multispectral and radar images, provide valuable information for Earth Observation purposes, from land cover, to atmospheric composition. Detailed segmentation maps can provide insights into various geospatial phenomena, supporting applications such as urban planning, environmental monitoring, disaster response, and agricultural analysis. However, segmenting aerial and satellite images presents unique challenges compared to conventional computer vision tasks on natural images. These often cover large geographic areas and exhibit high spatial and spectral variability. The objects of interest, such as buildings, roads, and vegetation, can have diverse appearances and scales, making it difficult to capture their contextual relationships. Moreover, annotating images for semantic segmentation is a time-consuming and intensive process, often requiring specific domain expertise. To address these challenges, researchers have explored various approaches to adapt and enhance segmentation techniques for remote sensing applications. These include leveraging the multi-scale and multi-modal nature of aerial data, incorporating domain-specific prior knowledge, and developing efficient annotation strategies. Additionally, the increasing availability of large-scale remote sensing datasets and advancements in deep learning architectures have enabled the training of more robust and generalizable models. This thesis directly addresses these challenges, adapting semantic segmentation solutions to varied remote sensing scenarios. We explore several techniques to effectively leverage the rich information derived from aerial and satellite sensors, tackling issues such as data scarcity and annotation costs to develop efficient and robust models. Our research contributions span multiple aspects of semantic segmentation in remote sensing, including regularization techniques, architectural changes, weakly supervised learning approaches, and domain adaptation frameworks.

Acknowledgements

The contributions presented in this dissertation were carried out in the context of the following projects: H2020 SHELTER (GA n.821282), H2020 SAFERS (GA n.869353), CEF DYDAS (Grant 2018-IT-IA-0101), HEU ATLANTIS (GA n.101073909), HEU OVERWATCH (GA n.101082320), PNRR NODES (Grant ECS00000036).

I would like to acknowledge LINKS Foundation for the opportunity to pursue this doctoral research program. I extend my gratitude to my supervisors, Prof. Barbara Caputo and Dr. Fabrizio Dominici, for their guidance and feedback throughout the research process, and my colleagues at PoliTo and LINKS for their support and availability.

My heartfelt appreciation goes to my all friends, especially Alessandro, Luca C. and Luca B. for their invaluable help and advice during these years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Remote sensing has emerged as a fundamental tool for Earth Observation, providing valuable information for a wide range of applications, such as land cover mapping, environmental monitoring, disaster response, and urban planning [88]. The increasing availability of high-resolution satellite and aerial imagery has opened up new possibilities for detailed analysis and interpretation of the Earth's surface [179]. However, the sheer volume and increased complexity of remote sensing data pose significant challenges in extracting insights and information from images. Semantic segmentation, a fundamental task in computer vision, has gained significant attention in remote sensing due to its ability to provide pixel-wise classification of images, enabling a detailed understanding of the scene's composition and spatial distribution of land cover and land use types [80]. By assigning a semantic label to each pixel in an image, semantic segmentation allows for the automatic delineation of objects and regions of interest, such as buildings, roads, vegetation, and water bodies. The advent of deep learning techniques, particularly convolutional neural networks (CNNs) [135] and Vision Transformers (ViTs) [61], has revolutionized the field of semantic segmentation. These approaches have demonstrated remarkable performance in capturing high-level features and learning complex patterns from large-scale datasets, surpassing traditional machine learning approaches [38]. The success of deep learning in computer vision tasks has inspired its application to remote sensing data, leading to significant advancements in semantic segmentation accuracy and efficiency [13]. However, semantic segmentation in remote sensing presents unique challenges compared to traditional computer vision scenarios. Remote sensing images often cover large geographic areas and exhibit high spatial and spectral variability [245, 16]. The objects of interest in these images can have diverse appearances, scales, and complex spatial relationships, making it challenging to capture their contextual dependencies and accurately delineate their boundaries.

Moreover, the limited availability of annotated data in remote sensing poses difficulties in training deep learning models, as the acquisition of pixel-level labels is time-consuming, costly, and often requires expert knowledge [142, 253]. To address these challenges, researchers have explored various approaches to adapt and enhance semantic segmentation techniques for remote sensing applications. One prominent direction is the development of specialized CNN architectures that can effectively capture the multiscale and multi-modal nature of remote sensing data [164]. Another important aspect is the exploration of weakly supervised learning approaches, which aim to reduce the reliance on expensive pixel-level annotations [99]. Weakly supervised methods leverage alternative forms of supervision, such as image-level labels, bounding boxes, or scribbles, to guide the training of semantic segmentation models. These approaches offer a more cost-effective and scalable solution for training models in scenarios where pixel-level annotations are scarce or impractical to obtain [165]. Domain adaptation techniques have also gained attention in the remote sensing community, as they enable the transfer of knowledge learned from one geographic location or from one sensor to another [125]. By minimizing the domain gap between source and target domains, domain adaptation methods can improve the generalization capability of semantic segmentation models, reducing the need for extensive fine-tuning or re-training when applying models to new areas or data sources [96]. This thesis aims to address these challenges, starting from current state-of-the-art solutions on natural images, and optimizing them for remote sensing scenarios with ad-hoc improvements. The main objective is to develop novel techniques and methodologies that can effectively leverage the unique characteristics of remote sensing data, overcome the limitations of existing approaches, and provide accurate and efficient semantic segmentation results in various scenarios, with an eye towards deployable solutions.

1.2 Research Contributions

This thesis presents a set of techniques and methodologies to address the unique challenges of semantic segmentation in remote sensing scenarios. The research contributions are organized into three main categories, each focusing on a specific challenge encountered in aerial images: (i) the aerial viewpoint, (ii) imbalance, both in terms of scales and classes, and (iii) domain robustness. For each challenge, we propose tailored methods and solutions to improve the accuracy and the generalization abilities of semantic segmentation models.

Top-down Viewpoint. Remote sensing images captured from an aerial perspective introduce unique challenges for semantic segmentation due to the top-down viewpoint and the arbitrary orientation of objects in the scene. To address these challenges, we propose a novel framework that combines Augmentation Invariance

(AI) regularization with an Adaptive Sampling (AS) strategy. The AI component guides the model to learn semantic representations that are invariant to photometric and geometric distortions, while the AS technique addresses class imbalance by dynamically selecting training samples based on class distribution and model confidence. We further extend these techniques to the context of incremental learning, introducing a contrastive distillation approach that enforces invariance to orientation changes and enhances the model's ability to incorporate new classes without forgetting previously learned features.

Scale and Class Imbalance. Aerial images often exhibit significant variations in object scale and a severe imbalance between classes, posing difficulties for accurate semantic segmentation. To tackle these issues, we focus on two specific applications: flood detection and photovoltaic panel segmentation. For flood detection, we construct a multimodal dataset combining SAR imagery, DEM data, and hydrography maps, and propose a multi-encoder architecture with entropy-weighted sampling to effectively fuse the different data modalities and address class imbalance. For photovoltaic panel segmentation, we introduce a multiscale regularization approach that encourages consistency between local and global features, improving the segmentation performance, especially for challenging categories. Additionally, we develop a post-processing algorithm to refine the segmentation output and generate cleaner polygonal representations of the detected objects.

Domain Robustness. Semantic segmentation models often struggle to generalize well to new geographic locations, sensors, or temporal conditions, requiring expensive pixel-level annotations for adaptation. To enhance domain robustness, reduce the need for manual annotations, and enable large-scale labeling, we explore four key techniques: unsupervised domain adaptation, learning from sparse annotations, multitask learning, and leveraging foundation models for automated annotation. We propose a hierarchical instance mixing strategy (HIMix) and a twin-head architecture for unsupervised domain adaptation, effectively aligning features across domains and improving segmentation performance. To learn from sparse annotations, we introduce SPADA, a framework that combines sparse ground truth labels with pseudo-labels generated by a teacher model, enabling effective training with limited annotated data. We then investigate the potential of multitask learning to improve the robustness and performance of semantic segmentation models, proposing RoBAD, a multitask learning framework that incorporates land cover classification as an auxiliary task to guide the training of the segmentation model. Lastly, we explore the use of foundation models as robust annotators with FMARS, a pipeline that employs state-of-the-art Large Vision Models (LVMs) to generate labels for remote sensing images at scale. FMARS addresses the challenge of creating extensive annotated datasets and demonstrates how these machine-generated labels can be used to train smaller, more manageable models for specific downstream tasks.

1.3 Thesis Outline

The thesis is organized into six chapters, following the structure of the research contributions presented in the previous section. The chapters are as follows: The current chapter, Chapter 1, provides an introduction to the research topic, highlighting the importance of semantic segmentation in remote sensing applications and the unique challenges posed by aerial images. It sets the context for the research contributions and outlines the objectives of the thesis. Chapter 2 presents the background and foundations necessary for understanding the research contributions of the thesis. It provides an overview of semantic segmentation techniques, including traditional methods and deep learning approaches. The chapter also introduces the fundamental concepts and datasets relevant to remote sensing and discusses the specific challenges encountered in this domain. Chapter 3 focuses on addressing the challenges arising from the aerial viewpoint in remote sensing images. It presents novel techniques, such as Augmentation Invariance (AI) regularization and Adaptive Sampling (AS), to learn semantic representations that are invariant to photometric and geometric distortions. The chapter also introduces a contrastive distillation approach for incremental learning, enabling the incorporation of new classes without forgetting previously learned features. Chapter 4 tackles the issues of scale and class imbalance in aerial and satellite images, in two application scenarios. It describes the construction of a multimodal dataset for flood detection and proposes a multi-encoder architecture with entropy-weighted sampling to effectively fuse different data modalities. For scale imbalance, the discussion focuses on photovoltaic panel segmentation, introducing a multiscale regularization approach and a post-processing algorithm to improve segmentation performance and generate cleaner polygonal outputs. Chapter 5 addresses the challenges of domain robustness and weak supervision in semantic segmentation for remote sensing. It explores techniques such as Unsupervised Domain Adaptation, learning from sparse annotations, multitask learning, and automated labeling through large vision models. The chapter presents a hierarchical instance mixing strategy (HIMix) and a twin-head architecture for unsupervised domain adaptation, a framework (SPADA) for learning from sparse annotations, a multitask learning framework (RoBAD) that incorporates land cover classification as an auxiliary task to guide the segmentation model, and an automated pipeline (FMARS) to generate labels on a large scale through LVMs. Finally, Chapter 6 concludes the thesis by summarizing the main contributions and findings for the field of remote sensing and semantic segmentation, and outlines potential future research directions and open challenges.

Chapter 2

Background and Foundations

2.1 Semantic Segmentation

Image segmentation is a fundamental problem in computer vision that aims to partition an image into multiple segments or regions, each corresponding to a different object or part of the scene. Formally, let $\Omega \subset \mathbb{R}^2$ represent the image domain. The goal of segmentation is to find a partition $\mathcal{R} = R_1, \ldots, R_n$ of Ω such that [85]:

1. $\bigcup_{i=1}^{n} R_i = \Omega$
2. $R_i \cap R_j = \emptyset$ for $i \neq j$
3. Each R_i is a connected set
4. $\mathcal{P}(R_i) = \text{TRUE}$ for some logical predicate $\mathcal{P}$
5. $\mathcal{P}(R_i \cup R_j) = \text{FALSE}$ for any adjacent regions R_i and R_j

S The first two conditions ensure that the regions cover the entire image domain without overlap. The third condition requires each region to be spatially connected. The fourth condition states that pixels in a segmented region must share some common property $\mathcal{P}$, such as intensity, color, texture, or semantic category. The last condition ensures that any merging of adjacent regions violates the property $\mathcal{P}$. There are three main types of image segmentation: semantic segmentation, instance segmentation, and panoptic segmentation [121, 88]. Semantic segmentation assigns each pixel a class label, but does not differentiate between different instances of the same class. For example, all pixels belonging to the "person" class would be labeled as such, without distinguishing between different people in the image. Instance segmentation, on the other hand, detects and delineates each distinct object of interest in the image. So different people would be segmented as separate instances.

Panoptic segmentation unifies the two tasks and aims to assign both a class label and an instance identifier to every pixel in the image.

Semantic segmentation can be seen as the task of assigning a semantic label to every pixel in an image, effectively partitioning the image into semantically meaningful regions that respect the rules above. Formally, let $\mathcal{I} : \Omega \rightarrow \mathbb{R}^c$ denote an input image, where $\Omega \subset \mathbb{R}^2$ is the image domain made of 2D pixel coordinates $\mathbf{p}$, and c is an arbitrary number of channels (i.e., 3 for typical RGB images). The goal of semantic segmentation is to infer a label mapping $\mathcal{Y} : \Omega \rightarrow 1, \ldots, K$ that assigns each pixel $\mathbf{p} \in \Omega$ to one of K predefined semantic categories [80]. In the deep learning paradigm, semantic segmentation is typically formulated as a dense pixel-wise classification problem. Given a training dataset $\mathcal{D} = (\mathcal{I}_i, \mathcal{Y}_i)i = 1^N$ consisting of image-label map pairs, the objective is to learn a mapping $f_\theta : \mathcal{I} \rightarrow \mathcal{Y}$ parameterized by θ, such that $f_\theta(\mathcal{I})$ approximates the true label map $\mathcal{Y}$ for a given test image $\mathcal{I}$. This is typically achieved by minimizing an empirical risk over the training set:

$$\theta^* = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathcal{I}_i), \mathcal{Y}_i) \tag{2.1}$$

where $\mathcal{L}(\cdot, \cdot)$ is a pixel-wise loss function, for instance the standard cross-entropy loss for multi-class classification.

Semantic segmentation has numerous applications across a wide range of domains. In autonomous driving, accurately understanding the drivable areas, pedestrians, vehicles, and other key objects in the scene is crucial for safe navigation [52, 63]. For intelligent transportation systems, semantic segmentation enables analysis of road scenes, detection of lanes, sidewalks, and obstacles to assist driver decision-making. In the medical field, segmenting anatomical structures and regions of interest from medical images such as MRI or CT scans aids in diagnosis, treatment planning, and surgical interventions [197]. Other applications include precision agriculture, where segmenting aerial or satellite imagery into categories such as crops, soil, and weeds facilitates monitoring of crop health and targeted treatments [49]. In robotics, semantic understanding of the environment is essential for tasks like grasping, manipulation, and human-robot interaction [205]. Automatic segmentation of images and video frames also enables content-based retrieval, efficient compression, and many other applications.

However, despite its wide-ranging utility, semantic segmentation of real-world imagery poses several challenges. Unlike image classification, which predicts a single label for the entire image, classifying each pixel is a significantly more complex task. Segmentation models must balance local detail with global context, as both high spatial resolution and long-range dependencies are needed for precise segmentation [39]. The model must handle objects with varying scales, poses, and appearances, and segment the boundaries between objects precisely. Another challenge is the large cost and time needed to annotate datasets with pixel-wise labels for training and evaluation. Compared to drawing bounding boxes or assigning image-level tags, meticulously tracing object boundaries is tedious and error-prone. As a result, semantic segmentation datasets tend to be smaller in scale or less diverse compared to image classification datasets [69, 27]. Models trained on limited data may not generalize well to the long tail of object appearances found in unconstrained environments. Segmentation is also more sensitive to small errors, as incorrectly labeling even a few pixels along boundaries can substantially alter the perceived quality of the result. Approaches that excel at classification may underperform on segmentation without architectural changes to capture fine-grained spatial information. Furthermore, many applications demand extremely efficient inference in order to segment images in real-time on low-power devices. Models designed solely for accuracy are often prohibitively expensive for effective deployment [112].

2.1.1 Metrics

Given the fine-grained pixel classification, semantic segmentation adopts different metrics from other machine learning tasks to better describe the performance of the models. The most widely used metrics are based on the Intersection-over-Union (IoU) and F1 score, which provide a quantitative measure of the overlap between the predicted and ground truth segmentation masks. The IoU, also known as the Jaccard index, is a similarity measure between two sets. For binary segmentation, it is defined as the size of the intersection between the predicted and ground truth masks divided by the size of their union:

$$\mathrm{IoU} = \frac{|\mathcal{Y} \cap \hat{\mathcal{Y}}|}{|\mathcal{Y} \cup \hat{\mathcal{Y}}|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2.2}$$

where $\mathcal{Y}$ is the ground truth mask, $\hat{\mathcal{Y}}$ is the predicted mask, and TP, FP, and FN denote true positives, false positives, and false negatives respectively. IoU ranges from 0 to 1, with 1 indicating perfect overlap. For multi-class segmentation with K classes, the IoU can be computed in several ways. The class-wise IoU is calculated separately for each class, treating it as a binary problem of the class versus the rest. The macro-average IoU, often simply called mean IoU (mIoU), is the unweighted average of the class-wise IoUs:

$$\text{macro-IoU} = \frac{1}{K} \sum_{i=1}^{K} \mathrm{IoU}_i \tag{2.3}$$

where IoU_i is the IoU of class i. Macro-average IoU treats all classes equally, regardless of their frequency. On the other hand, the micro-average IoU is calculated globally over all classes, which effectively weights each class by its frequency:

$$\text{micro-IoU} = \frac{\sum_{i=1}^{K} \text{TP}i}{\sum i = 1^{K}(\text{TP}_i + \text{FP}_i + \text{FN}_i)} \tag{2.4}$$

The F1 score is the harmonic mean of precision and recall. For binary segmentation, it is defined as:

$$\text{F1} = 2 * \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{2.5}$$

Similar to IoU, the multi-class F1 score can be computed as a macro-average (unweighted mean of class-wise F1 scores) or micro-average (globally over all classes). The macro-average F1 score is given by:

$$\text{macro-F1} = \frac{1}{K}\sum_{i=1}^{K} \text{F1}_i \tag{2.6}$$

where F1_i is the F1 score of class i. The micro-average F1 score is calculated as:

$$\text{micro-F1} = \frac{2 \cdot \sum_{i=1}^{K} \text{TP}i}{2 \cdot \sum i = 1^{K}\text{TP}i + \sum i = 1^{K}(\text{FP}_i + \text{FN}_i)} \tag{2.7}$$

Both IoU and F1 score penalize both over- and under-segmentation, providing a balanced measure of segmentation quality. In practice, macro-average IoU (mIoU) is the most commonly reported metric for semantic segmentation, as it treats all classes equally and is sensitive to even small segmentation errors along object boundaries. When not specified otherwise, *average IoU* or *mean IoU* typically refers to macro-average IoU. However, micro-average metrics may be more appropriate when class frequencies are highly imbalanced. In some cases, it is crucial to consider both types of averaging when comparing segmentation models, as they provide complementary information about performance across different classes.

2.1.2 Literature Review

Traditional methods. Early approaches to semantic segmentation relied heavily on hand-crafted features and heuristic rules. Low-level cues such as color, texture, and edges were used to group pixels into coherent regions, often based on similarity criteria or graph-based formulations [85]. For example, thresholding techniques aimed to separate foreground objects from the background based on intensity differences, while region growing methods iteratively expanded segments based on local similarity measures. While these techniques could produce reasonable results for simple images with clear boundaries and homogeneous regions, they often struggled

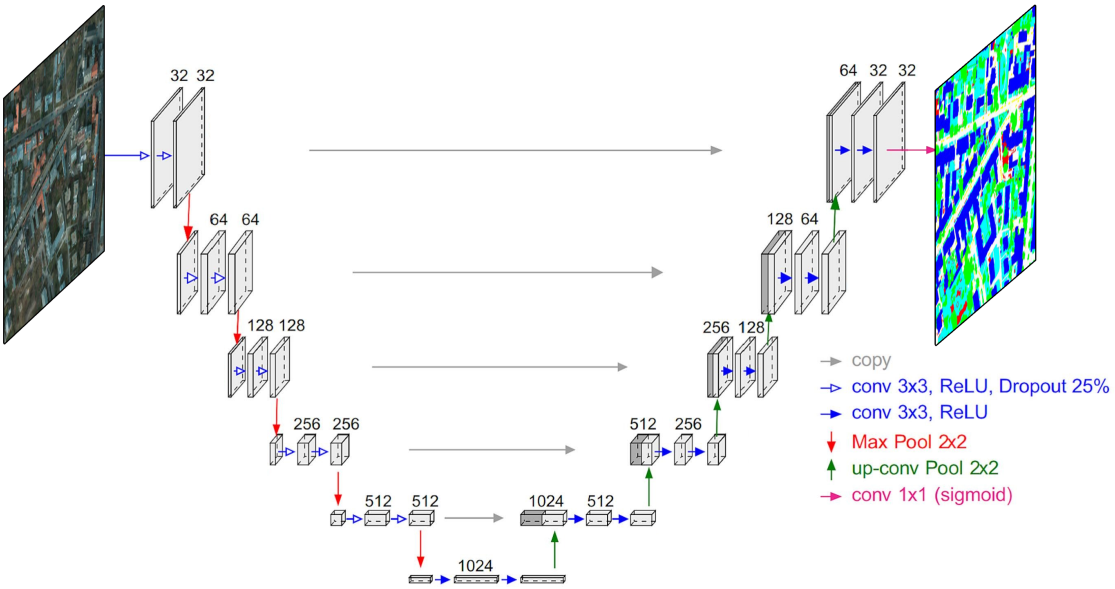


Figure 2.1: An example of U-Net architecture [197], the archetype of encoder-decoder model for semantic segmentation. The encoder (left) progressively downsamples the input image using convolutions and max-pooling, capturing higher-level features at each scale. The decoder (right) upsamples the feature maps using transposed convolutions and concatenates them with the corresponding encoder features through skip connections.

with more complex scenes containing occlusions, illumination changes, and intra-class variations [85]. The advent of machine learning brought more principled and data-driven approaches to semantic segmentation. Rather than relying solely on low-level cues, these methods aimed to learn a mapping from input features to output labels based on annotated training data. Popular techniques included Random Forests [210], boosting [101], and Support Vector Machines (SVMs) [243], which could learn discriminative classifiers for each semantic category. However, these methods still relied on hand-engineered features such as SIFT [137], HOG [55], and TextonBoost [115], which limited their generalization ability and robustness to variations in appearance and scale. Moreover, these approaches often treated each pixel independently, ignoring the rich contextual information and spatial dependencies present in natural images. To address this limitation, Conditional Random Fields (CRFs) [115, 70, 38] were widely used as a post-processing step to refine the segmentation maps produced by the classifiers. CRFs could model the pairwise compatibilities between neighboring pixels and encourage label consistency, leading to smoother and more coherent predictions. However, the performance of these methods was still heavily dependent on the quality of the initial segmentation and the choice of energy functions and inference algorithms.

Convolutional networks. The field of semantic segmentation has recently undergone a paradigm shift with the rise of deep learning, particularly Convolutional Neural Networks (CNNs). CNNs have the ability to learn hierarchical features directly from raw images, capturing both low-level appearance and high-level semantics in a unified framework. One of the first pioneering works introduced the concept of Fully Convolutional Networks (FCNs) [135], which adapted existing classification architectures such as AlexNet [61] or VGGNet [135] for dense prediction by replacing the fully-connected layers with convolutional ones. By leveraging the inherent spatial structure of convolutions and combining features from different depths, FCNs achieve significant improvements over previous state-of-the-art methods on challenging datasets like PASCAL VOC [69]. The downsampling operations in CNNs, such as pooling and stride convolutions, lead to a loss of spatial resolution in the feature maps, resulting in coarse and blurry segmentation boundaries. To alleviate this issue, several architectures based on the encoder-decoder paradigm have been proposed. These architectures typically consist of an encoder that progressively reduces the spatial resolution of the feature maps, and a decoder that gradually recovers the object details and spatial dimensions [15, 197]. The decoder often incorporates upsampling operations, such as transpose convolutions, and skip connections from the encoder to the decoder to combine low-level features with high-level semantic information. Another approach to address the loss of spatial resolution is to gather multiscale information, for instance through the use of dilated convolutions, which expand the receptive field without reducing the resolution of the feature maps [38, 39]. By using dilated convolutions, models can capture larger context while preserving the spatial information, leading to more precise segmentation boundaries. Some architectures, such as DeepLab [38], also integrate Conditional Random Fields (CRFs) as a post-processing step to further refine the segmentation results and enforce spatial consistency. As the depth and complexity of CNN architectures grow, overfitting becomes a major problem and the availability of large-scale annotated datasets becomes crucial. To improve the generalization and efficiency of semantic segmentation models, various techniques have been developed. Multiscale context aggregation modules, such as pyramid pooling [270] or atrous spatial pyramid pooling [39], are used to capture contextual information at different scales and resolutions.

Transformer architectures. Attention mechanisms have emerged as powerful tools for modeling long-range dependencies and capturing global context in semantic segmentation, since the emergence of the Transformer architecture [232]. Self-attention modules are used to adaptively integrate local features with their global dependencies, allowing the model to focus on relevant regions and capture contextual information at different scales [77, 262]. Some architectures introduce efficient attention mechanisms into convolutional models, such as axial-attention, which decomposes the global self-attention operation into separate attention operations along the spatial dimensions, making it computationally feasible for high-resolution feature maps [238]. Pure Transformer-based architectures have also been proposed, treating semantic segmentation as a sequence-to-sequence prediction task and leveraging the strong global context modeling capabilities of Transformers [271]. The advent of foundation models, large-scale pre-trained models that can be adapted to various downstream tasks, has further pushed the boundaries of semantic segmentation. These models, typically based on Transformers or hybrid CNN-Transformer architectures, are trained on massive amounts of unlabeled or weakly-labeled data using self-supervised learning objectives, and then fine-tuned on specific tasks with limited annotated data [61, 222, 131]. Some approaches introduce novel frameworks that unify different segmentation tasks, such as semantic and instance segmentation, using a Transformer decoder to predict a set of binary masks and their corresponding class labels [44]. Others extend this idea to panoptic segmentation, jointly predicting semantic and instance masks using a Transformer decoder with masked attention [45]. End-to-end Transformer architectures have also been designed specifically for semantic segmentation, achieving state-of-the-art performance. These architectures often incorporate novel mechanisms, such as overlapping patch embeddings and hierarchical feature extraction, to capture multiscale contextual information and produce high-resolution segmentation maps [255].

Benchmark datasets. The success of semantic segmentation methods on natural images heavily relies on the availability of high-quality annotated datasets. In the following paragraphs, we provide a non-exhaustive list of the most common benchmarks, considering natural images. PASCAL VOC [69] is one of the pioneering datasets for object detection and semantic segmentation, with 20 object categories and pixel-level annotations for a subset of images. Together with PASCAL VOC, Microsoft COCO [129] is probably the most well-known benchmark for natural images. This large-scale dataset contains 80 object categories with instance-level annotations, which can be converted to semantic segmentation labels. Another popular benchmark is ADE20K [272], a diverse dataset covering 150 semantic categories of objects, stuff, and parts, with annotations for both indoor and outdoor scenes. A large portion of semantic segmentation literature orbits around autonomous driving, also thanks to the availability of large-scale annotated datasets. For instance, Cityscapes [52] is a high-resolution dataset focused on urban scene understanding for autonomous driving, with 30 semantic classes and fine pixel-level annotations. It is also common to exploit synthetic data to improve generalization capabilities. For instance, SYNTHIA [198] is a large synthetic dataset of urban scenes with pixel-level semantic annotations, designed to facilitate research on domain adaptation and transfer learning [96]. These datasets have played a crucial role in advancing the state-of-the-art in semantic segmentation, providing standardized benchmarks for evaluating and comparing different approaches.

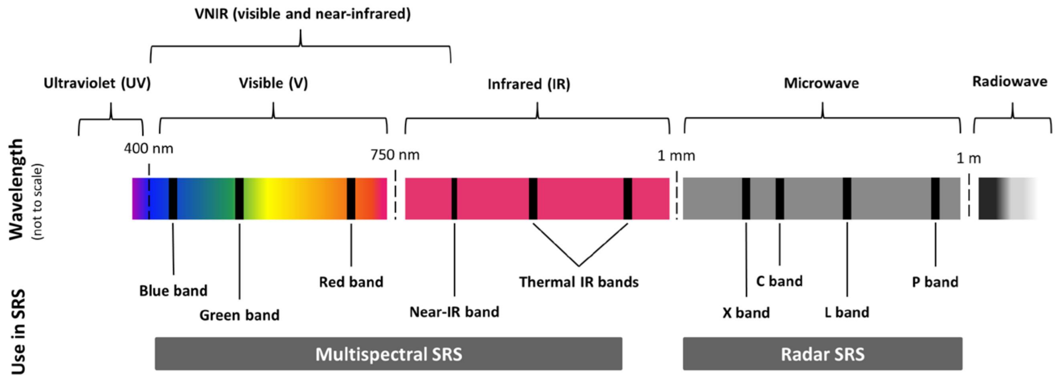


Figure 2.2: Simplified visualization of the electromagnetic spectrum, and distribution of the most common Satellite Remote Sensing (SRS) sensors [179].

2.2 Remote Sensing

Remote sensing is a field of study that focuses on obtaining information about the Earth's surface and atmosphere from a distance, typically through the use of satellite, airborne, or ground-based sensor technologies [195, 56]. This involves measuring the emanating energy using sensors on these above-ground platforms, and using their measurements to construct landscape images. The Earth's atmospheric composition and structure constrain the usable imaging wavelengths to the visible, infrared, thermal infrared, and microwave ranges [195], as displayed in Fig. 2.2. While the most common energy source remains reflected sunlight, other alternative sources can provide additional information, such as thermal emission, to inspect the surface heat and temperatures, or active microwave illumination, sensitive to the physical characteristics of the underlying terrain and capable of operating regardless of the weather. The sensors used to measure the radiation can be broadly classified into two categories: passive and active. Passive remote sensing relies on naturally occurring radiation, such as sunlight reflected from the Earth's surface or thermal energy emitted by the surface materials. Passive sensors detect and record this radiation without emitting any energy themselves. Examples of passive remote sensing include optical imagery captured by multispectral and hyperspectral sensors [56]. In contrast, active remote sensing systems generate their own electromagnetic radiation and measure the backscattered energy from the Earth's surface. Radar and LiDAR are examples of active remote sensing technologies, which use microwave and laser pulses, respectively, to illuminate the surface and record the returned signal [56].

The choice of instrument depends on the kind of interaction, that can be summarized into three forms, named *reflection*, *absorption*, or *emission*, depending on the properties of the surface materials and the wavelength of the radiation [56]. Reflection occurs when incident electromagnetic radiation bounces off the surface and is redirected back into the atmosphere. The amount of reflected energy depends on the surface's albedo, which is a measure of its reflectivity [195]. Surfaces with high albedo, such as fresh snow or white sand, reflect a large portion of the incident energy, while surfaces with low albedo, such as water bodies or dark soil, reflect less energy [56]. The reflected energy is often measured by passive remote sensing systems, such as multispectral and hyperspectral sensors, to characterize the surface properties and composition. This is the most common family of sensor, adopted for instance by Landsat [250] or Sentinel-2 constellations [65]. Absorption, on the other hand, occurs when incident electromagnetic radiation is taken up by the surface materials and converted into other forms of energy, such as heat [56, 184]. The degree of absorption varies depending on the wavelength of the radiation and the molecular structure of the surface materials. For example, vegetation absorbs a significant portion of the visible light for photosynthesis, while it reflects more energy in the near-infrared region [195, 209, 218]. By measuring the patterns of absorption and reflection across different wavelengths, remote sensing systems can provide information about the type, health, and distribution of vegetation on the Earth's surface [65]. Emission is the process by which surface materials release electromagnetic radiation due to their own thermal energy. All objects with a temperature above absolute zero emit radiation, with the wavelength and intensity of the emitted radiation depending on the object's temperature and emissivity [56]. Emissivity can be defined as a measure of a material's ability to emit thermal radiation compared to a perfect *black body* at the same temperature [18]. Remote sensing systems, particularly those operating in the thermal infrared region, can detect and measure the emitted radiation to determine the surface temperature and thermal properties of the Earth's surface [195]. This is for instance the case of the Moderate-Resolution Imaging Spectroradiometer (MODIS) [175] and Sentinel-3 [60] constellations.

2.2.1 Sensor types

In addition to the sensor type, remote sensing systems can also be classified based on the platforms that carry the sensors. The main types of remote sensing platforms include satellites, aircraft, and unmanned aerial vehicles (UAVs).

Satellite imagery. Satellite systems involve the use of sensors mounted on satellites in orbit to collect data about the Earth's surface and atmosphere. These satellites can be classified based on their orbit type, such as geostationary or polar-orbiting, and their sensor payload [195]. Geostationary satellites orbit the Earth at an altitude of approximately 36,000 km, maintaining a fixed position relative to the Earth's surface. These satellites are primarily used for weather monitoring and communications [56]. Polar-orbiting satellites, on the other hand, orbit the Earth at lower altitudes, between 600 and 800 km above the surface, and pass over the poles on each revolution. These satellites provide global coverage and are widely used for Earth observation, including land cover mapping such as Landsat [250] and Sentinel-2 [65], ocean monitoring such as Sentinel-3 [60], and atmospheric studies. Satellite remote sensing offers several key advantages, including the ability to provide consistent and global coverage of the Earth's surface, even in remote and inaccessible regions. These systems can collect data over large areas using a wide range of sensor types, resulting in diverse datasets suitable for various applications. Additionally, many satellite missions span several decades, ensuring long-term data continuity. However, satellite remote sensing also has some limitations, such as relatively coarse spatial resolution compared to aerial and UAV-based systems, or prohibitive costs for VHR data, fixed temporal resolution determined by the satellite's orbit and revisit cycle, and potential data quality issues caused by cloud cover and atmospheric interference.

Aerial imagery. Aerial remote sensing involves instead the use of sensors mounted on aircraft or helicopters, provides higher spatial resolution than satellite-based systems, enabling more detailed mapping of the surface and the detection of smaller features. They typically operate at lower altitudes compared to satellites, ranging from a few hundred meters to several kilometers above the ground [195]. These systems offer flexibility in data acquisition, as flight parameters and timing can be adjusted based on specific requirements. Aerial remote sensing also has the potential to capture oblique imagery, offering a different perspective compared to vertical imagery. Moreover, aerial surveys are generally less expensive and can be deployed faster than satellite missions. However, aerial remote sensing covers smaller areas compared to satellite-based systems and is more sensitive to weather conditions, which can impact flight schedules and data quality. Additionally, aerial remote sensing data may be subject to geometric distortions caused by aircraft motion and terrain variation.

UAV imagery. UAV-based remote sensing has gained popularity in recent years due to the increasing availability and affordability of small, lightweight drones equipped with high-resolution cameras and other sensors [56]. UAVs offer several advantages over traditional satellite and aerial platforms, including lower cost, higher flexibility, and the ability to collect data at very high spatial resolutions (often below 10 cm) [195]. UAV-based remote sensing is particularly useful for small-scale applications, such as precision agriculture, infrastructure monitoring, and environmental research. These systems can be deployed quickly and easily, allowing for frequent data collection and near-real-time monitoring of dynamic processes [29]. UAVs emerged as a popular choice in recent years, thanks to the increasing availability and affordability of small, lightweight drones equipped with high-resolution cameras and other sensors. These systems offer very high spatial resolution, enabling detailed mapping of small-scale features. UAVs provide high flexibility in data acquisition, as flight parameters and timing can be easily adjusted. They are also relatively low-cost and can be rapidly deployed compared to satellite and aerial platforms. Moreover, they have the potential to capture data beneath cloud cover and in areas that are inaccessible to larger platforms. However, UAV-based remote sensing has some limitations, such as limited coverage area due to battery life and range constraints, sensitivity to weather conditions (particularly wind and precipitation), and regulatory restrictions on UAV operation in certain areas, such as near airports or in urban environments.

2.2.2 Remote Sensing Resolutions

Remote sensing data is characterized by three key dimensions: spatial resolution, spectral resolution, and temporal resolution. These dimensions determine the level of detail captured in the data, the ability to distinguish between different features or materials, and the frequency at which observations are made. Understanding these characteristics is crucial for selecting the appropriate remote sensing data for a given application and interpreting the information derived from the data.

Spatial resolution. It refers to the size of the smallest object or feature that can be detected and distinguished in a remote sensing image. It is typically expressed in terms of the Ground Sample Distance (GSD), which represents the distance between the centers of adjacent pixels on the ground [195]. The smaller the GSD, the higher the spatial resolution of the image. The spatial resolution of remote sensing data varies depending on the sensor and platform used. For example, the Landsat-8 satellite has a spatial resolution of 30 meters for its multispectral bands and 15m for its panchromatic band [250], while Sentinel-2 provides a sampling of 10m per pixel. In contrast, VHR commercial satellites can provide imagery with a spatial resolution below 1m even for multispectral bands, such as Maxar data [89, 30], or Planet SkySat [203] (see Section 2.2.4). The choice of spatial resolution depends on the specific application and the level of detail required. High spatial resolution data is essential for tasks such as urban planning, infrastructure monitoring, and small-scale feature detection. However, higher spatial resolution also means larger data volumes and increased processing time. Lower spatial resolution data, on the other hand, is suitable for large-scale applications, such as global land cover mapping [78, 264], where larger and broader patterns are the main interest.

Spectral Resolution. Spectral resolution describes the ability of a remote sensing sensor to distinguish between different wavelengths of electromagnetic radiation, and it is determined by the number and width of the spectral bands captured by the sensor [56]. A spectral band is a range of wavelengths within the electromagnetic spectrum, such as visible light (400-700 nm), near-infrared (700-1100 nm), or thermal infrared (8-14 µm). Multispectral sensors typically have a few broad spectral bands, usually covering the visible and near-infrared regions of the spectrum. For example, the Landsat-8 Operational Land Imager (OLI) has nine spectral bands, including visible, near-infrared, and shortwave infrared bands [250]. These bands are selected to capture key features of the Earth's surface, such as vegetation health, water bodies, and mineral composition. Hyperspectral sensors, on the other hand, have hundreds or even thousands of narrow, contiguous spectral bands, providing a nearly continuous representation of the electromagnetic spectrum. This high spectral resolution allows for the detection and identification of specific materials or chemical compounds based on their unique spectral signatures [56]. Hyperspectral data is particularly useful for applications such as mineral exploration, vegetation species mapping, and water quality monitoring. Usually, spectral resolution is inversely proportional to spatial resolution: VHR sensors often provide visible (i.e., RGB) and optionally near-infrared (NIR) data, while lower resolution sensors such as Sentinel-2 compensate the lack of detail by providing more bands [65]. The choice of spectral resolution depends on the specific application and the level of spectral detail required. Multispectral data is sufficient for many applications, such as land cover classification and vegetation monitoring, while hyperspectral data is necessary for more specialized tasks that require the identification of specific materials or chemical compounds [195].

Temporal Resolution. Temporal resolution refers to the frequency at which a remote sensing system acquires data over a given area. This is determined by the revisit time of the sensor, intended as the time it takes for the satellite to return to the same location on the planet's surface [56]. The temporal resolution of remote sensing data can range from a few minutes to several weeks or even months, depending on the sensor and platform used. High temporal resolution is crucial for monitoring dynamic processes and changes, such as vegetation phenology, natural disasters, and urban development. For example, the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor aboard the Terra and Aqua satellites provides daily global coverage, enabling near-real-time monitoring of larger scale phenomena such as wildfires, floods, and algal blooms [195]. Sensors with lower temporal resolution, such as Landsat-8 (16-day revisit time) or Sentinel-2 (5-day revisit time), are still valuable for many applications, particularly those that focus on longer-term changes or trends [78, 142]. These sensors provide consistent, high-quality data that can be used to track changes in land cover, monitor agricultural productivity, and assess the impacts of climate change over time. The choice of temporal resolution depends on the specific application and the rate at which changes occur in the phenomena of interest. Once again, the temporal resolution is often directly tied to the spatial resolution: higher resolution sensors often have longer revisit times, also due to the larger amount of data to be processed and stored.

2.2.3 The Copernicus Programme

In the context of Earth Observation, remote sensing has emerged as a crucial tool for monitoring and understanding our planet. Among the openly available resources at the European level, the Copernicus Programme, coordinated by the European Space Agency (ESA), stands out as the most comprehensive and ambitious initiative. This program effectively revolutionized the access to Earth Observation data worldwide, providing a wealth of information to support environmental monitoring, emergency response, and scientific research. The Copernicus Programme, formerly known as Global Monitoring for Environment and Security (GMES), can be defined as the European Union's Earth observation program [12]. It serves as the central reference point for the EU to exploit remote sensing for environmental monitoring, emergency management, and civil security. In fact, Copernicus provides a comprehensive, unified system that collects and processes vast amounts of satellite data, supporting decision-makers, researchers, and other stakeholders with timely, accurate, and easily accessible information [105]. The primary objectives of the Copernicus Programme are threefold. First, it aims to monitor and understand the Earth's environment and climate change. Second, it provides support for emergency response and civil security. Finally, it fosters the development of innovative applications and services based on Earth observation data [12]. To achieve these objectives, Copernicus integrates data from a variety of sources, including dedicated satellites (Sentinels), contributing missions, and in situ sensors. The Programme consists of several interconnected components that work together to provide a comprehensive Earth observation system. The space component, managed by the ESA, represents its backbone. It includes the Sentinel family of satellites, specifically designed to meet the operational needs of the program [12]. The Sentinels provide a unique set of observations, covering various aspects of the Earth's environment. Sentinel-1 focuses on all-weather, day-and-night radar imaging (i.e., SAR) for land and ocean services, while Sentinel-2 provides high-resolution optical imaging for land monitoring. Sentinel-3 is a multi-instrument mission for marine and land monitoring, and Sentinel-4 and Sentinel-5 are dedicated missions for atmospheric monitoring. The Sentinel-5 Precursor aims to reduce data gaps between Envisat and Sentinel-5, and Sentinel-6 is a radar altimetry mission for measuring global sea-surface height [68]. ESA is responsible for the development, launch, and operation of the Sentinel satellites, as well as for coordinating access to data from contributing missions. The in situ component complements the space-based observations by providing ground-based, airborne, and seaborne measurements. These measurements are essential for calibrating and validating satellite data, as well as for providing additional information that cannot be obtained from space [105]. The services component of Copernicus transforms the raw data collected by the space and in situ components into value-added information products. These products are delivered through six thematic services: Land Monitoring, Marine Environment Monitoring, Atmosphere Monitoring, Climate Change, Emergency Management, and Security [105]. Each service provides specific information products tailored to the needs of users in their respective domains. The Copernicus Programme is an invaluable tool for computer vision research applied to remote sensing data, and it constitutes the primary source of data for several works presented in this manuscript, especially considering land applications (see Section 5.3) or disaster management (Section 4.2 and Section 5.4). The vast amounts of high-quality, multi-modal data provided by the Sentinel satellites and contributing missions offer the necessary inputs for advanced computer vision techniques applied to Earth observation problems, while more refined products such as delineation maps derived from the Emergency Management Service (EMS) [64], or land cover maps derived from the Land Monitoring service can be exploited as manually validated ground truth. Among the various missions, Sentinel-1 and Sentinel-2 have emerged as the most widely utilized. In the following sections, we will focus on these two missions, exploring their characteristics, data products, and their specific relevance to the field of computer vision in remote sensing, as they form the backbone of many data sources used in this thesis.

Sentinel-1

Sentinel-1, a constellation of two polar-orbiting satellites (Sentinel-1A and Sentinel-1B), is the first of the five Copernicus missions developed by ESA. As the European Radar Observatory for the Copernicus joint initiative, Sentinel-1 operates day and night in all weather conditions, carrying a C-band Synthetic Aperture Radar (SAR) instrument that enables the acquisition of imagery regardless of the weather [221]. The Sentinel-1 mission is designed to provide continuity of data from the European Remote Sensing (ERS) satellites and the Envisat mission, which operated C-band SAR instruments. With its improved revisit time, coverage, timeliness, and reliability of service, Sentinel-1 has become a crucial tool for operational applications and scientific research [221]. The two satellites, Sentinel-1A and Sentinel-1B, share the same orbital plane with a 180° orbital phasing difference, allowing for a revisit time of 6 days at the equator. The satellites fly in a near-polar, sun-synchronous orbit at an altitude of 693 km, with a 12-day repeat cycle and 175 orbits per cycle [221]. This configuration ensures consistent long-term data archives, which are essential for applications based on long time series.

Products. The SAR instrument aboard Sentinel-1 satellites can operate in four exclusive modes: Stripmap (SM), Interferometric Wide swath (IW), Extra-Wide swath (EW), and Wave (WV). From the data acquired in each mode, the Instrument Processing Facility (IPF) generates two types of Level-1 products: Single Look Complex (SLC) and Ground Range Detected (GRD). GRD products are further classified by their resolution into Full (FR), High (HR), and Medium (MR) [221]. Level-2 Ocean (OCN) products are also available, containing geophysical parameters such as ocean wind fields, swell spectra, and surface radial velocities [221]. SLC products contain focused SAR data that uses the full signal bandwidth and preserves the phase information. These products are suitable for applications requiring phase information, such as interferometry and coherence analysis. GRD products, on the other hand, are detected, multi-looked, and projected to ground range using an Earth ellipsoid model [221]. They are more suitable for applications that do not require phase information, such as backscatter analysis, classification, and change detection.

Applications. Sentinel-1 data is utilized in a wide range of applications. In land monitoring, it plays a crucial role in forestry, or agriculture [225, 181]. For forestry applications, Sentinel-1 can be used for clear-cut and partial-cut detection, forest type classification, biomass estimation, and disturbance detection. In agriculture, Sentinel-1 data helps in monitoring crop conditions, soil properties, and tillage activities, as well as in assessing land use and predicting harvests [225]. Terrain deformation mapping using Sentinel-1 interferometric SAR (InSAR) data allow for the detection of surface movements with millimeter-level accuracy, which is essential for monitoring land subsidence, structural damage, or underground construction [174]. For maritime monitoring, Sentinel-1 is used for ice monitoring, ship detection, oil spill monitoring, and the observation of marine winds and waves. Ship detection using Sentinel-1 data enables the identification of vessels not carrying Automatic Identification System (AIS) or other tracking systems, which is crucial for monitoring illegal activities such as illegal fishing and piracy [248]. Oil spill monitoring applications use Sentinel-1 data for gathering evidence of illegal discharges, analyzing the spread of oil spills, and prospecting for oil reserves by highlighting naturally occurring seepage [185]. In the context of emergency management, Sentinel-1 data is invaluable for flood monitoring, earthquake analysis, and landslide and volcano monitoring [161, 174]. SAR's capability to observe during cloud cover and Sentinel-1's frequent revisits make it ideal for flood monitoring, allowing for the assessment of flooded areas and the impact on human, economic, and environmental loss. SAR systems, however, including Sentinel-1, are subject to a characteristic phenomenon known as *speckle noise*. This noise appears as a grainy texture in SAR images, resulting from the coherent nature of radar imaging. Speckle occurs due to the constructive and destructive interference of radar waves reflected from multiple scatterers within each resolution cell. While speckle is often considered noise, it actually contains information about the imaged surface's structure. However, speckle can complicate image interpretation and analysis tasks, often requiring noise reduction techniques for an effective use [186].

InSAR provides the unique ability to produce medium and high-resolution maps of earthquake deformations, enabling the discovery of active fault lines and the study of potential risks. SAR interferometry can also locate areas prone to landslides and monitor surface deformation to provide early warning of potential disasters and monitoring of critical infrastructure. Pre-eruption uplift and post-eruption volcanic shrinkage can be monitored with similar interferometric techniques, complementing in-situ networks from volcano observatories [174]. As of August 4, 2022, the Copernicus Sentinel-1B satellite has reached the end of its mission due to an anomaly related to the instrument electronics power supply in December 2021. However, the mission continues with the fully operational Sentinel-1A satellite, and plans are in place to extend the constellation in the short term with a new Sentinel-1C satellite in the following months [66].

Sentinel-2

Band	**Central Wavelength** (nm)	**Bandwidth** (nm)	**Resolution** (m)
B1	443	20	60
B2	490	65	10
B3	560	35	10
B4	665	30	10
B5	705	15	20
B6	740	15	20
B7	783	20	20
B8	842	115	10
B8A	865	20	20
B9	945	20	60
B10	1375	30	60
B11	1610	90	20
B12	2190	180	20

Table 2.1: Spectral bands of the Sentinel-2 Multi-Spectral Instrument (MSI). The table lists the 13 spectral bands, their corresponding central wavelengths, bandwidths, and spatial resolutions.

Sentinel-2 is a wide-swath, high-resolution, multi-spectral imaging mission designed to provide continuity for the SPOT and Landsat missions. The full mission specification of the twin satellites, flying in the same orbit but phased at 180°, is designed to give a high revisit frequency of 5 days at the Equator [65]. As documented in Table 2.1, the mission carries a single Multi-Spectral Instrument (MSI) payload that samples 13 spectral bands, with four bands at 10m, six bands at 20m, and three bands at 60m spatial resolution, sensing at an orbital swath width of 290 km [65]. The MSI measures the Earth's reflected radiance in spectral bands from aerosol, to visible and infrared (both NIR and SWIR), with the design driven by the need for large swath high spatial and spectral resolution imagery [65].

Products. Sentinel-2 products are available at the public in two processing levels: Level-1C and Level-2A. Level-1C products provide Top-Of-Atmosphere (TOA) reflectance images derived from the associated Level-1B products. They are composed of 100×100 km^2 tiles in UTM/WGS84 projection, but each tile is provided with an additional 5 km overlap on each side, resulting in 110×110 km^2 images. Level-2A products, on the other hand, provide atmospherically-corrected Surface Reflectance (SR) images, derived from the Level-1C products. They are also provided as 110×110 km^2 tiles, and include additional outputs such as an Aerosol Optical Thickness (AOT) map, a Water Vapor (WV) map, and a Scene Classification (SCL) map [65].

Applications. Sentinel-2 data is used in a wide variety of applications, including land monitoring, emergency management, and security. In land monitoring, Sentinel-2 data is used for various purposes, such as mapping land cover and land use changes [78, 142], monitoring vegetation health and growth, and estimating geophysical parameters like Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) [90] and Leaf Area Index (LAI) [240]. Sentinel-2's frequent revisits and relatively high spatial resolution make it ideal for emergency management applications, such as rapid mapping in response to natural disasters [51, 72, 10]. In the security domain, Sentinel-2 data supports applications like maritime surveillance, border monitoring, and infrastructure monitoring [132]. Sentinel-2 data are also important for various Copernicus services, including the Land Monitoring Service, which uses Sentinel-2 data for a variety of applications such as spatial planning, forest management, water management, and agriculture [53, 225]. Another example of Copernicus downstream application is the Emergency Management Service, that uses Sentinel-2 data to provide rapid mapping in response to natural disasters and other emergencies [64].

2.2.4 Other Platforms and Constellations

In addition to the Copernicus programme and its associated Sentinel missions, there is a wide range of Earth observation satellite constellations and platforms that provide valuable data for remote sensing applications. These systems offer diverse capabilities across various spectral bands, spatial resolutions, and temporal frequencies, responding to the different needs of researchers and industries. In this section, we present a non-exhaustive list of some of the most popular satellite constellations and platforms, grouped by sensor type, namely optical, thermal, and radar.

Optical Sensors. Optical constellations are by far the most common, comprising the majority of the available satellite sources. Among these, the Landsat program, jointly managed by NASA and the United States Geological Survey (USGS), is the longest-running Earth observation satellite series, with the first satellite launched in 1972. Landsat satellites carry optical sensors, providing openly available moderate-resolution multispectral imagery of the whole planet. The most recent satellites in the series, Landsat 8 and Landsat 9, offer spatial resolutions of 15 meters for the panchromatic band and 30 meters for the multispectral bands [250]. Another open data source is MODIS, a key instrument aboard the NASA Terra and Aqua satellites. MODIS provides a comprehensive suite of global observations of the Earth's surface and atmosphere, with a spatial resolution ranging from 250 meters to 1 kilometer. The sensor acquires data in 36 spectral bands, covering visible, near-infrared, and shortwave infrared wavelengths. MODIS data is widely used for applications such as land cover mapping, vegetation monitoring, and ocean color studies [175].

Among higher resolution satellite constellations it is worth mentioning the *Satellite Pour l'Observation de la Terre* (SPOT), a series of high-resolution optical imaging satellites operated by the French space agency, Centre National d'Études Spatiales (CNES). The SPOT satellites provide multispectral imagery at spatial resolutions ranging from 1.5 to 20 meters, depending on the sensor and spectral band. SPOT data is commonly used for land cover mapping, urban planning, and commonly adopted for disaster management [47, 64].

Another very popular VHR data source is represented by Maxar Technologies [148], with its constellations WorldView and GeoEye. The former consists of high-resolution commercial Earth observation satellites, providing panchromatic imagery at spatial resolutions ranging from 31 to 46 centimeters, as well as multispectral imagery at resolutions from 1.24 to 1.85 meters. This source is also often used for a wide range of applications, from urban monitoring to disaster response [89]. GeoEye is another high-resolution commercial Earth observation satellite constellation, also operated by Maxar. GeoEye-1, the flagship satellite of the constellation, provides panchromatic imagery at a spatial resolution of 41 centimeters and multispectral imagery at 1.65 meters.

Among the European options, it is worth mentioning Airbus Pleiades. These are a series of high-resolution Earth observation satellites operated by France's CNES [84]. These satellites, Pleiades 1A and 1B, were launched in December 2011 and December 2012, respectively. They provide optical imagery with a resolution of 50 cm, enabling detailed monitoring and analysis for similar applications such as urban planning, agriculture, and disaster management.

Thermal Sensors. In addition to their optical capabilities, several satellite programs also include thermal infrared sensors. The Landsat program, for instance, incorporates thermal infrared sensors alongside its optical sensors. Specifically, Landsat 8 and Landsat 9 are equipped with the Thermal Infrared Sensor (TIRS) [250], which collects data in two thermal infrared bands at a spatial resolution of 100 meters. This thermal data is invaluable for various applications, such as surface temperature mapping, water resource management, and wildfire monitoring [200]. Similarly, the MODIS instrument, as mentioned earlier, acquires data in the thermal infrared region of the electromagnetic spectrum. MODIS thermal bands, with a spatial resolution of 1 kilometer, are employed for a range of applications, including land surface temperature monitoring, cloud detection, and atmospheric temperature and humidity profile estimation [175]. Furthermore, the Visible Infrared Imaging Radiometer Suite (VIIRS) instrument [31], operated by NOAA and NASA on the Suomi NPP and NOAA-20 satellites, collects thermal infrared data with spatial resolutions of 375 meters and 750 meters. VIIRS data is crucial for applications such as sea surface temperature measurement, wildfire detection, and atmospheric studies.

Another thermal source is the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) instrument aboard NASA's Terra satellite captures high-resolution thermal infrared data at 90-meter spatial resolution across five thermal bands. This data is used for applications such as volcanic activity monitoring, land surface temperature estimation, and mineral exploration [204].

Lastly, the Sentinel-3 satellite, part of the European Space Agency's Copernicus program, carries the Sea and Land Surface Temperature Radiometer (SLSTR). SLSTR measures thermal infrared radiation at a spatial resolution of 1 kilometer, supporting sea surface temperature monitoring, land surface temperature mapping, and wildfire detection [60].

Radar Sensors. Radar satellite constellations offer all-weather, day-and-night imaging capabilities. Among the available constellations, RADARSAT is operated by the Canadian Space Agency (CSA), and consists of SAR satellites providing radar imagery at various spatial resolutions and polarizations. The most recent addition, RADARSAT Constellation Mission (RCM), offers a spatial resolution of 1 meter in ultra-fine mode and a revisit time of up to four times per day. RADARSAT data is crucial for sea ice monitoring, ship detection, and disaster management [192, 162]. Similarly, the German Aerospace Center (DLR) operates the TerraSAR-X [182] and TanDEM-X satellites [273], which provide high-resolution radar imagery in the X-band (i.e., with a wavelength of around 3cm) with spatial resolutions ranging from 0.24 to 40 meters. The unique flying formation of these satellites enables the generation of high-precision global digital elevation models (DEM). Data from TerraSAR-X and TanDEM-X is utilized for topographic mapping, surface deformation monitoring, and urban planning [182, 273]. Another prominent radar constellation is COSMO-SkyMed, operated by the Italian Space Agency (ASI) and the Italian Ministry of Defense. This constellation consists of four SAR satellites that provide high-resolution radar imagery with spatial resolutions down to 1 meter. COSMO-SkyMed data is used for various applications, including emergency management, environmental monitoring, and defense and security [25].

2.3 Semantic Segmentation in Remote Sensing

Semantic segmentation plays a crucial role in the analysis and understanding of remote sensing imagery. This task of assigning a label to each pixel has become increasingly important in the field of remote sensing due to its wide range of applications, from land cover mapping [142, 53, 78], to urban planning [199, 59], environmental monitoring [79], or disaster management [22, 51, 72, 10]. The following sections introduce the concepts behind computer vision applied to Remote Sensing, highlightning the challenges at hand, and providing a brief albeit comprehensive view of the research landscape in this field, focusing on segmentation.

2.3.1 Characteristics and Challenges

Remote sensing imagery presents unique challenges compared to natural images when it comes to semantic segmentation. One of the primary differences lies in the viewpoint. Remote sensing images are typically captured from an aerial or satellite perspective, providing a top-down view of the Earth's surface. This viewpoint introduces variations in object appearances, scales, and orientations, which can significantly impact the performance of segmentation algorithms [91]. Another challenge in remote sensing semantic segmentation is the large-scale and high-resolution nature of the imagery, in terms of raw pixel dimensions. Remote sensing datasets often cover vast geographical areas and contain images with extremely high spatial coverage [89, 199]. This leads to scale imbalances, where different objects of interest may appear at vastly different scales within the same image or across different images. This is also tightly linked with class imbalance: while overrepresented or underrepresented categories are often common in the available datasets, disparities in average dimensions between classes (e.g., “car” vs “building”) directly translates into an inherent pixel imbalance, which pose difficulties in training segmentation models and achieving accurate results [261]. The complexity and heterogeneity of landscapes captured in remote sensing imagery further complicate the semantic segmentation task. Remote sensing scenes often encompass a wide variety of land cover types, such as urban areas, forests, agricultural fields, and water bodies, sometimes even across wildly different geographical areas. These landscapes exhibit diverse textures, patterns, and spectral characteristics, making it challenging to develop robust segmentation models that can generalize well across different domains. Domain adaptation techniques are often employed to address this issue, enabling the transfer of knowledge learned from one domain to another [6]. Compared to natural images, limited data availability remains a significant issue in remote sensing to this day. Collecting and annotating large-scale remote sensing datasets is a time-consuming and labor-intensive process, and the availability of labeled data is often limited in terms of scope, dimensions, and the number of annotated samples. This scarcity of labeled data hinders the training of deep learning models, which typically require vast amounts of annotated examples to achieve high performance. Several works have explored various approaches to mitigate this challenge, such as transfer learning, or weakly, and self-supervised learning techniques [3, 78, 245]. In the following section, we provide a comprehensive overview of the semantic segmentation landscape, applied to the remote sensing field, and the different approaches and solutions proposed for the aforementioned challenges.

2.3.2 Literature Review

Methods. In parallel with the methodological evolution in natural images, machine learning techniques have been widely explored for semantic segmentation in remote sensing imagery. Traditional supervised learning algorithms such as Support Vector Machines (SVMs) and Random Forests (RFs) have demonstrated effectiveness in capturing spatial and contextual information, handling high-dimensional feature spaces, and processing large datasets for remote sensing image segmentation [243, 191, 142]. Unsupervised learning methods, which do not require labeled data and aim to discover inherent patterns and structures in the input data, have also been commonly used in this context. Algorithms like K-means [141] and Self-Organizing Maps [73] have been proposed for different tasks. However, unsupervised methods often require post-processing steps to refine the segmentation results and align them with the desired semantic classes. Similarly, semi-supervised learning approaches have shown promise in enhancing the generalization capability of classifiers and learning meaningful feature representations from unlabeled data [120, 213]. Comparative studies have highlighted the superiority of deep learning approaches, particularly convolutional neural networks (CNNs), over traditional machine learning methods in terms of segmentation accuracy and generalization ability [166]. FCNs have been widely adopted for semantic segmentation tasks, replacing the fully connected layers in traditional CNNs with convolutional layers to generate pixel-wise predictions [135]. The U-Net architecture [197], originally proposed for biomedical image segmentation, has been successfully applied to remote sensing imagery in several contexts [72, 51, 59, 57]. Several variants and extensions of the standard encoder-decoder architecture have been proposed to improve segmentation performance, incorporating techniques such as residual connections [59], atrous convolutions [39], pyramid pooling [270], and attention mechanisms [232] to enhance the network's ability to capture multiscale contextual information, focus on relevant features, and suppress irrelevant ones [160, 2, 98, 238]. Transformers have recently gained traction in remote sensing, with architectures such as ViT [61] and Swin Transformer [131] showing promising results for image classification tasks and semantic segmentation in aerial scenarios [218] and satellite scenes [27, 234]. Deep learning-based approaches have also been explored for domain adaptation and transfer learning in remote sensing, aiming to bridge the gap between different data domains and improve segmentation performance across different datasets [242]. Among techniques aimed at mitigating data scarcity, it is worth mentioning incremental learning approaches, [217] and multitask learning frameworks [246].

Applications. Land Use and Land Cover (LULC) mapping is one of the most common applications of semantic segmentation in remote sensing, aiming to classify the planet surface into categories such as urban, forest, water, and agricultural land [142, 57]. Accurate and up-to-date LULC maps are crucial for various purposes, including urban planning, environmental monitoring, and natural resource management. Recent datasets, such as the ESA World Cover [264], provide valuable information for monitoring land cover changes and enabling downstream applications [10]. Semantic segmentation also plays a vital role in infrastructure mapping, such as building footprint extraction and road network delineation [46]. Identifying and monitoring man-made structures from satellite imagery is essential for urban development, transportation planning, and disaster response. Deep learning architectures and large-scale datasets have been proposed to tackle these challenges, enabling the development and evaluation of semantic segmentation models for infrastructure monitoring [46, 21, 89]. In the agriculture domain, semantic segmentation is widely used for precision farming and crop monitoring [74], allowing farmers to optimize resource allocation and maximize crop yield by accurately delineating field boundaries, identifying crop types [225], and detecting weeds or diseases [49]. Pixel-wise segmentation approaches and large-scale aerial image databases have been introduced to facilitate the development and evaluation of semantic segmentation models in this field[74, 49, 209, 218]. Disaster management, such as flood mapping and damage assessment, also heavily relies on semantic segmentation. Rapid and accurate mapping of flood extent and affected areas is essential for emergency response and resource allocation [189, 106]. Georeferenced datasets and large-scale damage assessment datasets have been created to support the development of deep learning algorithms for flood detection and post-disaster damage assessment [22, 89, 161]. Despite the advancements in semantic segmentation for remote sensing, several challenges remain, such as the large variation in viewpoint and scale of objects, class imbalance problem, and domain adaptation issues. Researchers have proposed various techniques to address these challenges, including rotation equivariant detectors, entropy-based sampling approaches, and incremental learning methods [91, 122, 217].

Datasets. To train and evaluate semantic segmentation models for remote sensing, many larger and smaller datasets have been proposed in literature, in different contexts and application scenarios. Some of these sources were mentioned above and refer to specific application scenarios, such as ESA WorldCover [264] (which is itself a product of machine learning), Agriculture-Vision [49], Sen1Floods11 [22], or xBD datasets [89].

The ISPRS Vaihingen and Potsdam datasets [199] have been widely used as benchmarks for semantic segmentation in remote sensing, similar to the role that the PASCAL VOC dataset has played in the computer vision community. These datasets provide VHR aerial imagery and corresponding pixel-wise annotations for several object classes, such as buildings, roads, and vegetation. The Vaihingen dataset consists of 33 patches of varying sizes, while the Potsdam dataset comprises 38 patches, each covering an area of 6000×6000 pixels. Building upon these sources, several larger-scale and more diverse datasets have been introduced in recent years to address the growing needs for resources.

DeepGlobe [57] was introduced as part of a challenge for parsing the Earth through satellite images. It consists of three sub-datasets: road extraction, building detection, and land cover classification. The dataset covers a diverse range of geographic locations and provides high-resolution satellite imagery along with pixel-wise annotations for each task.

DOTA (Dataset for Object deTection in Aerial images) [252] is a large-scale dataset designed for object detection in aerial images. It covers 15 object categories, such as planes, ships, vehicles, and bridges, with annotations provided in the form of oriented bounding boxes. The dataset captures objects at various orientations and scales in high resolution, making it a good source for aerial pretraining.

LoveDA [242] is a land cover dataset designed for domain adaptive semantic segmentation. It consists of high-resolution aerial images from multiple cities and sensors, covering different geographic locations and imaging conditions. The dataset provides pixel-wise annotations for several land cover classes, such as building, road, water, and vegetation.

More recently, thanks to the increasing popularity and effectiveness of large foundation models [112], many large-scale datasets have been proposed in the literature, either for benchmarking or pretraining purposes. For instance, OpenEarthMap [253] provides additional labels for a general-purpose LULC mapping, reusing the images from other task-specific datasets and effectively obtaining a large-scale worldwide benchmark. Given the amount of data comprising remote sensing resources, these sources may need high storage requirements. For instance, SSL4EO-S12 [245] provides instead a worldwide collection of Sentinel-1 and Sentinel-2 co-registered tiles for pretraining, with more than 1TB of data. SAtlas Pretrain [16] is perhaps the prime example of large-scale pretraining source, with more than 30 TB of information, with diverse tasks and images, with worldwide coverage.

Chapter 3

The Aerial Viewpoint

3.1 Introduction

Aerial imagery has become a valuable tool for a wide array of applications, however the unique characteristics of aerial imagery pose significant challenges for existing semantic segmentation models, which are often adapted from other domains such as autonomous driving or medical imaging [135, 197]. One of the most prominent challenges is the top-down perspective inherent to aerial images, where scenes are captured from a bird's eye view. This perspective allows for arbitrary rotations of the sensor around the vertical axis, resulting in the same scene being captured from different angles. Consequently, the appearance of objects and their spatial relationships can vary significantly across different images, making it difficult for traditional segmentation models to generalize effectively [19]. To address these challenges, in this chapter we propose specific approaches that explicitly account for these unique properties. One option is to exploit this invariance with respect to rotations of the remote sensing images through regularization techniques. While leveraging geometric and photometric augmentations to train models can improve the robustness of models with respect to changes in image orientation, we can explicitly model that into the training procedure.

In Section 3.2, we introduce a framework that implements this concept as an Augmentation Invariance (AI) regularization, combined with an Adaptive Sampling (AS) strategy. The AI component guides the model to learn semantic representations that are invariant to photometric and geometric distortions commonly found in aerial imagery. The AS technique addresses the class imbalance problem by selecting training samples based on the pixel-wise distribution of classes and the current confidence of the model. This approach proves to be effective on the Agriculture-Vision dataset, consistently outperforming the baseline methods. These techniques can be further extended, and they can also be applied in other settings, such as incremental learning. In Section 3.3, we propose a contrastive regularization technique that compares the segmentation features produced by an input image and its augmented version (i.e., flipped, or rotated). By minimizing the difference between these features, the model learns to be invariant to orientation changes. This approach not only improves the robustness of the model to different aerial viewpoints but also improves the model distillation phase, allowing the model to incorporate new classes without forgetting previously learned features.

The contributions presented in this chapter led to the publication of two works:

- Arnaudo A. Cermelli F., Tavera A., Rossi C., Caputo B., *A Contrastive Distillation Approach for Incremental Semantic Segmentation in Aerial Images*, In Proceedings of the 21st International Conference of Image Analysis and Processing (ICIAP 2022), Lecce, Italy, May 23–27, 2022 (pp. 742-754).

- Tavera A., Arnaudo A., Masone C., Caputo B., *Augmentation Invariance and Adaptive Sampling in Semantic Segmentation of Agricultural Aerial Images*, In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, (pp. 1656-1665).

3.2 Augmentation Invariance in Aerial Images

The use of semantic segmentation for environmental monitoring through aerial imagery has continuously expanded in recent years, from urban mapping [199] to cropland analysis [49]. In these contexts, as with other computer vision challenges, the use of deep learning has contributed to revolutionary results. However, the majority of deep learning models applied to semantic segmentation were originally designed for different use cases mostly involving natural images, such as navigating autonomous vehicles [52, 198] or completely different domains, such as inspecting medical acquisitions [197]. Given the similarity between scenarios, these models are often repurposed for remote sensing analysis without significant modifications to accommodate the unique characteristics inherent to aerial imagery, which can negatively impact the effectiveness of these frameworks when applied outside their original context.

Focusing on the distinct challenges posed by aerial imagery, one of the main aspects that distinguishes it from other domains is the top-down perspective. In the remote sensing field, images that are typically captured from an aerial viewpoint utilize cameras mounted on aircraft and usually positioned orthogonal to the Earth's surface to fully capture the scene with limited obstructions. This perspective is unique in that it lacks traditional depth cues and reference points that are often present in terrestrial or street-level photography. Furthermore, the ability to capture images from any rotational angle around the vertical axis adds another layer of complexity [91]. This flexibility in viewpoint results in images where the spatial organization of semantic elements is far less predictable than in datasets designed for natural images, where structural elements like roads and sky are consistently positioned in specific regions of the image. This fundamental difference in perspective requires tailored approaches in the analysis and processing of aerial images to effectively understand and interpret the complex and dynamic landscapes captured from above.

Another important aspect, especially in environmental scenarios, is represented by class unbalance. This is a well-known issue in the computer vision domain: this is already pronounced in the context of aerial imagery, where the objects of interest vary greatly in size (i.e., from small vehicles to expansive natural regions), and it is brought to the extreme in challenging tasks such as detecting defective cropland patterns [49], given the inherently infrequent nature of the entities to locate. Consequently, this issue translates into widely different raw pixel counts, that need to be taken into consideration for a balanced training.

To address the unique challenges posed by aerial imagery, we propose a set of regularization to be applied at training time, specifically aimed at semantic segmentation on remote sensing imagery to enhance performance. Our proposed approach incorporates two techniques: *Augmentation Invariance* (AI) and *Adaptive Sampling* (AS). The former method exploits various augmentations to train with a contrastive approach, where the same inputs are subtly modified in different ways so that the model can learn features that are robust to changes in appearance and perspective. Among these, rotations around the vertical axis remain the predominant focus, so that the obtained features remain invariant to the point of view. Overall, this technique aims to enable the model to focus on semantic content rather than just visual appearance. Adaptive Sampling tackles instead the issue of class imbalance using two main parameters: first it statically conditions the selection of training images based on the distribution of pixels among the available classes, then it dynamically adjusts the class sampling weights by exploiting the model's current level of confidence in its predictions. This method ensures that underrepresented classes receive more focus during training, promoting a more balanced learning process. Both strategies are integrated into a cohesive end-to-end training framework, enhancing the model's ability to interpret complex aerial images effectively.

In summary, this section provides the following contributions: (i) a study of the aerial viewpoint as an exploitable characteristic to obtain better results through regularization, proposing the AI strategy to manage the unique perspective challenges of aerial data, helping the model to better differentiate between semantic information and visual details, and (ii) an adaptive sampling (AS) method aimed at mitigating the severe class imbalance commonly encountered in aerial image datasets. This approach dynamically adjusts the sampling of training data based on real-time assessments of class distribution and model confidence, ensuring a fairer representation of all classes during the learning process. An ablation study is further carried out on the overall framework to analyze the impact of each solution introduced, aiming to validate their effectiveness and utility comprehensively.

Last, (iii) we conduct a thorough series of experiments using the Agriculture Vision dataset [49], a large aerial dataset that encompasses multiple semantic classes over agricultural fields with complex patterns, focusing on recent transformer-based architectures such as SegFormer as main backbone [255], given their robustness in other contexts. Moreover, the experiments were designed to also evaluate the effectiveness of training models using only RGB images versus incorporating Near-Infrared (NIR) data, which can provide additional valuable information, especially considering vegetation. The relevant code to reproduce the results presented here is available at `https://github.com/edornd/agrivision-2022`.

3.2.1 Related Works

In the domain of aerial and remote sensing, semantic segmentation has been applied to various target environments, including urban areas [59, 166, 13], land cover [21, 242, 57], and agricultural scenarios [152, 257, 49]. Each application comes with its own set of challenges and requirements. For instance, urban monitoring tasks often necessitate high-resolution imagery and may involve temporal change detection [46, 140]. Land cover segmentation, on the other hand, must handle significant variations in semantic category sizes and visual differences across domains, which can be addressed through multi-level or multiscale feature aggregation [260] and domain adaptation techniques [242, 19]. In agricultural settings, traditional segmentation approaches rely on vegetation indices like NDVI [249]. However, the current trend is shifting towards more robust computer vision techniques, such as automated fusion of multi-spectral data [209] and precise crop segmentation [74]. Agricultural aerial images often include additional spectral bands, such as near-infrared (NIR), alongside the visible spectrum. To effectively utilize this multi-modal data, researchers have proposed solutions like duplicating input weights [49, 177] or employing early or late fusion strategies [260, 181]. Another critical aspect of aerial imagery is the arbitrary and uncertain camera orientation, which can significantly impact the performance of semantic segmentation models. While this issue has been addressed in incremental learning [8] and classification tasks [187], this can also be applied in semantic segmentation, as shown in the following sections.

3.2.2 Method

Our framework, illustrated in Fig. 3.1, focuses on the SegFormer architecture [255] by incorporating two additional modifications. In the first place, we implement a specialized regularization function designed to align pixel embeddings from the Transformer network for both the original and augmented images, named Augmentation Invariance. This alignment aims to ensure that the semantic representations learned by the model are robust against common photometric distortions and changes of perspective encountered in aerial imagery. Second, we devise an

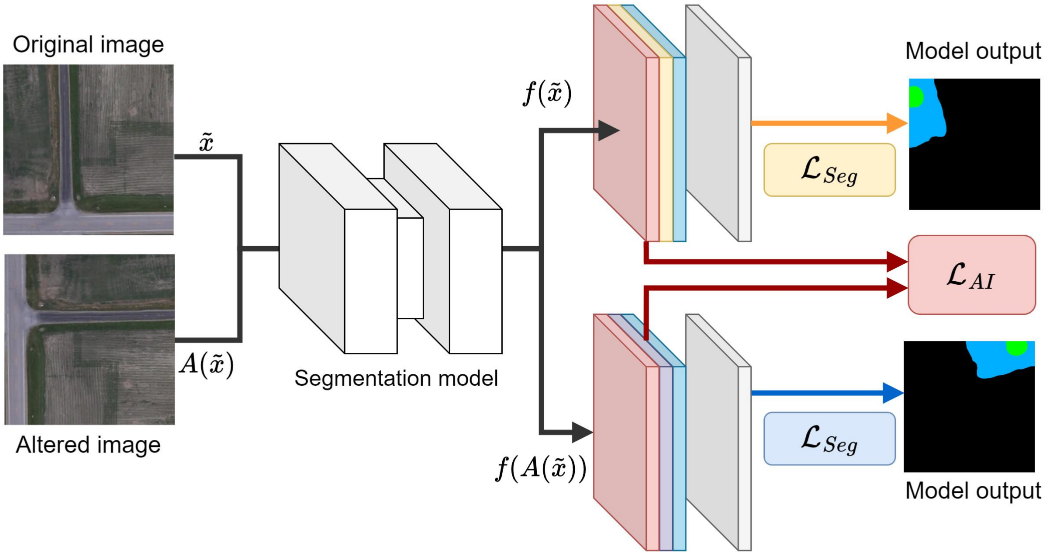


Figure 3.1: Overview of the Framework Architecture. The process begins with the *Adaptive Sampling* process dynamically selecting a sample, followed by the creation of its augmented version. These images are then processed by the segmentation model, which calculates the standard segmentation loss, $\mathcal{L}_{seg}$. Simultaneously, the $\mathcal{L}_{AI}$ objective drives the model to identify and extract identical features from both the original and the augmented images in a contrastive way through *Augmentation Invariance*, ensuring consistency in feature recognition across transformations.

Adaptive Sampling strategy that selects training samples based on the prior knowledge of class distribution and the current confidence levels of the network. In the subsequent sections, we begin by defining the problem setting and then elaborate on these two mechanisms.

Problem Statement

Here, we address the problem of detecting agricultural cropland defects from aerial images using semantic segmentation. Our training data consists of triplets $< x, y, z >$ of images organized as $S = \{(x \in \mathbb{X}, y \in \mathbb{Y}, z \in \mathbb{Z})\}$, where $\mathbb{X}$ denotes the set of RGB images, $\mathbb{Z}$ refers to the Near-Infrared (NIR) component, and $\mathbb{Y}$ comprises the semantic labels that link each pixel to a specific class c within a fixed and well-defined array of semantic classes $\mathbb{C}$. For clarity, we refer to $\mathbb{I}$ to refer to the collective set of pixels present in each image and mask. In order to feed the model with a four-dimensional image containing every channel, i.e., red, green, blue and NIR, we further construct $\tilde{x} \in \tilde{\mathbb{X}}$ as a four-channel image by concatenating x and z along the channel dimension.

Considering the RGB-NIR images, the aim of this work is to develop a mapping function $f_\theta : \tilde{\mathbb{X}} \to \mathbb{R}^{|\mathbb{I}| \times |\mathbb{C}|}$ that, using a fixed set of parameters θ adjustable via training, assigns a probability to each pixel in RGB-NIR inputs to indicate its likelihood of belonging to any given semantic category in $\mathbb{C}$. In every experiment presented in this section, we first focus on minimizing the standard cross-entropy loss L_{seg} as base objective function :

$$L_{\text{seg}}(\tilde{x}, y) = -\frac{1}{|\mathbb{I}|} \sum_{i \in \mathbb{I}} \sum_{c \in \mathbb{C}} y_i^c \log(p_i^c(\tilde{x})), \tag{3.1}$$

where $p_i^c(\tilde{x}) = f_\theta(\tilde{x})[i, c]$ is the output of our model for a pixel i in class c, and y_i^c indicates the correct classification for that pixel and class. We note that the same process can be applied to the images in the RGB space, namely x, to obtain $p_i^c(x) = f_\theta(x)[i, c]$.

Augmentation Invariance

Typical frameworks designed for semantic segmentation may experience significant performance declines when applied to aerial imagery, due to the variable perspectives inherent in aerial images, the potential for substantial distortions caused by different camera angles, and substantial photometric variations across various scenes.

To address these challenges, we introduce a contrastive regularization approach named Augmentation Invariance (AI). This mechanism employs augmentations to train the model to learn representations that remain invariant to these perspective and appearance variations. The process operates as follows: in the first place, given an input image x, pixel-wise features $f_i(x)$ are extracted from the penultimate layer of the model architecture, intentionally bypassing the final layer used for pixel-wise segmentation. This yields raw pixel-wise information that is not yet classified into a specific category, however already represents a localized information that could potentially be mapped to those classes by the subsequent linear layer. In short, the feature maps should already contain a robust representation of the scenery, with high vector similarities between semantically comparable pixels. At the same time, a duplicate of x is subjected to a series of augmentations, including a random pipeline of geometric transformations A_g (such as *horizontal flipping*, *vertical flipping*, and *random rotation*) and photometric augmentations A_p (such as *color jitter*). For brevity, we denote the composite augmentation process as $A_p \circ A_g = A$. The augmented image $A(x)$ is processed through the model to extract features $f_i(A(x))$. This process can be equally applied to both RGB and RGB-NIR images, respectively named x and $\tilde{x}$.

Once both the original and augmented feature maps have been obtained, we enforce that the information extracted from the original image x align with those extracted from the augmented image $A(x)$, ensuring invariance to the applied transformations. However, due to the application of geometric transformations, the obtained tensors have to be transformed back in a common reference system by either applying the same geometric modifications A_g to the feature maps of the original image $f_i(x)$, or by applying an inverse transformation to the feature maps of the augmented image $f_i(A(x))$. In our work, the alignment is achieved using a pixel-wise mean squared error loss, defined as:

$$L_{AI}(x, A(x)) = \frac{1}{|I|} \sum_{i \in I} (f_i(x) - A_g^{-1}(f_i(A(x))))^2 \tag{3.2}$$

where A_g^{-1} denotes the inversion of the geometric augmentations applied to x, ensuring that features from corresponding pixels are compared.

Additionally, we preserve the ground truth annotations of the augmented images, applying the segmentation loss to both the original and augmented images. The total training loss is formulated as:

$$L_{tot} = L_{seg}(x, y) + L_{seg}(A(x), A_g(y)) + \lambda L_{AI}(x, A(x)) \tag{3.3}$$

where $A_g(y)$ represents the geometric transformation applied to the ground truth annotation y, and λ is a balancing factor.

This approach is distinct from conventional data augmentation, as it pairs the original and transformed images to provide robust guidance during training in a contrastive fashion, with the aim of enhancing the model's ability to learn representations that are robust to variations in perspective and appearance.

Adaptive Sampling

As further discussed in Chapter 4, a significant challenge in semantic segmentation of aerial imagery is the severe imbalance in the pixel-wise distribution of semantic classes, with some classes appearing extremely rarely while others are overly represented. To mitigate this issue and ensure that the model receives a balanced exposure to all semantic categories during training, we introduce an Adaptive Sampling (AS) technique that operates in tandem with the AI approach described above. The core idea behind the AS method is to dynamically select training samples based on two primary factors: the global, pixel-wise class distribution, and the current network confidence for each category. By considering these factors, the sampling mechanism prioritizes images that contain underrepresented classes and those for which the network exhibits the least confidence. This approach helps to regularize the training process and prevents the model from being biased towards the more frequent classes. More formally, the AS method assigns a probability AS_c to each class c at every iteration, defined as:

$$AS_c = \sigma((1 - dist * conf)^{\gamma}), \tag{3.4}$$

where *dist* represents the global class distribution, stored as an array with length $|C|$, $conf$ denotes the current class-wise network confidence, again stored as a list with equal length, σ is a normalization function using a simple *min-max* approach, and γ is a relaxation parameter. For γ values greater than 1, the sampling probability distribution becomes more skewed towards the underrepresented classes, while using $\gamma < 1$, the distribution becomes more uniform, giving less emphasis to the rare classes.

After selecting a semantic class c based on this dynamically updated probability, an image is randomly sampled from a subset of data X_c that contains pixels belonging to the chosen class. To compute the class distribution *dist*, we exploit the supervised learning setting and statically calculate the pixel count for each semantic class $c \in C$ as a preprocessing step. The resulting array, which encodes the class distribution, is normalized to the range $[0, 1]$ and denoted as *dist*. The min-max normalization function is applied to ensure consistency to these class priors and allow for a probability-like factor at sampling time. The network confidence $conf$ is instead computed *online*, during the training process, and maintained in an array of size $|C|$. At each iteration step t, the pixel-wise Softmax probabilities are computed on the current batch of prediction logits. The mean confidence value for each class c is then derived from the available ground truth labels by averaging the pixels belonging to that category. To smooth out the confidence estimates, therefore avoiding sudden changes in the sampling mechanism, the actual network confidence is computed as the exponential moving average of the prior confidence at step $t-1$:

$$conf_t = \alpha conf_{t-1} + (1-\alpha) conf_t, \tag{3.5}$$

where α represents a smoothing factor. The higher the value of α remains, the more the network gives weight to the historical estimates. Likewise, lower values emphasize the current estimate more. By actively selecting training samples based on the global class distribution and the current network confidence, the Adaptive Sampling approach exposes the model to a more balanced distribution of semantic classes. This, in turn, leads to improved performance on underrepresented categories and helps to prevent the model from being dominated by the more frequent classes. The integration of both AI and AS into the training loop allows learning robust and invariant representations for accurate semantic segmentation in unbalanced aerial settings.

3.2.3 Experiments

Implementation details

We evaluate our proposed approach on the Agriculture-Vision dataset [49], a novel large-scale aerial agricultural image dataset designed for advancing research in agricultural pattern analysis and semantic segmentation. The dataset consists of 94,986 high-quality aerial images collected from 3,432 farmlands across the United States between 2017 and 2019. Each image is accompanied by dense pixel-level annotations of nine types of important field patterns, including *double plant*, *dry-down*, *end row*, *nutrient deficiency*, *planter skip*, *storm damage*, *water*, *waterway*, and *weed cluster*. These labels were obtained by manual annotation, carried out by domain experts. The aerial images in Agriculture-Vision have spatial resolutions ranging from 10 cm to 15 cm per pixel, significantly higher than typical open satellite images [65]. Furthermore, each image in this dataset contains four spectral channels: Red, Green, Blue (RGB), and Near-infrared (NIR), providing rich spectral information for vegetation pattern analysis. Agricultural experts annotated the aforementioned nine types of field patterns on these images using a commercial labeling software. The annotated images were then cropped into patches using a sliding window approach to generate the final dataset, which consists of 56,944 training images, 18,334 validation images, and 19,708 test images. The dataset exhibits significant class imbalance, with some field patterns, such as *dry-down* and *weed cluster*, occupying much larger areas than others. The size and shape of the same field pattern can also vary substantially across different images. Moreover, the scarcity of some field patterns, such as *storm damage*, poses additional difficulties for model training and evaluation. Thanks to its high-resolution, multi-spectral imagery, large-scale annotations, and the peculiar use case, the Agriculture-Vision dataset provides a robust benchmark for semantic segmentation. Due to the unavailability of the test set at the time of these study, we assess the performance of our method on the provided validation set. We conduct two sets of experiments: the first using only RGB images for training and testing, and the second exploiting the combination of RGB and NIR data. The images are provided in a tiled format, with each tile having a resolution of 512 × 512 pixels.

In terms of metrics, we adopt the standard mean Intersection over Union (mIoU) [69] to evaluate the performance of our method in all the experiments, in line with previous works [49, 209], as in Eq. (2.3).

Given its effectiveness in the context of natural images, we build our framework on top of the SegFormer architecture [255], utilizing a MiT-B5 encoder pretrained on the ImageNet-1k dataset as backbone. We compare our method against a diverse set of state-of-the-art semantic segmentation techniques, including FCN [135], DeepLab V3 [40], DeepLab V3+ [39], UperNet [254], FPN [128], PSPNet [270], HRNetV2 [241], HRNetV2+OCR [262], and the vanilla SegFormer [255]. These baselines are trained using ResNet-50 or HRNetV2-W18 backbones pretrained on ImageNet, depending on the specific architecture. We develop our framework and reproduce all the baselines using the *mmsegmentation* [158] library, in turn based on the *PyTorch*. The experiments are conducted on two NVIDIA Tesla V100 GPUs, each with 16 GB of RAM. During training, we apply various data augmentation techniques, including random resizing from 1× to 2× the original size, random horizontal and vertical flipping, and random crops subsequently resized back to 512×512. For the evaluation phase, we perform inference on the raw data without any additional preprocessing. All the baselines and our model are trained for 80,000 iterations using AdamW as optimizer, with a learning rate of 6×10^{-5}, a weight decay of 0.01, and beta parameters set to the default value of 0.9 and 0.999 respectively. We employ a polynomial learning rate decay with a factor of 1.0 and an initial linear warm-up for 1,500 iterations. We do not use class-balanced loss or Online Hard Example Mining (OHEM) approaches, as in the original SegFormer [255], since our setup focuses on evaluating the goodness of the proposed methods, without introducing further balancing options. However, these additional techniques could be potentially included to further improve the performance in imbalanced scenarios. When training on the $\tilde{x} \in \tilde{\mathcal{X}}$ examples that provide an extra NIR channel, we expand the network input to four bands by duplicating the input weights of the first channel, belonging to red, the most similar in terms of wavelength [49]. For the AI variants, we apply additional transformations to the input images, including horizontal and vertical flipping, random rotation from 0° to 360° with a step of 90°, and photometric and perspective distortion with a strength of 0.1. The probability for each transformation is set to $p = 0.5$, meaning each augmentation has a 50% chance of being activated for each input. Based on the hyperparameter study detailed in the next section, the value of λ in Eq. (3.3) is set to 0.75 for every experiment. We also evaluate the performance of our model using different combinations of γ and α values on both the RGB and NIR-RGB experimental settings. For the relaxation parameter γ reported in Eq. (3.4), we considered the values in the set $\{1, 2, 4, 6\}$. This parameter adjusts the distribution of sampling probabilities, with higher values of γ increasing the emphasis on underrepresented classes and lower values resulting in a more balanced distribution. Similarly, for the smoothing factor α in Eq. (3.5), we explored a fixed set of values in $\{0.75, 0.85, 0.90, 0.968, 0.99\}$. Higher values of α assign more importance to past estimates, resulting in a more stable and smooth confidence curve, while lower values prioritize the current estimate and allow for faster adaptability. Based on this search, we empirically found that the combination $\gamma = 4$ and $\alpha = 0.968$ consistently produced the highest mIoU scores across both settings.

Results

We consider two experimental settings: using only RGB images and using a combination of RGB and Near-Infrared (NIR) images. Our method aims to address the unique challenges posed by aerial imagery, such as the need for handling multi-spectral data, extreme class imbalance, and large-scale variations in aerial patterns. Results for the RGB setting are displayed in Table 3.1. Our approach achieves a mIoU of 46.41, surpassing all baseline models by a significant margin. The improvement is particularly notable for underrepresented classes such as *double*

Method	Backgr.	Db. Plant	Dry-down	End row	Nutr. Def.	Pl. Skip	Water	Waterways	Weed Cl.	mIoU
FCN	69.99	16.91	45.55	0.18	13.66	6.62	42.27	0.52	8.50	22.91
DeepLab V3	66.27	17.01	40.64	9.46	16.40	10.04	17.06	12.29	9.97	22.13
DeepLab V3+	68.55	16.31	46.36	6.46	16.05	4.56	16.61	19.10	13.89	23.10
UperNet	65.84	15.79	38.03	10.12	17.31	11.09	4.47	15.45	16.94	21.67
SFPN	69.65	10.61	49.49	2.70	11.46	4.80	35.68	9.89	11.16	22.83
PSPNet	68.11	16.93	45.77	4.89	18.99	8.54	11.31	17.64	17.20	23.26
HRNetV2	71.21	16.81	55.10	5.22	18.63	13.26	13.03	21.23	14.07	25.39
HRNetV2+OCR	72.42	19.46	56.79	12.31	17.30	21.31	28.36	24.62	18.05	30.07
SegFormer	74.93	33.19	**59.65**	18.28	**31.64**	39.20	77.97	**41.45**	28.31	44.96
Ours	**75.47**	**36.97**	58.49	**22.69**	31.29	**41.39**	**80.23**	40.07	**30.42**	**46.41**

Table 3.1: Results of the experiments carried out on the RGB set in terms of IoU over the Agriculture-Vision dataset.

Method	Backgr.	Db. Plant	Dry-down	End row	Nutr. Def.	Pl. Skip	Water	Waterways	Weed Cl.	mIoU
FCN	68.35	9.40	47.57	0.54	15.16	9.97	53.74	0.47	10.17	23.93
DeepLab V3	69.03	19.97	43.94	5.85	23.98	17.86	46.74	29.03	11.36	29.75
DeepLab V3+	68.29	17.18	48.07	7.48	24.17	19.57	19.43	24.58	13.22	26.89
UperNet	67.43	15.63	36.40	10.73	20.37	14.57	34.21	25.28	14.54	26.57
SFPN	68.69	5.99	48.71	0.18	22.74	17.21	44.50	18.30	12.79	26.57
PSPNet	66.92	17.73	29.87	10.24	28.01	18.66	13.90	29.83	11.99	25.24
HRNetV2	71.28	16.99	54.30	4.52	27.90	15.74	21.66	25.47	17.88	28.42
HRNetV2+OCR	72.60	17.98	56.69	11.97	27.91	23.79	48.99	27.73	22.06	34.42
SegFormer	76.17	33.63	58.96	18.92	40.57	38.93	80.56	42.85	27.88	46.50
Ours	**76.19**	**37.32**	**61.75**	**24.57**	**42.75**	**42.01**	**81.32**	**43.71**	**31.75**	**49.04**

Table 3.2: Results of the experiments carried out on the RGB-NIR set in terms of IoU over the Agriculture-Vision dataset.

plant and *end row*, which experience an increase of 3.78 and 4.41 in IoU, respectively, compared to the best-performing baseline, SegFormer. To further enhance the segmentation performance, we incorporate NIR information alongside the RGB channels. The results for this configuration are shown in Table 3.2. In this setting, our method achieves an even higher mIoU of 49.04, outperforming the baselines by a substantial margin. The inclusion of NIR data proves to be advantageous for all classes, with the most significant improvements observed in categories such as *nutrient deficiency*, *dry-down*, and *waterways*, which exhibit respective IoU gains of 11.41, 3.26, and 3.64 compared to the RGB-only setting. These findings highlight the importance of multi-spectral information in computer vision applied to vegetation and agricultural analysis, in consistency with previous studies in the literature [257].

Method	Backgr.	Db. Plant	Dry-down	End row	Nutr. Def.	Pl. Skip	Water	Waterways	Weed Cl.	mIoU
SegFormer	76.17	33.63	58.96	18.92	40.57	38.93	80.56	42.85	27.88	46.50
SegFormer + AI	76.62	35.26	61.24	20.74	43.45	43.49	80.41	45.10	33.12	48.82
SegFormer + AS	75.89	35.86	59.23	22.5	41.25	40.72	77.98	40.85	30.99	47.25
SegFormer + AI + AS	76.19	37.32	61.75	24.57	42.75	42.01	81.32	43.71	31.75	**49.04**

Table 3.3: Ablation study conducted on both AI and AS components, applied to the RGB-NIR setup, highlighting the effectiveness of the combination of both techniques.

Method	Backgr.	Db. Plant	Dry-down	End row	Nutr. Def.	Pl. Skip	Water	Waterways	Weed Cl.	mIoU
0.1	76.60	33.92	60.24	18.84	41.92	41.28	82.23	42.45	31.70	47.69
0.25	76.54	35.26	60.70	20.55	42.22	43.84	80.60	43.16	33.25	48.46
0.5	76.48	35.79	59.71	20.34	42.65	40.03	81.12	44.52	32.00	48.07
0.75	76.62	35.26	61.24	20.74	43.45	43.49	80.41	45.10	33.12	**48.82**
1	76.57	34.42	60.25	20.32	41.95	40.03	82.14	43.51	32.22	47.93

Table 3.4: Hyperparameter study on the influence of λ applied to the RGB-NIR setting.

Ablation study. To further validate the effectiveness of our approach, we conduct an ablation study to assess the individual contributions of the AI and AS components. As shown in Table 3.3, the Augmentation Invariance technique provides a significant boost to the overall framework, confirming our hypothesis about the importance of addressing the specific challenges in agricultural aerial imagery. The addition of AS further enhances the performance, particularly for underrepresented classes, by dynamically selecting training samples based on the global class distribution and current network confidence. The combination of AI and AS yields the best results, demonstrating the combined effect of these components in tackling the unique characteristics of the aerial domain. We further perform a comprehensive analysis of the influence of the λ hyperparameter, which controls the intensity of the AI loss. As reported in Table 3.4, the best results are obtained with $\lambda = 0.75$, therefore applying a 25% reduction factor with respect to the standard segmentation loss, highlighting the importance of balancing between CE and feature consistency. Notably, even with suboptimal hyperparameter settings, our AI component outperforms all the baselines, highlighting its effectiveness. The qualitative results presented in Fig. 3.2 provide visual evidence of the superiority of our approach compared to the baselines. The segmentation maps generated by our AI and AS method exhibit more precise and accurate delineation of field patterns, especially for challenging cases with small, scattered, or irregularly shaped anomalies.

Performance costs. Inevitably, the implementation of AI and AS introduces some computational overhead, albeit with minimal impact. AI effectively doubles the batch size, as it processes both the original and augmented versions of each image. This primarily results in increased GPU memory usage, as the augmentations can be kept on the GPU device with minimal additional computational cost. AS, on the other hand, operates on a 1D vector of length C (where C is the number of classes), updating class confidence after each batch. This update process is computationally efficient, leveraging matrix operations on the GPU, with only the final update requiring CPU intervention. Moreover, these computational overhead only needs to be applied at training time, and can be dropped once the model is completely trained.

In summary, our experimental results and analysis on the Agriculture-Vision dataset demonstrate the effectiveness of the proposed AI and AS techniques in addressing the unique challenges of semantic segmentation in aerial imagery. By leveraging multi-spectral data, handling extreme class imbalance, and learning invariant representations, our approach achieves state-of-the-art performance, surpassing a wide range of baseline models.

3.3 Rotation Invariance applied to Incremental Learning

Semantic segmentation plays a crucial role in aerial image processing in many fields, including environmental and agricultural monitoring [13], as shown in previous sections. Despite the rapid progress, adapting these models to evolving environments and integrating new knowledge over time remains a challenge. The majority of state-of-the-art solutions are designed to operate on a static set of categories, relying on end-to-end training with the assumption that all relevant classes are known *a priori*. However, this assumption does not hold in real-world scenarios where new classes may emerge, or the domain may shift over time. When presented with new training data containing previously unseen categories, deep neural networks are prone to catastrophic forgetting [149], a phenomenon in which the model's performance on previously learned classes deteriorates significantly. This limitation hinders the practical deployment of semantic segmentation models in more complex environments, where the ability to incrementally learn and adapt is key. Incremental learning is a recurring problem in machine learning research, where several techniques have been proposed over the years to mitigate the catastrophic forgetting. These techniques can be broadly categorized into replay-based methods [217], parameter isolation approaches [143], and regularization-based methods [193]. While these approaches have displayed impressive results in tasks such as image classification and object detection, their application to semantic segmentation has been limited. The unique challenges posed by this downstream task, such as the need for dense pixel-wise predictions and the presence of imbalanced class distributions, often require ad-hoc solutions. In the context of aerial imagery, the problem of incremental learning is particularly relevant. Aerial datasets are often scarce and expensive to annotate, making it impractical to collect exhaustive labeled data. Moreover, the images themselves may be acquired incrementally over time (e.g., satellite acquisitions), effectively requiring to update and expand the model's knowledge as new information becomes available.

One of the key challenges in incremental learning for semantic segmentation is the need to balance the stability of previously acquired knowledge with the dynamism required to incorporate new information. Existing approaches, such as those based on knowledge distillation [35], provide an excellent starting point by

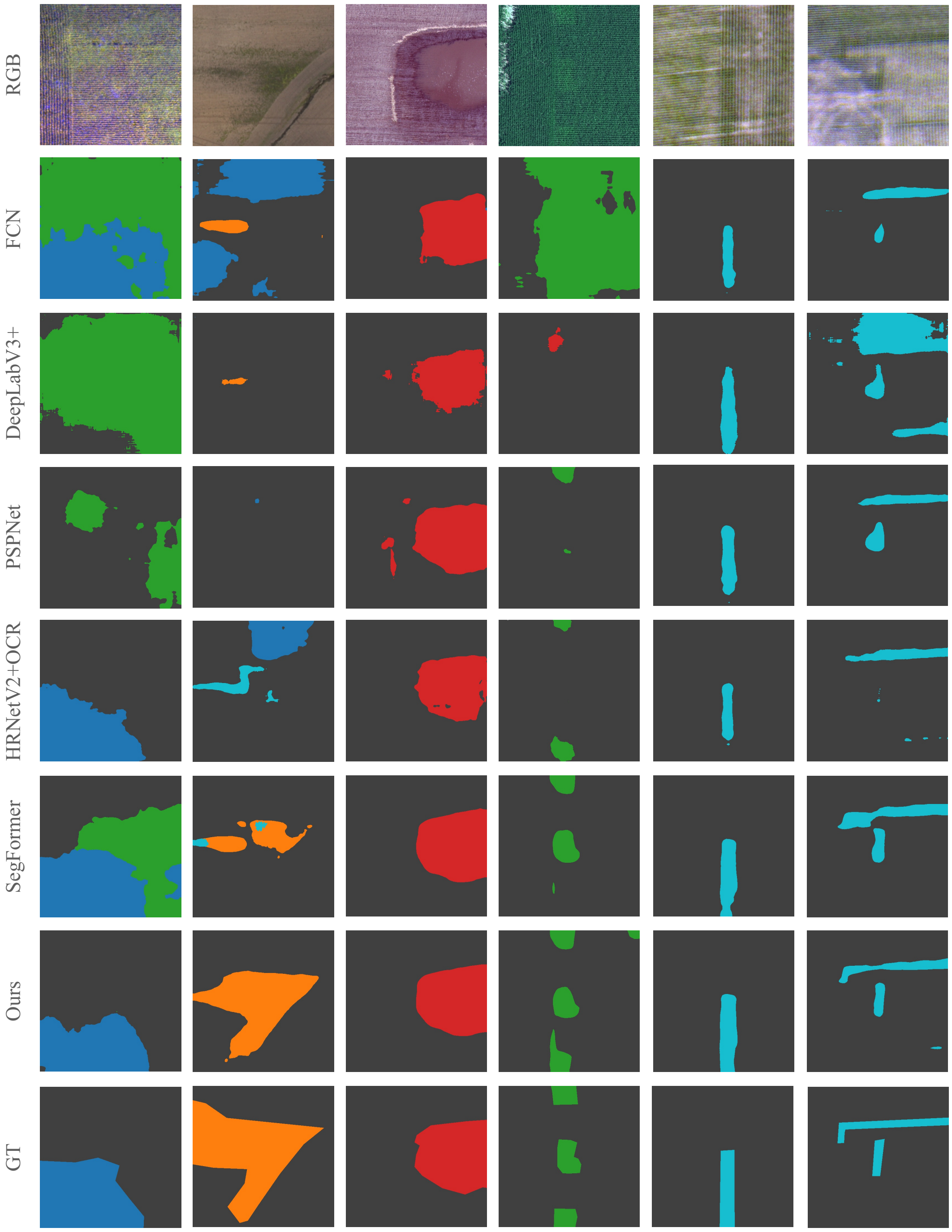


Figure 3.2: Qualitative results obtained on the Agriculture-Vision dataset.

leveraging the outputs of a previous model to guide the learning of a new model. However, these approaches often struggle to capture the unique characteristics of aerial imagery, such as the arbitrary orientation of objects and the presence of large-scale patterns. For these reasons, we propose here a novel Incremental Class Learning (ICL) framework for semantic segmentation, specifically designed to tackle the challenges of remote sensing images. Similar to the solution presented in Section 3.2, we introduce a contrastive regularization scheme that explicitly models the orientation invariance of aerial imagery. By encouraging the model to learn representations that are robust to rotations and reflections, we aim to improve the stability and generalization of the incremental learning process. Pivotal to our approach is the idea of contrastive representation learning, which has recently emerged as a powerful paradigm for unsupervised learning [41, 93]. Contrastive learning aims to learn representations that are invariant to certain transformations while being discriminatory between different classes. In the context of aerial imagery, we hypothesize that learning orientation-invariant representations can lead to more robust incremental learning, as the model can better capture the inherent structure and patterns in the data regardless of the specific orientation of objects. To this end, we propose a two-stage contrastive regularization scheme. In the first stage, we apply contrastive learning within each incremental step, encouraging the model to learn representations that are invariant to rotations and reflections of the input. This is achieved by comparing the activations of the model on transformed pairs of aerial images and minimizing their difference. In the second stage, we extend this idea to the incremental learning setting by comparing the activations of the current model with those of the previous model, thereby promoting consistency and stability across incremental steps. We evaluate the effectiveness of our proposed approach on the Potsdam dataset [199], a widely-used benchmark for aerial semantic segmentation. Through extensive experiments, we demonstrate that our contrastive regularization scheme consistently improves the performance of incremental learning, outperforming robust incremental baselines in various settings. We also provide insights into the learned representations and analyze the impact of different design choices on the model's performance. In short, the main contributions of these sections are: (i) an improved incremental learning framework for semantic segmentation exploiting the peculiar viewpoint of remote sensing imagery as regularization, (ii) an extensive set of experiments on the Potsdam dataset to provide empirical evidence of the effectiveness of our approach, and (iii) an analysis of the impact of the different design choices, highlighting strengths and limitations.

3.3.1 Related Works

Semantic segmentation of aerial images presents distinct challenges compared to natural images, such as arbitrary orientations and additional spectral bands beyond the visible spectrum. Deep learning techniques have been successfully applied to this domain [13, 166], addressing the multi-modal nature of the data through strategies like input weight duplication [177] or late fusion of features [181, 260].

Considering regularization methods, several solutions have been proposed to handle the rotational invariance of aerial images, improving classification performance [187, 236]. Our work leverages this property by applying contrastive regularization to both the segmentation and incremental learning stages, enhancing the segmentation and the knowledge distillation process. Incremental learning aims to sequentially learn new information while mitigating catastrophic forgetting [149] of previously acquired knowledge. Existing approaches include replay-based methods that utilize exemplars from old classes [193], parameter isolation techniques that selectively prune weights [143], and memory-based methods that consolidate important parameters [265]. Knowledge distillation, often employing a teacher-student paradigm, has emerged as one of the most effective strategies [124, 35]. For semantic segmentation of aerial images, hybrid approaches combining knowledge distillation and exemplar replay [217] and methods that reinforce internal representations across learning steps [75] have been proposed. Recognizing the unique challenges posed by the background class in semantic segmentation, unbiased losses and regularization have been introduced to address the associated distributional shift [35]. Contrastive learning has recently gained importance as a powerful technique for representation learning, nearly closing the gap between supervised and self-supervised approaches [155, 41, 93, 32] and even surpassing supervised methods by learning more robust representations [111]. The core idea is to cluster representations of similar samples together while pushing dissimilar instances apart. Contrastive learning is often applied using pretext tasks based on image transformations [155, 41], image reconstruction [167, 213], or cross-modal techniques [134, 228]. These auxiliary tasks can be combined with supervised objectives like semantic segmentation to improve performance [213, 228], handle low-resource settings [36], or incorporate additional modalities [180]. Inspired by these successes, we propose augmenting inputs and using the resulting representations to regularize the model during both standard training and knowledge distillation, promoting invariance to the applied transformations.

3.3.2 Method

Our approach builds upon the MiB framework [35], a robust baseline for ICL in semantic segmentation tasks, and introduces a novel Contrastive Distillation technique that leverages the unique viewpoint of aerial imagery to improve the model's ability to learn new classes while retaining knowledge of previously learned classes. The key contributions of our methodology lie in the introduction of a *contrastive regularization* term and a *contrastive distillation* term. The former exploits the orientation invariance property of remote sensing imagery by applying random transformations to the input images and enforcing the model to produce similar representations for the original and transformed images, to learn features that are robust to variations in orientation. The latter aims instead to transfer orientation-invariant knowledge from the previous model to the current model. This is achieved by minimizing the difference between the representations of the current model on the transformed image and the transformed features of the previous model on the original image. This enables the model to learn new classes while retaining knowledge of previously learned categories, improving on previous methods. In the following subsections we provide a detailed description of our methodology, including the baseline framework, contrastive distillation technique, the training procedure, and the experimental setup.

Problem Statement

Here we focus on the problem of Incremental-Class Learning (ICL) for Semantic Segmentation in the context of aerial images. The goal is to develop a model that can learn new classes incrementally without forgetting the previously learned knowledge. We consider a scenario where different subsets of data are provided sequentially, each containing a distinct set of labels. First, let us define Semantic Segmentation as a pixel-wise classification task, where each pixel x_i in an image $x \in \mathcal{X}$ with fixed dimensions $\mathcal{H} \times \mathcal{W}$ is assigned a label $y_i \in \mathcal{Y}$ representing its semantic category. In some cases, pixels may also be associated with a generic background class $b \in \mathcal{Y}$. The objective is to learn a function f_θ, where θ represents the learnable parameters, that maps the image space $\mathcal{X}$ to the pixel-wise label space $\mathcal{Y}$, i.e., $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times |\mathcal{Y}|}$. In the ICL setting, the learning process is divided into multiple sequential training phases, which we refer to as learning *steps*. At each step t, we receive a new training set $\mathcal{D}_t$ and a set of new labels $\mathcal{Y}_t$. The label space is expanded by combining the previous labels $\mathcal{Y}_{t-1}$ with the new labels $\mathcal{Y}_t$, resulting in an updated label set $\mathcal{C}_t = \mathcal{Y}_{t-1} \cup \mathcal{Y}_t$. The pixel-wise labels y_i in $\mathcal{D}_t$ belong to either one of the current categories $\mathcal{Y}_t$ or the generic background class b. The challenge lies in training a new model f_θ^t on the entire label set $\mathcal{C}_t$ while preserving the knowledge learned from previous steps. To achieve this, we derive the old labels from the outputs of the previous model $f_\theta^{t-1} : X \rightarrow \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times |\mathcal{Y}_{t-1}|}$ and learn the new labels through standard training on the current dataset $\mathcal{D}_t$. The ultimate goal is to obtain a single model f_θ^t that performs well on both old and new classes, i.e., $f_\theta^t : X \rightarrow \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times |\mathcal{C}_t|}$.

Baseline Framework

As a robust baseline for our methodology, we adopt the MiB framework proposed in [35]. The MiB approach addresses the challenge of catastrophic forgetting by addressing the main peculiarity of Semantic Segmentation tasks, namely the concept of *background* class. The core idea behind the MiB framework is to employ a dual-loss strategy that combines a supervised cross-entropy loss and a knowledge distillation loss. The supervised cross-entropy loss focuses on learning the new classes introduced in the current incremental step, while the knowledge distillation loss aims to preserve the knowledge acquired from previous classes. Formally, let $L_{CE}(\theta_t)$ denote the cross-entropy loss for the new classes at incremental step t, and let $L_{KD}(\theta_t)$ represent the knowledge distillation loss that transfers knowledge from the previous model $f_{\theta_{t-1}}$ to the current model f_{θ_t}. The overall loss function optimized by the MiB framework at each incremental step t is defined as:

$$L(\theta_t) = L_{CE}(\theta_t) + \lambda L_{KD}(\theta_t) \tag{3.6}$$

where λ is a hyperparameter that controls the balance between the two loss terms.

The knowledge distillation loss $L_{KD}(\theta_t)$ plays a crucial role: it is designed to minimize the discrepancy between the predictions of the current model and the predictions of the previous model for the old classes. Specifically, the knowledge distillation loss is computed as follows:

$$L_{KD}^{\theta_t}(x, y) = \frac{1}{N} \sum_{i \in x} \sum_{c \in Y_{t-1}} q_x^{t-1}(i, c) \log(q_x^t(i, c)) \tag{3.7}$$

where x is an input image, i denotes a pixel in the image, c represents a class label from the set of old classes Y_{t-1}, and N is the total number of pixels. The terms $q_x^{t-1}(i, c)$ and $q_x^t(i, c)$ represent the predicted probabilities of a pixel i belonging to the class c for the previous model $f_{\theta_{t-1}}$ and the current model f_{θ_t}, respectively.

This can be defined as a standard KD loss formulation, however one of the key challenges in ICL for semantic segmentation is the presence of a background class that is shared among different incremental steps. The contents of the background class can in fact vary significantly across different steps, leading to a distributional shift. In MiB, L_{CE} and L_{KD} are provided in an *unbiased* formulation, where the predicted probability for the background class is redefined as the sum of the probabilities of the old classes for the current model and the sum of the probabilities of the new classes for the previous model, as detailed in Eq. (3.8).

$$q_x^t(i, b) = \begin{cases} \sum_{k \in Y_{t-1}} q_x^t(i, k) & \text{for CE loss} \\ \sum_{k \in Y_t} q_x^t(i, k) & \text{for KD loss} \end{cases} \tag{3.8}$$

This modification helps to mitigate the bias induced by the background class and improves the model's ability to handle any variations.

Furthermore, the MiB framework employs a specific weight initialization strategy for the final classifier layer of the model: the weights corresponding to the new classes are initialized in a way that their outputs are uniformly distributed around the background class. This initialization scheme facilitates the convergence of the model during incremental learning steps and helps to prevent the bias towards new classes.

Contrastive Distillation

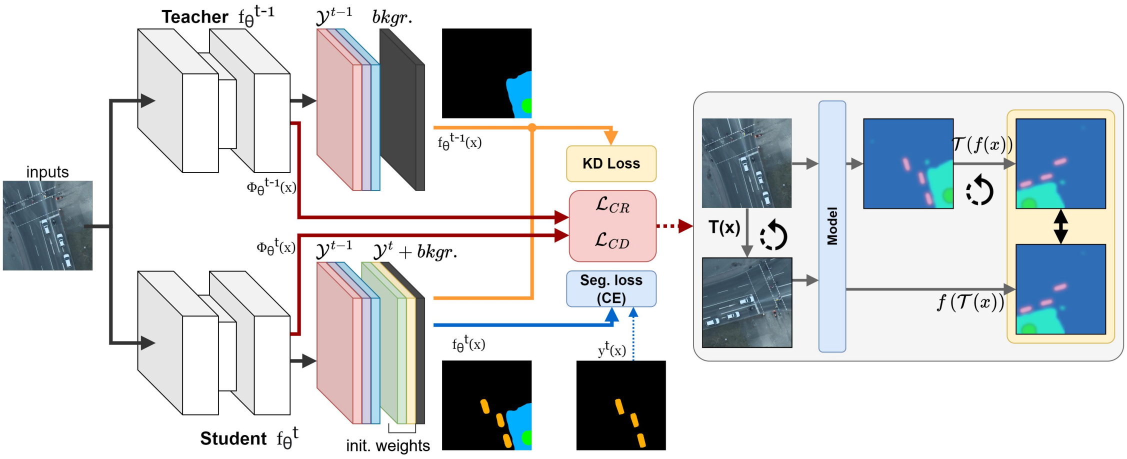


Figure 3.3: Schema representing our incremental learning framework. At each incremental step t, the pair composed of the input image x and the transformed $\mathcal{T}(x)$ is processed by both the previous model $f_{\theta_{t-1}}$ and the current model f_{θ_t}. The new classes are learned through supervised training using the available ground truth (depicted in blue), while the knowledge of old classes is transferred via distillation (shown in yellow). Additionally, the representations of the transformed inputs are contrasted with the transformed features of the original input, regularizing both the supervised training ($\mathcal{L}_{CR}$) and the distillation process ($\mathcal{L}_{CD}$) (illustrated in red).

Building upon this baseline framework, we introduce a novel contrastive distillation approach that leverages the unique characteristics of aerial imagery to enhance the incremental learning process. Similar to Section 3.2.2, our approach is motivated by the observation that aerial images exhibits an invariance to orientation: unlike natural images, where the orientation of objects is typically fixed and plays a crucial role in their recognition, aerial images capture scenes from a top-down perspective, resulting in arbitrary orientations of objects. This property suggests that the semantic content of an aerial image should remain consistent regardless of its orientation. We propose to exploit this orientation invariance property by incorporating a contrastive learning mechanism into the distillation process itself. Contrastive learning has emerged as a powerful paradigm for learning discriminative feature representations. It works by encouraging the model to produce similar representations for positive pairs of samples, while pushing apart the representations of negative pairs [41, 93]. In the context of ICL for semantic segmentation of aerial images, we adapt the contrastive learning principles to promote the learning of orientation-invariant features and improve the knowledge transfer across incremental steps. Our approach consists of two key components: a *Contrastive Regularization* (CR) term and a *Contrastive Distillation* term. The CR term is designed to encourage the model to learn features that are invariant to orientation, without yet considering the ICL context. Given an input aerial image x, we apply a random transformation $\mathcal{T}$ to obtain a transformed version of the image, denoted as $\mathcal{T}(x)$. The transformation $\mathcal{T}$ can include operations such as rotation and flipping, which preserve the semantic content of the image while altering its orientation, but also photometric distortions such as the ones applied in Section 3.2.3. We then extract the feature representations of both the original image and the transformed image using the current model f_{θ_t}, denoted as $\phi_{\theta_t}(x)$ and $\phi_{\theta_t}(\mathcal{T}(x))$, respectively. To enforce the learning of orientation-invariant features, we introduce a Contrastive Regularization term $\mathcal{L}_{CR}$ that minimizes the mean squared error (MSE) between the feature representations of the transformed image and the transformed features of the original image:

$$\mathcal{L}_{CR} = \frac{1}{N} \sum_{i=1}^{N} \left(\phi_{\theta_t}(T(x_i)) - T(\phi_{\theta_t}(x_i))\right)^2 \tag{3.9}$$

where N is the number of samples in the batch, and x_i denotes the i-th sample. By minimizing this term, we encourage the model to produce similar representations for the original and transformed versions of the image, thus boosting the learning of orientation-invariant features. In addition to the CR term, we propose a Contrastive Distillation term $\mathcal{L}_{CD}$ that aims to transfer the orientation-invariant knowledge from the previous model $f_{\theta_{t-1}}$ to the current model f_{θ_t}. Similar to the previous formulation, we apply the transformation $\mathcal{T}$ to the input image and extract the features using both the previous and current models. The CD term is defined as the MSE between the representations of the current model on the transformed image and the transformed features of the previous model on the original image. Formally:

$$\mathcal{L}_{CD} = \frac{1}{N} \sum_{i=1}^{N} \Big(\phi_{\theta_t}(T(x_i)) - T(\phi_{\theta_{t-1}}(x_i))\Big)^2 \tag{3.10}$$

By minimizing this term, we encourage the current model to learn feature representations that are consistent with the orientation-invariant knowledge captured by the previous model, improving the knowledge transfer across incremental steps. The overall loss function for our contrastive distillation approach is therefore defined as:

$$L(\theta_t) = \mathcal{L}_{CE}(\theta_t) + \lambda\mathcal{L}_{KD}(\theta_t) + \eta\mathcal{L}_{CR}(\theta_t) + \rho\mathcal{L}_{CD}(\theta_t) \tag{3.11}$$

where $\mathcal{L}_{CE}(\theta_t)$ and $\mathcal{L}_{KD}(\theta_t)$ are the cross-entropy loss and knowledge distillation loss from the MiB framework, respectively, and η and ρ are hyperparameters that control the contribution of the contrastive regularization and distillation terms.

3.3.3 Experiments

In our experiments, we evaluate the effectiveness of our proposed contrastive distillation approach on the Potsdam dataset, a widely-used benchmark for semantic segmentation of aerial images. The Potsdam dataset is a high-resolution aerial image dataset captured over the city of Potsdam, Germany, and is part of the ISPRS 2D Semantic Labeling Contest [199]. The dataset consists of 38 patches, each covering an area of 6000 × 6000 pixels with a ground sampling distance (GSD) of 5 cm. The dataset includes multiple modalities, such as RGB (true color), infrared (IR), and digital surface model (DSM) data, offering rich spectral and spatial information for semantic segmentation tasks. The Potsdam dataset provides pixel-wise annotations for six semantic classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter. These classes represent the main urban land cover categories commonly found in aerial imagery. The dataset is highly detailed and captures the complex urban landscape of Potsdam, including a mix of residential, commercial, and industrial areas, as well as green spaces and transportation networks. Here, we leverage the RGB and infrared modalities of the Potsdam dataset to evaluate our proposed solution under different input settings.

Implementation Details

We build upon the MiB framework [35] as a strong baseline for ICL in semantic segmentation, however we note that the simplicity of this approach allows its application in any context. To simulate the incremental learning scenario, we divide the Potsdam dataset into multiple incremental steps, each introducing new semantic classes. We consider two different experimental configurations: (i) a three-step scenario named *(3S)*, where the initial step includes the classes *building* and *tree*, the second step adds *impervious surfaces* and *low vegetation*, and the final step introduces the *car* class; and (ii) a five-step scenario (*5S*), where each class is introduced sequentially in this order: *building*, *tree*, *impervious surfaces*, *low vegetation*, and *car*. For the 5S configuration, we exclude the *clutter* class from our experiments since it is not part of the official benchmark evaluation [199]. To ensure a fair evaluation, we create a disjoint split of the dataset for each incremental step, such that each split contains only a single semantic class. This setup guarantees that the model is exposed to new and unseen images at each step, providing a more challenging and realistic incremental learning scenario. We employ an encoder-decoder architecture based on the Res-UNet model [59] as the backbone for our experiments. To optimize the memory requirements, we replace the ResNet encoder with a more efficient TResNet architecture [196] pretrained on ImageNet. For experiments involving multi-modal data (e.g., RGB+IR), we adapt the input layer of the network by duplicating the weights of the red channel, following the approach briefly described in Section 3.2.3 [177]. We train the model for 80 epochs at each incremental step using the AdamW optimizer with a learning rate of 10^{-3} and a cosine annealing scheduler. The learning rate is reduced to 10^{-4} for the final steps to facilitate fine-tuning. We employ a batch size of 8, and due to the contrastive augmentation, the effective batch size is effectively doubled to 16 at runtime. Given the large size of the Potsdam patches, we divide them into smaller tiles of size 512×512 pixels with an overlap of 12 pixels to ensure a balance between computational efficiency and spatial context. Besides the additional regularization, we extensively apply data augmentation to further enhance the robustness of the final model. We focus on geometric transformations, including random flipping, random rotations with varying degree and around a randomly selected center, and random scaling factors to zoom in and out. Given the remote sensing context, we further apply reflection padding to the areas of the images that fall outside the frame after these transformations. Following previous experiments [59], this method leverages the symmetry commonly found in urban aerial images, making it particularly effective for our application. For the CR and CD terms, we empirically set the weight factors η and ρ to 0.1 in order to introduce these soft penalties while keeping the focus on CE and KD during training. In terms of augmentations, we use random vertical flipping, horizontal flipping, and 90-degree rotations as regularization. We allocate 15% of the training set as the validation set to monitor the model's performance during training. To evaluate the performance of our approach, we use the F1 score as in Eq. (2.5) as the primary metric following previous works [59], computed on the official test set of the Potsdam dataset. Similar to the IoU metric, the F1 score provides a balanced measure of precision and recall, making it suitable for assessing the quality of the semantic segmentation results.

Results

Method	Building	Tree	Clutter	Surf.	Low veg.	Car	Avg.
MiB (RGB)	0.9116	0.8217	0.2766	0.8918	0.7589	0.8500	0.7517
MiB + CRCD (RGB)	0.9209	0.8085	0.3119	0.9021	0.7619	0.8541	**0.7599**
MiB (RGBIR)	0.8708	0.8062	0.2682	0.8773	0.7414	0.8176	0.7303
MiB + CRCD (RGBIR)	0.9178	0.8190	0.3128	0.8950	0.7635	0.8515	**0.7598**

Table 3.5: Class-wise and average F1 scores obtained after three incremental steps using the 3S configuration, where vertical lines separate the label groups introduced at each step. We compare the performance of the MiB baseline and our proposed regularizations (MiB + CRCD) using both RGB and multi-spectral (RGBIR) input modalities.

Table 3.5 presents the class-wise and average F1 scores obtained after the three incremental steps in the 3S configuration. The results demonstrate that the MiB

Method	Building	Tree	Surfaces	Low veg.	Car	Avg.
MiB (RGB)	0.8451	0.7449	0.7912	0.7011	0.6759	0.7810
MiB + CRCD (RGB)	0.9015	0.7515	0.8848	0.7313	0.8287	**0.8195**
MiB (RGBIR)	0.8564	0.7007	0.8575	0.6862	0.8228	0.7847
MiB + CRCD (RGBIR)	0.8770	0.7740	0.8755	0.7343	0.8437	**0.8209**

Table 3.6: Class-wise and average F1 scores obtained after five incremental steps using the 5S configuration. We evaluate the performance of the MiB baseline and our proposed approach (MiB + CRCD) using RGB and multi-spectral (RGBIR) input modalities.

baseline achieves a competitive performance, indicating its effectiveness as a framework specifically designed for semantic segmentation tasks. However, our contrastive distillation approach consistently improves upon the MiB baseline in every experiment and across all incremental steps. These improvements are observed for both RGB and multi-spectral (RGBIR) input settings. It is worth noting that in the simpler 3S configuration, the RGB baseline performs on par with the regularized version. We hypothesize that this is due to both the effectiveness of the standard approach and the robustness of the backbone network pretrained on RGB images, which could also justify the performance drop with the additional infrared channel. However, in more challenging scenarios such as the 5S configuration, the contribution of the additional regularization becomes more pronounced. As shown in Table 3.6, our contrastive distillation approach achieves a significant improvement over the MiB baseline, with an average increase of approximately 4% in the F1 score.

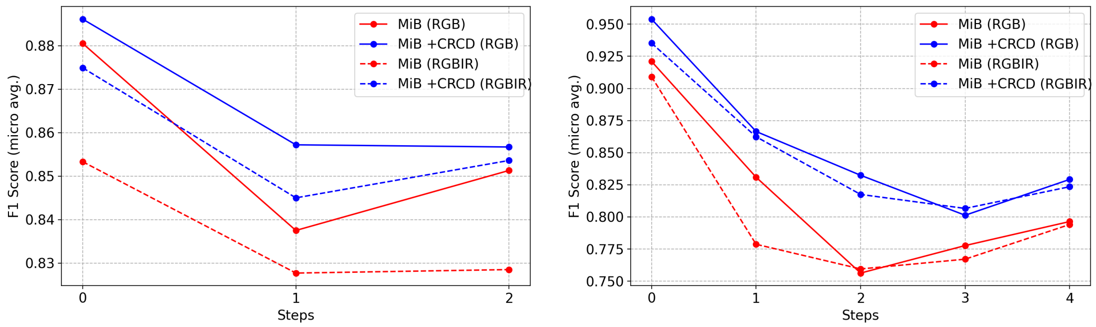


Figure 3.4: Micro-averaged F1 scores over the incremental steps for the 3S configuration (left) and 5S configuration (right). The proposed CRCD solution often outperforms the strong MiB baseline, with the multi-spectral (RGBIR) variant (dashed lines) providing further improvements.

Fig. 3.4 illustrates the micro-averaged F1 scores across the incremental steps for both the 3S and 5S configurations. The results highlight the performance improvements of our CRCD solution over the MiB baseline, especially in the latter setup. The proposed method consistently achieves higher F1 scores throughout the incremental learning process, indicating its effectiveness in mitigating catastrophic forgetting and enabling the model to successfully learn new classes while retaining knowledge of previously learned classes. This is true in both the multi-spectral (RGBIR) and standard (RGB) scenario, thus regardless of the spectral information.

Table 3.7: Ablation study results on the 5S configuration, as class-wise and average F1 scores for the last incremental step.

$\mathcal{L}_{CE}$	$\mathcal{L}_{KD}$	$\mathcal{L}_{CR}$	$\mathcal{L}_{CD}$	**Building**	**Tree**	**Imp. surf.**	**Low veg.**	**Car**	**Average**
				0.0	0.0	0.0	0.0	0.8708	0.1742
✓		✓	✓	0.6118	0.4927	0.6924	0.2909	0.5275	0.5231
✓	✓			0.8491	0.7625	0.8480	0.6751	0.7703	0.7810
✓	✓	✓		0.8178	0.7452	0.8514	0.6781	0.8186	0.7822
✓	✓		✓	0.9079	0.7522	0.8815	0.7011	0.7895	0.8064
✓	✓	✓	✓	0.9015	0.7515	0.8848	0.7313	0.8287	**0.8196**
Offline				0.9510	0.8535	0.9063	0.8415	0.8942	0.8893

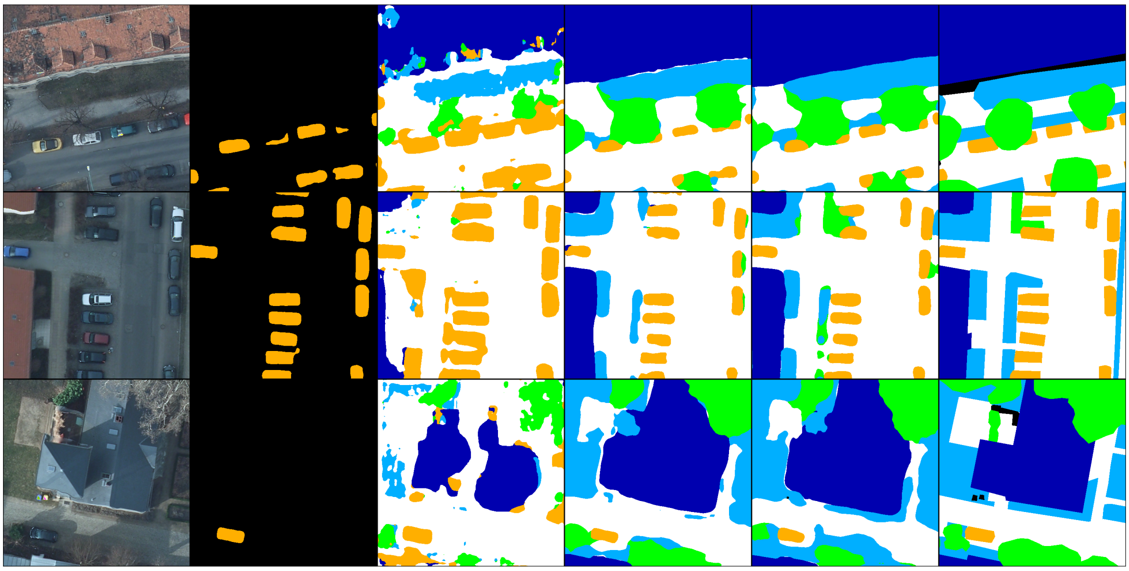


Figure 3.5: Qualitative comparison of segmentation results. From left to right: input image in RGB format, a standard finetuning, finetuning with CD and CR terms, MiB baseline, MiB with regularization terms (MiB + CRCD), and ground truth. Our approach produces more accurate and coherent segmentation maps compared to the baseline.

Ablation study. To better assess the individual contributions of the proposed regularization terms and their impact on the overall performance, we conduct an ablation study. Table 3.7 presents the results of this study, focusing on the 5S configuration with RGB input modality. We assign a column for each of the necessary regularization terms, namely unbiased Cross-Entropy ($\mathcal{L}_{CE}$), unbiased Knowledge Distillation loss ($\mathcal{L}_{KD}$), Contrastive Regularization ($\mathcal{L}_{CE}$), and Contrastive Distillation ($\mathcal{L}_{CD}$). We begin by establishing a simple finetuning baseline, where the model is trained on new classes without considering any previous knowledge and no regularizing factors. As expected, this approach leads to catastrophic forgetting, resulting in poor performance on all classes except the last one. The finetuning baseline serves as a lower bound, demonstrating the necessity of employing techniques to mitigate forgetting in incremental learning contexts. To further validate the distillation capabilities of our regularization terms, we conduct an additional experiment where we apply finetuning with $\mathcal{L}_{CE}$ and our approach, considering both CR and CD, without the actual distillation loss. The results demonstrate that the proposed regularization terms actively contribute to the preservation of previous knowledge, even in the absence of explicit distillation. Next, we evaluate the performance of the MiB framework, which incorporates knowledge distillation and unbiased cross-entropy loss to address the challenges of incremental learning. The MiB framework achieves significantly better results compared to the finetuning baseline, with an average improvement of over 60%. This highlights its effectiveness as a strong baseline for ICL in semantic segmentation tasks. To investigate the impact of our proposed regularization terms on top of that, we start by naively introducing the $\mathcal{L}_{CR}$ term, which focuses on the current incremental step only. While this regularization improves the performance on the last class, as expected, it has a negative effect on the previously seen categories, which are not explicitly considered. This underscores the importance of addressing the knowledge transfer across incremental steps. On the other hand, applying the $\mathcal{L}_{CD}$ term alone, which compares the activations between the current and previous models, yields higher scores for the previous categories. This regularization term facilitates the transfer of knowledge from the old model to the new one, resulting in an average score increase of 2%. However, it does not provide any performance boost for the classes introduced in the current incremental step. The best performance is obtained when combining both regularization terms, $\mathcal{L}_{CR}$ and $\mathcal{L}_{CD}$. By jointly optimizing for consistency within the current step and knowledge distillation across incremental steps, we achieve an improvement of around 4% over the baseline. This performance boost is observed for both the current and previous classes, demonstrating the effectiveness of the combination of our regularization terms. Last, it is worth noting that the results obtained by our approach are close to the theoretical upper bound of the offline setting, where the model is trained on the entire dataset at once.

Computational cost. Similar to AI+AS (Section 3.2), including two additional regularization factors inevitably introduces a performance overhead. Nevertheless, once again, the two penalties are applied during a standard segmentation loss pass, and distillation pass, duplicating the batch into two comparable sets of image. Except for the unavoidable additional memory usage, the actual computational cost can be minimized by keeping every processing on the GPU device.

3.4 Summary

In this chapter, we explored the unique challenges posed by the aerial viewpoint in semantic segmentation tasks and proposed novel techniques to address them. We focused on two key aspects: the arbitrary orientation of scenes captured from a top-down perspective and the severe class imbalance commonly found in aerial imagery datasets. In Section 3.2, we introduced a framework that combines Augmentation Invariance (AI) regularization with an Adaptive Sampling (AS) strategy. The AI component guides the model to learn semantic representations that are invariant to photometric and geometric distortions, while the AS technique addresses class imbalance by dynamically selecting training samples based on class distribution and model confidence. Experiments on the Agriculture-Vision dataset demonstrated the effectiveness of this approach, consistently outperforming baseline methods. The ablation studies and qualitative results confirmed the importance of explicitly addressing the unique characteristics of aerial imagery to improve performance and generalization capabilities. We further extended these techniques to the context of incremental learning in Section 3.3. We proposed a contrastive distillation approach that compares the segmentation features of an input image and its augmented version to enforce invariance to orientation changes. This not only improves the model's robustness to different aerial viewpoints but also enhances the model distillation phase, enabling the incorporation of new classes without forgetting previously learned features. The experimental results on the Potsdam dataset highlighted the effectiveness of the contrastive distillation in incremental learning scenarios, outperforming robust incremental baselines in various settings. The findings presented in this chapter provide valuable insights and effective solutions for semantic segmentation in aerial imagery, both in standard and incremental learning settings. By leveraging invariance to photometric and geometric transformations, as well as tackling class imbalance, the proposed techniques consistently improve performance and offer a promising direction for future research in this field. However, we acknowledge that the proposed methods were evaluated on Agriculture-Vision and Potsdam. Future works may be aimed at evaluating their effectiveness on other aerial imagery datasets with different characteristics, or at assessing which augmentation techniques and hyperparameters may require further optimization for optimal performance in different scenarios. Another interesting direction for future work is the investigation of similar techniques in unsupervised or semi-supervised learning approaches, to leverage the vast amounts of unlabeled aerial imagery available, reducing the reliance on large-scale annotated datasets.

Chapter 4

Class and Scale Imbalance in Remote Sensing

4.1 Introduction

In the context of remote sensing data, we often face the challenge of class and scale imbalance, which can significantly impact the performance of machine learning models. Class imbalance refers to the uneven distribution of categories within a dataset, where certain groups are disproportionately represented compared to others. On the other hand, scale imbalance pertains to the variations in the size and spatial extent of objects or regions of interest within the image. Addressing both class and scale imbalance is crucial for developing robust and accurate models: large-scale datasets often exhibit class unbalance with a power-law probability distribution [11] and some categories only provide a handful of samples. Without countermeasures, the model may tend to favor the majority class, leading to biased predictions and poor performance on the minority classes. Similarly, scale unbalance can affect the model's ability to effectively capture and represent objects at different scales, resulting in suboptimal feature extraction and bad generalization capabilities. In this chapter, we tackle these challenges focusing on two widely different application scenarios: flood delineation applied to satellite images, and photovoltaic panel delineation from VHR aerial imagery.

In the first case, we construct our own dataset, named *MMFlood*, using Sentinel-1 as main input and the Copernicus Emergency Management System (CEMS) activations as ground truth source [64]. Given the nature of the problem at hand, the dataset exhibits a significant class imbalance, with the majority of pixels representing non-flooded areas and only a small portion corresponding to flooded regions. This is inherent to the nature of flood events, where the affected areas typically cover a limited portion of the land surface compared to the overall extent of the analyzed region. To address the class imbalance problem in flood delineation, we investigate various strategies, including an Entropy-Weighted Sampling (EWS) technique.

EWS leverages the concept of information entropy to assign higher importance to samples with more informative content. By considering the distribution of pixels among the flooded and non-flooded classes within each label, EWS calculates a weight for each sample based on its information content. Samples with higher entropy are given higher weights, indicating their significance in the training process. This weighted sampling strategy enables the model to learn more effectively from the available data and improves its ability to accurately delineate flooded regions. We evaluate the goodness of this approach on MMFlood, a multimodal dataset specifically designed for flood delineation, that combines Synthetic Aperture Radar (SAR) imagery from Sentinel-1 with additional modalities such as Digital Elevation Model (DEM) and hydrography maps. The inclusion of these complementary data sources provides additional context of more accurate delineations. On this dataset, we carry out a comprehensive benchmark evaluation using standard deep learning model architectures, comparing the introduced methodologies.

In the second application scenario, we focus on the delineation of photovoltaic (PV) panels from VHR aerial imagery. The main challenge of this task is instead related to scale imbalance, as PV panels appear in a wide range of scales, from small residential components to industrial-grade plants spanning entire fields. To address this issue, we start by constructing a tailored dataset covering the Piedmont region in Italy, comprising 105 large-scale aerial images and more than 9,000 manual annotations. We then propose ad-hoc modifications to the semantic segmentation model architecture, introducing a local-contextual training paradigm that guides the model towards multiscale consistency by encouraging high-resolution and low-resolution features to be similar. We evaluate the effectiveness of our approaches through a comprehensive benchmark on the custom PV panel dataset, comparing the performance of different semantic segmentation architectures and assessing the impact of our modifications. The benchmark results demonstrate the benefits of the multiscale training paradigm and the inclusion of additional modalities, such as the infrared band, on the final scores. Furthermore, we introduce a post-processing algorithm tailored for PV panel delineation, which leverages prior knowledge about their geometry to refine the segmentation output and produce cleaner and more precise boundaries. In summary, this chapter presents contributions that led to the publication of the following works:

- Montello F., Arnaudo E., Rossi C., *MMFlood: A Multimodal Dataset for Flood Delineation from Satellite Imagery*, IEEE Access, vol. 10, pp. 96774-96787, 2022.
- Arnaudo E., Blanco G., Monti A., Bianco G., Monaco C., Pasquali P., Dominici F., *A Comparative Evaluation of Deep Learning Techniques for Photovoltaic Panel Detection From Aerial Images*, IEEE Access, vol. 11, pp. 47579-47594, 2023.

4.2 Class Unbalance in Flood Delineation

Floods are among the most devastating natural disasters, causing significant damage to infrastructure, agriculture, and human lives. The increasing frequency and intensity of flood events due to climate change [147] have emphasized the need for accurate and timely flood delineation methods. Synthetic Aperture Radar (SAR) imagery has emerged as a valuable tool for this purpose, thanks to its ability to penetrate clouds and operate in all weather conditions [24]. However, flood delineation from SAR imagery remains a challenging task due to several factors. First, the complex interactions between SAR signals and the Earth's surface lead to speckle noise (see Section 2.2.3) and other artifacts, hindering accurate delineation [146]. Second, the heterogeneous nature of flood events occurring in different landscapes and environmental conditions makes it difficult to develop robust and generalizable algorithms. Finally, and most importantly, the class imbalance problem, where the number of pixels representing flooded areas is much smaller than the number of non-flooded pixels, can lead to biased models that underestimate the extent and severity of flood events. To address the class imbalance problem in flood delineation, we apply a sampling mechanism, Entropy Weighted Sampling (EWS). EWS is a technique that assigns higher probabilities to more informative samples, i.e., those with higher entropy, during the training process. By focusing on the inputs with more information content, EWS aims to mitigate the bias towards the majority class and improve the model's performance on the minority class. In the context of flood mapping, EWS can help the model better learn the characteristics of flooded areas, even when they represent a small fraction of the total pixels in the dataset. To evaluate the effectiveness this sampling mechanism in the context of flood delineation, we introduce MMFlood, a large-scale multi-modal dataset featuring SAR imagery, digital elevation models (DEMs), and hydrography information for a diverse set of events worldwide, sourced from the Copernicus EMS [64] platform. MMFlood provides a challenging benchmark for machine learning algorithms, as it encompasses a wide range of geographical regions and environmental conditions, and exhibits significant this inherent class imbalance. In this work, we conduct extensive experiments on the MMFlood dataset to assess the performance of EWS in combination with standard deep learning architectures for semantic segmentation. We compare the results obtained using EWS with those of standard training strategies and demonstrate the benefits of our approach in terms of improved accuracy and robustness on the minority class. Furthermore, we explore the potential of leveraging the complementary information provided by the multiple data modalities in MMFlood, namely SAR imagery and DEM, exploiting a deep architecture that effectively fuses these data sources to enhance the overall performance. Relevant code and the associated datasets described in this section are available at https://github.com/edornd/mmflood.

4.2.1 Related Works

Dataset	Source	Modalities	Geoloc.	Task	Images	Img. size	Res.
Vaihingen [199]	Aerial	RG-IR, DSM	✓	S	33	<2,500×2,000	10cm
Potsdam [199]	Aerial	RGB-IR, DSM	✓	S	38	6,000×6,000	10cm
BigEarthNet [216]	S2	12 bands	✓	C	590,326	120×120	10-60m
LandCoverNet [21]	S2	12 bands	✓	C	9,000	256×256	10-60m
Agriculture-Vision [49]	Aerial	RGB-NIR	×	S	94,986	512×512	10-20cm
HRSID [248]	S1, TerraSAR-X	HH,VV,HV	✓	OD	5,604	800×600	1-5m
DeepGlobe [57]	Maxar	RGB	×	S	1,146	2,448×2,448	50cm
xBD [89]	Maxar	RGB	×	OD	9,168	1,024×1,024	80cm
FloodNet [189]	Aerial	RGB	×	S, C, VQA	2,343	4,500×3,000	1.5cm
SEN12FLOOD [190]	S1, S2	VV-VH (S1), 12 bands (S2)	✓	C	336*	512×512	10-60m
sen1floods11 [22]	S1, S2, JRC	VV-VH (S1), 12 bands (S2), Hd.	✓	S	4,831	512×512	10-60m
ETCI-2021 [102]	S1, NASA	VV-VH, Hd.	×	S	33,405	256×256	20m
MM-Flood	S1, MapZen, OSM	VV-VH, DEM, Hd.	✓	S	1,748	<2,000x2,000	20m

Table 4.1: Comparison between MMFlood and other remote sensing datasets, highlighting their characteristics and limitations with focus on flood mapping. The table presents information on data sources (S1: Sentinel-1, S2: Sentinel-2), modalities (RGB, DSM, DEM, SAR polarizations, hydrography), georeferencing, task types (C: classification, S: segmentation, OD: object detection, VQA: visual question answering), number of images, image sizes, and spatial resolutions.

Datasets. Despite the growing interest in remote sensing applications, the availability of large-scale aerial and satellite datasets remains limited, if we consider specific tasks such as flood delineation. Many of the most widely used datasets in the Earth Observation (EO) domain, such as the Vaihingen and Potsdam datasets [199] described in previous sections, focus primarily on urban land cover classification and its related tasks. These datasets often cover a single area, or contain a small number of images. Some larger datasets, including BigEarthNet [216], or DeepGlobe [57], aim to address more general-purpose tasks. However, they still have some limitations and limited reusability. BigEarthNet, for instance, provides coarse labels for multiple land cover classes across ten countries, but does not include annotations for semantic segmentation. Similarly, DeepGlobe covers various geographical areas and offers pixel-level semantic annotations for seven categories, but lacks location data, is limited to the visible spectrum, and provides a simplified taxonomy. When considering datasets specifically designed for disaster management, the list of available resources is even shorter. Datasets like xBD [89] and FloodNet [189] provide annotated images for flood events, but they are limited to optical imagery and may lack geographical diversity. A major drawback remains the absence of SAR data, which is particularly useful in contexts where cloudy weather is expected. Among the few datasets that include SAR imagery for flood delineation, SEN12-FLOOD [190] and Sen1floods11 [22] are notable examples. However, SEN12-FLOOD covers a limited number of flood events and geographical regions, while Sen1floods11, despite its worldwide coverage, relies on a mix of manual and automated annotations, which may result in incomplete or noisy ground truth data. Additionally, the reliance on Sentinel-2 imagery for flood delineation in Sen1floods11 can lead to missing annotations due to cloud coverage. Another dataset worth mentioning is the ETCI 2021 dataset [102], which contains a large number of SAR images from five different geographical regions. Despite their undeniable value, most flood mapping datasets have some limitations. Firstly, many existing datasets contain a limited number of images representing actual flood events, with the majority of the data depicting water basins. Secondly, the absence of a Digital Elevation Model (DEM) can hinder the models' ability to distinguish between flooded areas and permanent water bodies [113, 168]. To address the limitations of existing datasets, we introduce MMFlood, a large-scale multi-modal dataset specifically designed for flood delineation using SAR imagery. Our dataset leverages the CEMS activations as ground truth, generating high-quality masks from activations that have been produced using SAR imagery, manually validated by experts, and most importantly, only include flooded areas. In addition to the SAR imagery and flood masks, MMFlood incorporates DEM and hydrography data when available, covering the same areas of interest. These can be provided to the models as additional inputs, or simply applied in post-processing to further clean the final result. To provide a more comprehensive overview of the limitations and characteristics of existing datasets, we present a comparative analysis in Table 4.1, highlighting the key features and shortcomings of each dataset in the context of flood mapping.

Flood delineation. Flood delineation works encompass a wide range of techniques, utilizing various algorithms and data sources. Among the satellite instruments available, Synthetic Aperture Radar (SAR) data from networks such as TerraSAR-X [116], RADARSAT [192], COSMO-SkyMed [25], and Sentinel-1 [221] have been extensively used in the flood mapping literature. Sentinel-1, in particular, has emerged as one of the most convenient options due to its worldwide coverage at medium-high spatial resolution, short revisit times, and open data availability. Early approaches in this field primarily relied on masking and thresholding techniques, coupled with meticulous data preprocessing [146, 233, 138, 7] or Fuzzy Logic approaches [226, 186, 145]. As Artificial Intelligence and Deep Learning techniques gained traction in the Computer Vision domain, numerous supervised machine learning classifiers were developed and applied to these tasks, including Support Vector Machines [103, 20], Fully Convolutional Neural Networks [106], Bayesian Networks [54], Deep Belief Networks [17], and Random Forests [176]. Currently, deep learning solutions have primarily concentrated on flood delineation at ground level [263] or through drone and aerial imagery [189]. However, significant research has been conducted on remote sensing SAR data for various applications, such as image despeckling [119, 159], detection of large objects like ships [37, 248], and land cover classification [216, 225]. SAR imagery is frequently augmented with information derived from additional sources, including optical data like Sentinel-2 [216] or even vastly different modalities such as Automatic Identification Systems (AIS), which are often exploited in vessel detection through domain adaptation [118] or ad-hoc data fusion [79]. While classical machine learning techniques provide robust results across various tasks, Convolutional Neural Networks (CNN) remain the dominant architecture in this field. Encoder-decoder modules such as U-Net [176] and multiscale extraction networks like DeepLab [39] have been successfully employed for semantic segmentation tasks in remote sensing imagery. These architectures have demonstrated their effectiveness in capturing spatial and contextual information, enabling accurate delineation of flooded areas.

4.2.2 Dataset

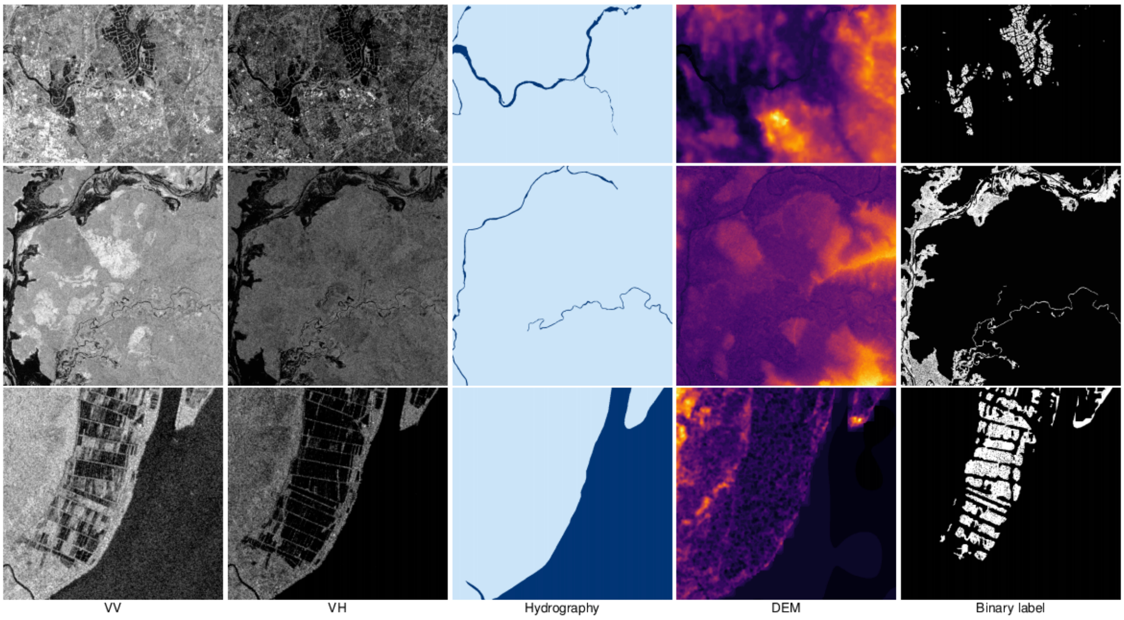


Figure 4.1: Samples from the MMFlood dataset, showcasing the variety of flood events and data modalities. Each row represents a different flood event, while the columns depict the various data types: Sentinel-1 SAR imagery (VV and VH polarizations), hydrography information, Digital Elevation Model (DEM), and the corresponding binary flood mask.

The foundation of the MMFlood dataset lies in the Copernicus Emergency Management Service (EMS) [64], which serves as the primary source for identifying and delineating flooded areas. Copernicus, an EU program aimed at developing European information services based on satellite Earth Observation (EO) and in-situ data, monitors and forecasts the state of the environment to support climate

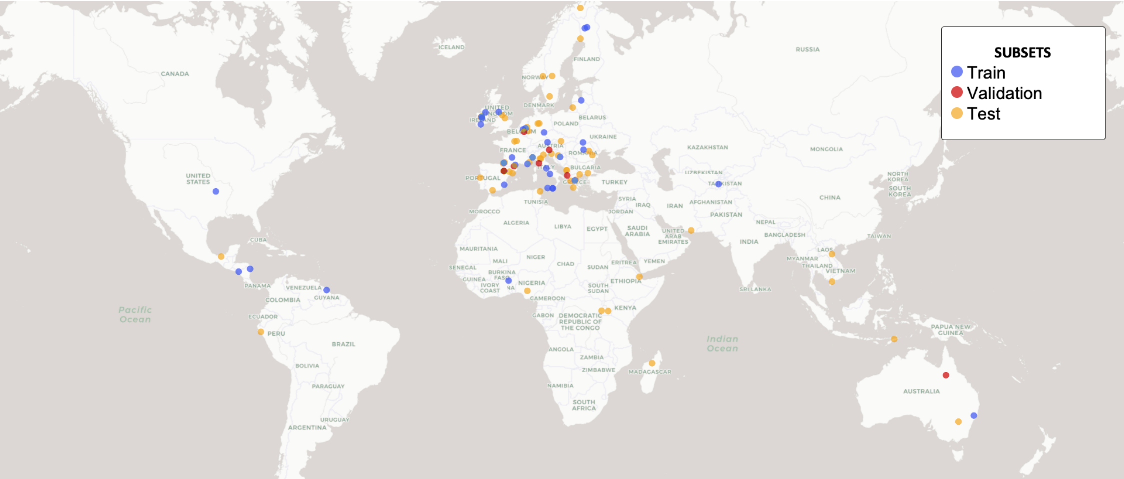


Figure 4.2: Geographical distribution of flood events in the MMFlood dataset. The map depicts the locations of the major floods derived from Copernicus EMS activations, with different colors representing the dataset splits: training (blue), validation (red), and testing (yellow).

Italy	11	Sweden	3	Djibouti	1	Mexico	1
France	10	Nicaragua	2	Slovenia	1	Iran	1
Spain	7	Netherlands	2	Portugal	1	Nigeria	1
Germany	6	Norway	2	Croatia	1	Madagascar	1
Greece	6	Vietnam	2	Moldova	1	Belgium	1
Ireland	5	Uganda	2	Lithuania	1	Austria	1
United Kingdom	4	Ukraine	2	Timor-Leste	1	Slovakia	1
Australia	3	Latvia	1	Guyana	1	Bosnia and Hze.	1
Albania	3	Tunisia	1	Peru	1	United States	1
Finland	3	Belgium	1	Tajikistan	1		
Romania	3	Togo	1	Honduras	1		

Table 4.2: Countries present in the MMFlood dataset and their number of activations. The dataset covers a total of 42 countries and 96 events, with Italy, France, and Spain having the highest number of mapped flood events.

change mitigation and adaptation strategies. The EMS, one of the information tools provided by Copernicus, offers a comprehensive collection of geospatial information related to various disaster types, including meteorological and geophysical hazards, man-made disasters, and humanitarian crises. To construct the MMFlood dataset, we programmatically retrieve the vector packages containing flood delineation products from the EMS Rapid Mapping service. These packages, comprising sets of geographical files, are used to assess the extent of flood events and serve as ground truth for training machine learning models. Due to inconsistencies in package structure, naming conventions, and geometry, a small portion of the available activations were discarded. Subsequently, we conducted a thorough manual inspection of the remaining activations to ensure correctness and uniformity. The final dataset encompasses a total of 95 individual flood events spanning seven years, from 2014 to 2021, and covering 42 different countries, as detailed in Table 4.2. The geographical distribution of the selected events is depicted in Fig. 4.2, with different colors representing the various dataset splits. Although the Copernicus products inherently focus on European regions, with Italy, France, and Spain having the highest number of activations, a substantial portion of the data is distributed across the globe. South America, with two unique flood events, is the least represented continent in terms of absolute numbers. For each EMS activation, we extract essential information to provide accurate annotations and contextual details. This information includes the estimated flood start date, the general location of the area of interest, the bounding box of the event, and the polygons delineating the actual flooded areas. As each activation often groups together multiple disasters occurring in the same region and caused by the same agent, it may contain one or more vector packages defining different flooded sub-regions within the event. To maximize the dataset's coverage, we consider each sub-region separately during the image acquisition and rasterization phase. The flood polygons provided by the EMS have undergone manual validation by both the service provider responsible for generating the delineations and the Joint Research Centre (JRC) services to ensure a high level of accuracy [5]. From these polygons, we automatically derive the minimum bounding box fully enclosing the flooded areas, which is necessary for the image retrieval process. To account for potential inaccuracies in the delineations and to include contextual information, we expand the bounding box in each direction by a variable amount corresponding to 10% of its maximum extent.

From the obtained polygons, we can derive a raster dataset by downloading the corresponding remote sensing images in that specific window in time and space. Our primary focus is on Synthetic Aperture Radar (SAR) imagery due to its unique properties that make it particularly suitable for flood delineation tasks, especially in cloudy environments. We select Sentinel-1, as it offers worldwide coverage, relatively short revisit times, and open access to its data, making it an ideal choice for our dataset. The Sentinel-1 constellation consists of two twin satellites, Sentinel-1A and Sentinel-1B, which share the same orbital plane and carry a C-band SAR instrument. The revisit time varies depending on the latitude, with an estimated 3 days at the equator and less than 1 day at the poles. For the purpose of flood delineation, we focus on Level-1 Interferometric Wide (IW) Ground Range Detected (GRD) products. These provide the signal amplitude without the phase information, maintaining an approximately square spatial resolution and pixel spacing without further resampling. Additionally, the multi-look processing applied to these products reduces speckle noise, which is beneficial for our task. To streamline the image retrieval process and ensure that we only acquire data for the desired areas and time periods, we leverage the Sentinel-Hub platform [212]. This third-party service offers advanced filtering and compositing capabilities, providing direct access to various open remote sensing data sources. For each bounding box obtained from the Copernicus EMS activations, we retrieve the corresponding raw SAR signal in two polarization modes: VV (Vertical transmit and Vertical receive) and VH (Vertical transmit and Horizontal receive). To minimize discrepancies between the SAR imagery and the flood delineations, we restrict the acquisition window to a maximum of 4 days from the reported event date. In cases where a single Sentinel-1 tile does not fully cover the area of interest, we utilize the mosaicking features of Sentinel-Hub to automatically merge neighboring acquisitions. In the rare event that adjacent tiles are not available, we mask out the missing pixels in both the SAR acquisition and the corresponding ground truth to maintain consistency. The resulting images are provided as orthorectified, georeferenced GeoTIFF files with a spatial resolution of 20m/pixel and 32-bit floating-point precision. To enhance the contextual information available for flood delineation, we incorporate additional modalities into the dataset. We collect Digital Elevation Model (DEM) data for the same areas of interest using Mapzen's terrain tiles, which are primarily based on the Shuttle Radar Topography Mission (SRTM30) [229]. The DEM data provides information about the terrain elevation, which in turn should aid the identification of areas more susceptible to flooding. Although the DEM data is static and may not capture small variations in terrain due to concurrent natural events like landslides, it offers wider coverage and reduced noise compared to dynamic data sources. We resample the DEM data to match the 20m/pixel resolution of the SAR imagery and include it as a separate single-band GeoTIFF file. Furthermore, we extend the dataset by including hydrography maps of the areas of interest. Given the global scope of the EMS activations, we utilize OpenStreetMap (OSM) data for this purpose. By leveraging the OSM Overpass API, we extract polygons representing water layers within the image bounds, whenever available. To maintain pixel-wise alignment among the dataset tuples, we also rasterize the hydrography vectors at 20m/pixel, matching the resolution of the SAR and DEM data. While the availability and quality of hydrography data may vary across different regions, we successfully obtain hydrography rasters for more than half of the SAR acquisitions. To ensure the quality and relevance of the dataset, we perform a thorough manual filtering process: for each EMS activation, we retrieve Sentinel-1 images within a 4-day window starting from the reported event date. Among these images, we select the one that visibly matches the flood delineation mask and discard the others. Additionally, we manually verify the DEM rasters and filter out tuples with invalid elevation data to maintain data integrity. Regarding hydrography data, we acknowledge that its availability may be limited in certain areas due to missing data or the absence of water basins. However, we choose to retain all samples, even those without corresponding hydrography data, to support future studies and enable optional post-processing steps, such as subtracting permanent water areas from the model predictions if required. The final MMFlood dataset comprises 1,748 tuples, each containing SAR, DEM, and binary flood mask images. Among these tuples, 1,012 (57.9%) also include the corresponding hydrography images. Each tuple is accompanied by metadata specifying the EMS activation code, event date, and country of occurrence. Moreover, the SAR GeoTIFF is annotated with the acquisition date of the satellite signal, which may differ from the actual event date. On average, the SAR acquisition occurs 1 day and 21 hours after the reported event date. To accommodate the requirements of semantic segmentation tasks, we preserve the original image size for each area of interest while ensuring that the minimum input dimensions of 512 × 512 pixels are maintained across the entire dataset. The smallest image size in the dataset is 531 × 524 pixels, while the largest is 1,944 × 1,944 pixels. For benchmarking and training purposes, we split the dataset into three subsets: training, validation, and testing. To ensure a robust and unbiased split, we perform a random division based on the EMS activations, taking into account the geographical distribution of the events. This approach guarantees that each subset contains a proportionally equal number of examples from different locations worldwide, mitigating potential biases arising from specific land types or geographical areas. The resulting splits, visualized in Fig. 4.2, consist of 54 activations for the training set, 34 for the test set, and 7 for the validation set.

4.2.3 Method

The MMFlood dataset presents several challenges that must be addressed to ensure effective training and utilization of the data. These challenges arise from the inherent characteristics of the task at hand, such as the class imbalance, the nature of SAR data, and the inclusion of multiple modalities. The primary challenge remains the significant class disparity between flooded and non-flooded pixels, with flooded areas often constituting only a small portion of the segmentation mask. To tackle this issue, we explore different approaches, including filtering and downsampling strategies based on the flood-to-background ratio, and an Entropy-based Weighted Sampling (EWS) technique that assign higher weights to samples with more balanced class distributions. Another challenge stems from the inherent noise in SAR data. While typical preprocessing pipelines for SAR imagery involve radiometric correction and speckle noise reduction, we deliberately avoid these manual steps to maintain the dataset's reusability and expect deep learning algorithms to learn relevant features from the raw data. To address the noise, we employ data augmentation techniques and investigate the effectiveness of incorporating DEM data alongside the SAR imagery. The inclusion of multiple modalities, such as SAR data in VV and VH polarizations and elevation rasters from DEM, presents both opportunities and challenges in effectively combining and utilizing the complementary information they provide. We explore various approaches to merge the modality-specific features, including expanding input weights and employing early and late fusion strategies. In this section, we describe the methodology and training configurations employed to address these challenges and exploit the unique characteristics of the dataset constructed.

Problem Statement

In this work, we focus on the flood delineation task, considering only SAR and DEM data as training modalities, as the hydrography information can be utilized as a post-processing step to further refine the results. We formulate the flood delineation problem as a binary segmentation problem, where the objective is to classify each pixel in an image as either *flood* or *background*.

Let X be a set of samples, where each sample consists of a pair of images x_s and x_d, representing the SAR and DEM inputs, respectively. These images have matching and constant dimensions of $H \times W$ pixels. Additionally, let Y be a set of corresponding labels, where each label $y \in Y$ has the same dimensions as the input images. For each pixel i in an image, the label provides a binary annotation $y_i \in \{0, 1\}$, where 0 indicates a background pixel and 1 represents a flooded pixel.

The goal of binary segmentation training is to learn a model f_θ with parameters θ that maps from the image space to the label space, i.e., $f_\theta : X \to \mathbb{R}^{|H \times W|}$. In the multimodal setting, we employ a model architecture with two separate encoders, g_S and g_D, which extract features from the SAR and DEM inputs, respectively. These features are then fused through a shared decoder h to obtain the final model $f = h(g_S(x_S) \oplus g_D(x_D))$, where $\oplus$ denotes the feature fusion operation.

Handling Class Imbalance

To mitigate the imbalance problem, we investigate two complementary approaches based on downsampling and upsampling techniques. In the downsampling approach, we apply a threshold on the percentage of flood pixels in each image tile, effectively filtering out tiles that are almost entirely covered by background pixels. This allows us to focus on tiles that contain a higher proportion of informative flood pixels. Formally, given a list of preprocessed tiles, we compute a flood pixel ratio τ, obtained as:

$$\tau = \frac{\sum_i \mathbb{1}[y_i = 1]}{\sum_i \mathbb{1}[y_i]} \tag{4.1}$$

where y_i represents the label of the pixel i, and $\mathbb{1}[\cdot]$ is the indicator function. In other words, τ is the ratio of flood pixels to the total number of pixels in the label.

At training time, we simply select a cutoff value of τ and conduct the experiments on the subset of samples. This provides a baseline for the subsequent experiments.

In the upsampling approach, we leverage the concept of entropy from information theory to estimate the informative content of each label. Entropy can be interpreted as the amount of information contained in a sample, with higher entropy indicating more informative content [122]. We exploit this property to perform EWS, where each sample is weighted by a factor w_j based on its information content, computed as:

$$w_j = \lambda \left(- \sum_{y} p(y) \log_2 p(y) \right), \quad \forall j \in |X| \tag{4.2}$$

where $p(y)$ represents the distribution of pixels among the two classes (flood and background) in the label y, λ is a modulating factor between 0 and 1, and j is the index of each sample in the dataset X. By assigning higher weights to samples with higher entropy, we prioritize the selection of more informative tiles during training.

Integrating Multiple Modalities

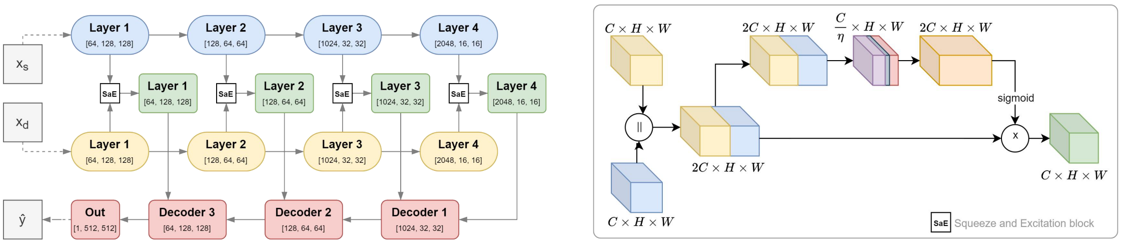


Figure 4.3: Multi-encoder architecture for integrating SAR and DEM modalities. The network consists of two separate encoders (blue and yellow) for each modality, which are merged layer-wise using Squeeze-and-Excitation (SaE) blocks to produce multimodal feature maps (green). The decoder (red) is agnostic to the multi-encoder setup.

Another challenge in the MMFlood dataset is the integration of multiple modalities, specifically SAR and DEM data, which provide complementary information for flood delineation. Given the matching spatial extents of both modalities, we explore two approaches to effectively combine these data sources: input channel expansion, with a single-encoder model (SE), and a multi-encoder (ME) approach. In the baseline SE approach, we simply concatenate the SAR bands and the DEM image along the channel dimension, creating a 3-channel input for the network. The network is trained from scratch, allowing it to learn features from each modality independently.

For the improved ME approach, we design a custom residual U-Net architecture [59] with two separate, lightweight encoders, one for each modality. Inspired by previous works, [228], we incorporate early and late fusion mechanisms using Squeeze-and-Excitation (SaE) blocks [98]. At each layer i of the encoders, we obtain feature maps with dimensions $H_i \times W_i \times C_i$. The SaE block merges these feature maps by first concatenating them along the channel dimension, resulting in a tensor of size $H_i \times W_i \times 2C_i$. This tensor is then calibrated using a weight map generated by the second branch of the SaE block. The weight map is obtained by squeezing the input tensor into a channel descriptor with $\frac{C_i}{\eta}$ channels and then expanding it back to $2C_i$ channels. The expanded tensor is passed through a sigmoid activation function to scale the values between 0 and 1, and then multiplied element-wise with the concatenated feature maps. Finally, the calibrated feature maps are reduced to the original dimensions $H_i \times W_i \times C_i$ using a convolutional layer, ensuring compatibility with the decoder architecture. This multi-encoder approach enables the network to learn modality-specific features while leveraging the different information content provided by SAR and DEM data. Fig. 4.3 illustrates the multi-encoder architecture, with the SaE block depicted on the right and the overall network structure on the left.

4.2.4 Experiments

In this section, we present a comprehensive analysis of the experiments conducted on the MMFlood dataset and discuss their outcomes. We begin by outlining the preprocessing pipeline designed to transform the raw images into suitable input data for our models. This is followed by a series of benchmark tests aimed at identifying the most effective encoder-decoder combination to serve as the foundation for further experimentation. Subsequently, we test the introduction of additional components described in Section 4.2.3, namely EWS, the integration of DEM as an extra input channel, and the multi-encoder (ME) architecture compared against the single-encoder architecture (SE). We examine the impact of these enhancements on the model's performance, assessing their strengths and limitations in the context of flood delineation.

Implementation Details

In order to prepare the MMFlood dataset for training, we first apply a series of preprocessing steps to the raw images. For the SAR data, we take the base-10 logarithm of the acquisitions, converting the values to decibels and reducing the impact of high backscatter peaks. As for the DEM data, which represents the raw altitude in meters, we limit the range to $[-100, 6000]$ to minimize errors and keep values in a reasonable range. Furthermore, we normalize each input by calculating the mean and standard deviation for each channel, ensuring consistency across all experiments.

The MMFlood dataset contains images of varying sizes and scales, which poses a challenge for training. To address this issue and maintain a standardized setup, we adopt an offline tiling approach [49]. We divide each image and modality into 512×512 tiles, extracting the minimum number of tiles needed to fully cover the original image. This tile size strikes a balance between preserving pixel-wise class balance and capturing sufficient contextual information. In situations where the image dimensions are not evenly divisible by the tile size, we use dynamic overlapping, extending the tiles by the minimum number of pixels required to cover the image entirely without padding. This approach maximizes the visual content within each tile. After preprocessing, the training set consists of 6,182 tiles, with an additional 560 tiles reserved for validation, excluding the test partition. To maintain the integrity of the pixel count, we apply the tiling procedure only to the training and validation sets, preserving the full-sized images for testing. During the testing phase, we employ online tiling with a fixed overlap and merge the predictions to create a single output with dimensions matching those of the input images. In addition to preprocessing, we utilize various online data augmentation techniques to reduce overfitting during training. For experiments involving multiple modalities, we apply a shared set of transformations to each input, including random rotation, cropping, horizontal and vertical flipping, each with a probability of $p = 0.5$. We also include grid and elastic deformations, commonly used in medical image analysis [34], to generate diverse flood patterns and simulate terrain variations, possibly enhancing the model's ability to generalize across different geographical contexts. To tackle the inherent speckle noise in SAR images, we apply further pixel-level transformations, such as random Gaussian blur and multiplicative noise.

Throughout our experiments, we train the models for 100 epochs, employing an early stopping criterion with patience of 30 epochs. Given the unique characteristics of the MMFlood dataset, we do not use any pre-trained weights. We use a batch size of 16 for most experiments, with the exception of UNet combined with ResNet50 or DenseNet121, where we reduce the batch size to 12 to manage memory constraints. It is worth noting that due to the incompatible stride sizes between the outputs of DenseNet and the DeepLabV3+ decoder, we exclude this particular combination from our baseline results. For optimization, we use the AdamW optimizer with an initial learning rate of $\lambda = 1 \times 10^{-3}$ and a weight decay coefficient of 0.01. We also employ a polynomial learning rate scheduler with $\gamma = 3$, decreasing the learning rate to $\lambda = 10^{-4}$ by the end of training.

Considering the imbalanced nature of the flood delineation task, we select the focal Tversky loss [2] as our loss function for all experiments. This loss extends the soft Dice score by incorporating a focal component, making it particularly suitable for handling imbalanced problems. Formally, this objective function can be defined as:

$$\mathcal{L}(p, y) = (1 - \mathcal{I}_T(p, y))^{\gamma} \tag{4.3}$$

where $\mathcal{I}_T$ refers to the Tversky Index, in turn computed as:

$$\mathcal{I}_T(p, y) = \frac{\sum_{i=1}^{N} p_i y_i}{\sum_{i=1}^{N} p_i y_i + \alpha \sum_{i=1}^{N} (1 - p_i) y_i + \beta \sum_{i=1}^{N} p_i (1 - y_i)} \tag{4.4}$$

Here, N is the total number of pixels, p_i and y_i are the predicted and ground truth labels for a pixel i, and α and β are hyperparameters that control the trade-off between false positives and false negatives. γ is a focal parameter that adjusts the weight of easy and hard examples. We set the hyperparameters of the loss to $\alpha = 0.6$, $\beta = 0.4$, and $\gamma = 2$. To thoroughly evaluate the performance of our models, we consider Precision, Recall, IoU, and F1 Score as metrics. We calculate the mean value of these metrics across all test images, giving equal weight to each tile. All experiments and procedures outlined in the next section were carried out on a workstation featuring an Intel Xeon Silver 4216 CPU and four Nvidia GTX 2080Ti GPUs. The implementation was developed using Python, leveraging the PyTorch library for model training and testing.

Results

Encoder	**Decoder**	**Precision**	**Recall**	**IoU**	**F1**
Otsu		0.2895	0.4627	0.1963	0.2895
ResNet50	UNet	0.6910	0.8710	0.6269	0.7706
ResNet50	DLV3+	0.6733	0.9031	0.6279	**0.7714**
ResNet50	PSPNet	0.6659	0.8858	0.6132	0.7603
TResNet	UNet	0.6151	0.9178	0.5830	0.7366
TResNet	PSPNet	0.7376	0.7462	0.5897	0.7419
TResNet	DLV3+	0.6331	0.6299	0.4614	0.6315
EfficientNet	UNet	0.6856	0.6242	0.4853	0.6534
EfficientNet	DLV3+	0.4491	0.2050	0.1638	0.2815
EfficientNet	PSPNet	0.7143	0.6211	0.4976	0.6645
DenseNet121	PSPNet	0.6050	0.8967	0.5656	0.7225
DenseNet121	UNet	0.5954	0.9054	0.5605	0.7184

Table 4.3: Comparative analysis of various encoder-decoder combinations for flood delineation on the MMFlood test set. The table presents the performance metrics (precision, recall, IoU, and F1 score) for each model, with the Otsu method serving as a baseline. The ResNet50 encoder paired with UNet and DeepLabV3+ (DLV3+) decoders achieves the best overall performance.

Model	τ	**DEM**	**EWS**	**Precision**	**Recall**	**IoU**	**F1**
SE	2%			0.6733	**0.9031**	0.6279	0.7714
SE	2%	✓		0.6814	0.8979	0.6324	0.7748
SE	×		✓	0.6976	0.8955	0.6451	0.7843
SE	×	✓	✓	0.7173	0.8893	0.6585	0.7941
ME	×	✓	✓	**0.7319**	0.8794	**0.6652**	**0.7989**

Table 4.4: Incremental performance improvements achieved by incorporating DEM data, entropy-weighted sampling (EWS), and a multi-encoder (ME) architecture, compared to the single-encoder (SE) baseline. The pixel ratio threshold (τ) variation is also reported here to compare its impact on the model's performance.

τ	**EWS**	**Precision**	**Recall**	**IoU**	**F1**
0%	×	0.4118	0.2274	0.1717	0.2930
2%	×	0.6733	0.9031	0.6279	0.7714
5%	×	0.6301	0.9097	0.5930	0.7445
0%	✓	0.6976	0.8955	**0.6451**	**0.7843**
2%	✓	0.6700	0.9024	0.6247	0.7690
5%	✓	0.6486	0.9096	0.6094	0.7573

Table 4.5: Hyperparameter study examining the impact of the flood pixel ratio threshold (τ) and the effectiveness of EWS on the model's performance. The experiments reported here were executed on the single-encoder (SE) variant.

Our initial experiments focus on establishing a set of baseline results by evaluating the performance of different combinations of encoders and decoders on the MMFlood dataset. Table 4.3 presents the average Precision, Recall, IoU, and F1 scores achieved by each model variant on the test set, considering only the SAR imagery as input and applying a flood threshold $\tau = 2\%$ to improve the training stability (see Section 4.2.4).

The results indicate that the choice of encoder has a more significant impact on the model's performance than the decoder. ResNet variants, namely ResNet50 and TResNet, consistently outperform other encoder architectures, achieving mean IoU scores of 0.62 and 0.59, respectively. In contrast, EfficientNet and DenseNet variants yield suboptimal results, with a maximum IoU of 0.56 for DenseNet121, despite their increased complexity.

Among the decoders, UNet and DeepLabV3+ demonstrate comparable performance, attaining an average IoU of 0.77. Although PSPNet slightly lags behind with a 1% performance gap, it exhibits the most stable results across all experiments. Compared to the Otsu thresholding baseline, deep learning approaches

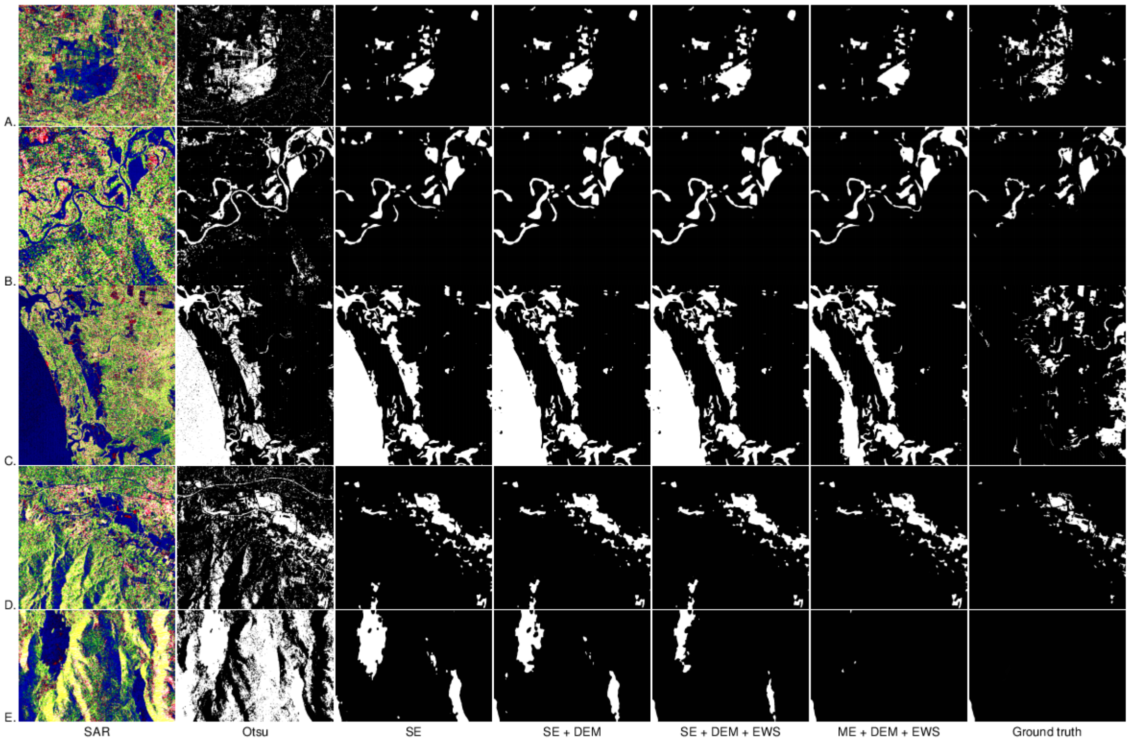


Figure 4.4: Qualitative results on the MMFlood test set. From left to right: SAR input, Otsu thresholding baseline, single-encoder model with SAR only (SE), single-encoder model with SAR and DEM (SE+DEM), single-encoder model with SAR, DEM, and entropy-weighted sampling (SE+DEM+EWS), multi-encoder model with SAR, DEM, and entropy-weighted sampling (ME+DEM+EWS), and the ground truth mask.

prove to be significantly more robust, achieving an improvement of +0.3 in most metrics. This superiority is further emphasized by the qualitative results presented in Fig. 4.4, where the Otsu baseline produces noticeably noisier outputs than its neural network counterparts. WE note that careful preprocessing could potentially mitigate this issue, however deep learning models can effectively handle noisy inputs without additional manual steps, and the purpose of deep learning is in fact to obtain an end-to-end solution, without further manual intervention.

To investigate the influence of incorporating additional modalities and addressing class imbalance, we conduct a series of experiments with different combinations of DEM data and class balancing techniques, as shown in Table 4.4. The inclusion of DEM data as an extra input channel in the SE variant (i.e., without any further processing or considerations) leads to a performance improvement of 1.5% over the baseline model, achieving an IoU of 0.63 while maintaining the flood threshold at $\tau = 0.02$. This result highlights the significance of terrain information in the context of flood segmentation, where elevation and the nominal extents of water basins, often visible and well-defined in the DEM, play an important role. The performance gain is further amplified by the introduction of EWS, which guides the sampling of the most informative labels during training. With the combination of DEM and EWS, the model attains an IoU of 0.65, demonstrating the effectiveness of this approach in addressing class imbalance. The impact of EWS is particularly evident in terms of precision, which increases from 0.67 to 0.71 when compared to the baseline model. Finally, we evaluate the ME architecture, which achieves the highest IoU score at 0.66. While numerically comparable to the SE variant, the qualitative results in Fig. 4.4 reveal that the ME setup exhibits greater resilience to terrain changes, where shadows on mountain ranges or large permanent water basins could be erroneously classified as floods by other model configurations. This observation suggests that the careful fusion of SAR and DEM data during training, despite the limited loss in recall, allows for more precise output with an effective utilization of the visual information provided by the DEM modality.

Hyperparameter Study. To gain further insights into the effectiveness of different sampling procedures, we conduct a hyperparameter study on the influence of an efficient sampling at training time, as presented in Table 4.5. We evaluate four downsampling thresholds ($\tau \in \{0\%, 2\%, 5\%\}$) to assess the impact of reducing the training set size on the model's performance. Additionally, we examine the influence of EWS on each variant, effectively combining oversampling of informative inputs with downsampling of tiles containing mostly background pixels. In this context, the results highlight that, without any kind of oversampling mechanism (i.e., with EWS disabled), the optimal threshold for this task is $\tau = 0.02$, as lower values hinder the model's performance, while higher values introduce a bias towards flooded areas, leading to less precise predictions. In fact, the precision start lowering above the $\tau = 0.02$ threshold, while the gain in terms of recall becomes negligible.

Considering now the last three rows in Table 4.5, the inclusion of EWS yields a substantial performance boost in the absence of downsampling, improving the baseline IoU by 4.5% to reach 0.64. However, the benefits of EWS are diminished in the threshold-based approaches, likely due to the reduced number of training samples.

In summary, these results suggest that EWS is most effective when applied to the entire dataset, as it can fully exploit the available data by providing a more diverse set of samples in a more guided way, while a threshold can still be effective when no particular sampling technique is applied.

Discussion. Our experiments on the MMFlood dataset demonstrate the effectiveness of deep learning techniques for flood delineation, particularly when combined with additional modalities such as DEM data and class balancing strategies like EWS. The multi-encoder architecture emerges as the most promising approach, leveraging the complementary information provided by SAR and DEM data to achieve robust and accurate flood segmentation. However, we acknowledge certain limitations that could be addressed in future works. First, MMFlood primarily focuses on SAR imagery, which, although advantageous in terms of cloud penetration and day-and-night operation, may not always provide the highest resolution or the most comprehensive view of the flood situation. Additional optical or radar sources could greatly improve the mapping quality. Second, while the inclusion of DEM and hydrography data proved beneficial, the static nature of these modalities may not fully capture the dynamic changes in terrain and water levels during flood events, therefore limiting the performance improvements. Introducing ground sensor data, time-varying elevation information (e.g., LiDAR or InSAR information), or simply a time series of acquisitions, may better represent the evolving scene during the event. Last, the current benchmark evaluation focuses on a relatively narrow set of deep learning architectures and training strategies: future research could explore a wider range of models, including state-of-the-art Vision Transformers. Moreover, the investigation of advanced training techniques, such as self-supervised learning, could help in improving the robustness and generalization capabilities of the models on these specific downstream tasks, in presence of limited labeled data.

4.3 Scale Unbalance in Aerial Images

Another main issue affecting the performance of semantic segmentation models in remote sensing is the scale imbalance of the objects present in the scene. In particular, especially when considering very high resolution (VHR) acquisitions, the ratio between object sizes and the level of detail can vary by several orders of magnitude. This phenomenon can be observed in many real-world scenarios, where objects of interest may appear at different scales within the same image. For instance, in medical imaging, entities to be delineated can have vastly different sizes depending on the target. This is true also in remote sensing, where buildings and roads can span hundreds of pixels, while smaller objects such as cars or trees may only occupy a few dozen pixels [242]. Scale imbalance poses a significant challenge for semantic segmentation models, as they need to accurately classify objects at all scales while maintaining a high level of spatial resolution. To address this issue, several approaches have been proposed in the literature, from multiscale architectures, such as PSPNet [270] or DeepLab [39], to a combination of local and global processing, in order to gather information about details and the overall context. This can be achieved through a two-stage pipeline, where the first stage identifies candidate regions containing objects of interest, and the second stage refines the segmentation within those regions [42].

In this section, we tackle the scale imbalance problem in the specific case of PV panel segmentation, where this issue is particularly relevant due to the wide range of installation sizes encountered. Large industrial plants can cover entire fields, while smaller agricultural or domestic installations may only partially cover roofs. This variability in scale can make it difficult for a single model to accurately segment all types of panels, leading to poor performance on either large or small plants. To mitigate this issue, we propose a multiscale training approach, based on representation consistency between global and local features [42]. Specifically, we process each input tile at two different resolutions: a local scale, where the image is split into smaller patches, and a global scale, where the entire tile is downsampled to a lower resolution. The model is then trained to produce consistent predictions across both scales, encouraging it to learn features that are robust to changes in object size. This is achieved through a consistency loss that penalizes differences between the local and global predictions, acting as a regularization term during training. We evaluate the effectiveness of our approach on a custom dataset of VHR aerial imagery, specifically created for the task of PV panel segmentation. The dataset covers a large area in Italy, specifically in the Piedmont region, and includes manual annotations for over 9,000 PV panels, ranging from large industrial installations to small domestic ones. Our experiments show that the proposed multiscale training strategy significantly improves the performance of semantic segmentation models, particularly in terms of their ability to handle the scale imbalance present in the data. Moreover, we demonstrate that incorporating additional information from the infrared band can further boost the accuracy of the segmentation, especially for challenging object categories.

In this section, inspired by previous works such as GLNet [42], and regularization such as Augmentation Invariance, presented in Chapter 3, we propose a multiscale regularization approach to improve the consistency between local and contextual features. Second, we demonstrate the effectiveness of this technique on a custom aerial dataset, purposely constructed for the PV panels delineation task. Last, we also explore the use of domain-specific priors to further refine the output and produce more meaningful results for downstream analysis. We release the constructed benchmark dataset and the code to reproduce the results at `https://github.com/links-ads/access-solar-panels`.

4.3.1 Related Works

Multiscale feature extraction. In recent years, deep learning-based approaches have achieved remarkable success in semantic segmentation tasks, especially considering Convolutional Neural Networks (CNNs) that learn hierarchical features from the input image and produce dense pixel-wise predictions. However, one of the main challenges in aerial segmentation is capturing contextual information at multiple scales, which is crucial for accurately classifying objects of different sizes and resolving ambiguities in local regions.

To address this issue, various multiscale feature aggregation techniques have been proposed in the literature. One popular approach is to use an encoder-decoder architectures, where the encoder progressively downsamples the input image to extract high-level features, while the decoder gradually upsamples the feature maps and combines them with lower-level features to recover spatial details. The U-Net [197] is the reference example of this design, which has been widely adopted and extended in many subsequent works [77, 59, 72]. Another common strategy is to employ pyramid pooling modules that capture context information at multiple scales. The Pyramid Scene Parsing Network (PSPNet) [270] introduces a pyramid pooling module that aggregates features from different regions of the feature map using adaptive average pooling, followed by upsampling and concatenation. This allows the network to incorporate global context information and improve its ability to handle objects of various sizes. Similarly, the DeepLab family of models [39, 40] utilize atrous spatial pyramid pooling (ASPP) to capture multiscale context by applying convolutional filters with different dilation rates.

Attention mechanisms have also been explored as a means of selectively focusing on relevant features and suppressing irrelevant ones, both in channel and space dimensions. The Dual Attention Network (DANet) [77] introduces a double attention module that captures dependencies along both spatial and channel dimensions, allowing the network to adaptively integrate local features with their global dependencies. The Criss-Cross Network (CCNet) [100] proposes a criss-cross attention module that computes the attention weights between each pixel and its surrounding pixels in a criss-cross path, enabling the network to capture long-range dependencies more efficiently.

More recently, transformer-based architectures have gained popularity in semantic segmentation due to their ability to model long-range dependencies and capture global context. The Vision Transformer (ViT) [61] adapts the self-attention mechanism from natural language processing to computer vision tasks, treating an image as a sequence of patches and learning relationships between them. Subsequent works such as the Swin Transformer [131] and Segmenter [215] or SegFormer [255] have further improved upon this design by introducing hierarchical architectures and integrating convolutional layers. Similarly, CNNs have also benefited from the design improvements, with efficient and effective architectures such as ConvNext [133].

While these techniques have shown impressive results on standard benchmarks on natural images, they often yield suboptimal performances when applied in scenarios where objects exhibit significant scale variations. This is particularly evident in applications such as remote sensing, where the resolution of satellite and aerial imagery can vary greatly depending on the sensor and acquisition parameters. In the specific use case of photovoltaic (PV) panel segmentation from aerial imagery, the scale imbalance between large industrial installations and small domestic panels poses a major challenge for existing methods. To tackle this issue, some works have proposed multiscale training strategies that process the input image at different resolutions and encourage the model to learn scale-invariant features. For example, GLNet [42] introduces a collaborative global-local network that consists of a global branch that captures context information from a downsampled version of the input and a local branch that processes the full-resolution image in a patch-wise manner. The two branches are jointly optimized using a consistency loss that enforces similar predictions across scales.

Dataset	**Images**	**Annotations**	**Area (km^2)**	**Resolution (cm)**
DOTA [252]	2,806	188,282	5,261	15-150
California [23]	601,095	19,863	1,698	30
ISPRS [199]	23	4,488	3.2	7.5-15
Ours	**105**	**9,462**	**2,940**	**30**

Table 4.6: Comparison of our dataset with a non-exhaustive list of similar aerial image datasets for object detection and segmentation. While our dataset is focused on PV panel detection, it is comparable in scale to other large-scale datasets like the California PV panels dataset and DOTA, which is instead designed for general object detection tasks.

Delineating PV panels. In the context of PV panel segmentation, deep learning techniques have shown great promise in this regard, especially thanks to CNNs due to their resource efficiency and ability to learn hierarchical features directly from the input imagery. Several early works [28] propose CNN-based methods for segmenting PV panels from aerial imagery, demonstrating the effectiveness of deep learning for this task. The DeepSolar framework [259] presents a comprehensive approach for mapping PV installations across the United States using a combination of satellite imagery and deep learning, resulting in a database covering over 1 million PV installations. Subsequent works have explored various strategies for improving the performance and efficiency of deep learning-based detection methods, such as incorporating additional information from other data sources [81] or employing transfer learning techniques to leverage pre-trained models and reduce the amount of labeled data required for training [107].

Given the increasing application of deep learning techniques on downstream tasks, there has been growing interest in the development of large-scale datasets and benchmarks for PV panel detection. These datasets cover a wide range of geographical areas and vary in terms of their size and resolution. However, the most extensive datasets, both in terms of the number of annotations and the total surface area covered, primarily focus on the United States. Notable examples include the California dataset [23] and the DeepSolar framework [259], which provide large-scale annotations for PV installations across multiple cities and states. While these datasets provide undoubted valuable resources, their applicability to other geographical regions may be limited due to differences in visual characteristics, such as landscapes, vegetation types, and building architectures. These variations can significantly impact the performance of automatic detection systems trained on data from a specific region. When considering the European continent, the availability of large-scale PV panel datasets is relatively limited compared to the United States. Existing studies have primarily focused on smaller geographic areas, such as the Netherlands [108, 183] or Switzerland [33], and have relied on a combination of satellite imagery and aerial photography. The scarcity of extensive datasets for Europe can be attributed to the higher costs associated with acquiring very high resolution (VHR) aerial and remote sensing imagery, which is essential for accurate PV panel detection and delineation. In contrast, open data sources like Sentinel satellite imagery, while freely available, may not provide sufficient spatial resolution for this task. In this work, we attempt to bridge this gap by creating a high-quality PV panel dataset specifically focused on the Piedmont region of Italy. As shown in Table 4.6, our dataset is comparable in scale to existing resources such as the California dataset [23], providing over a hundred VHR images across two large provinces in the region, along with thousands of manually annotated PV panels.

4.3.2 Dataset

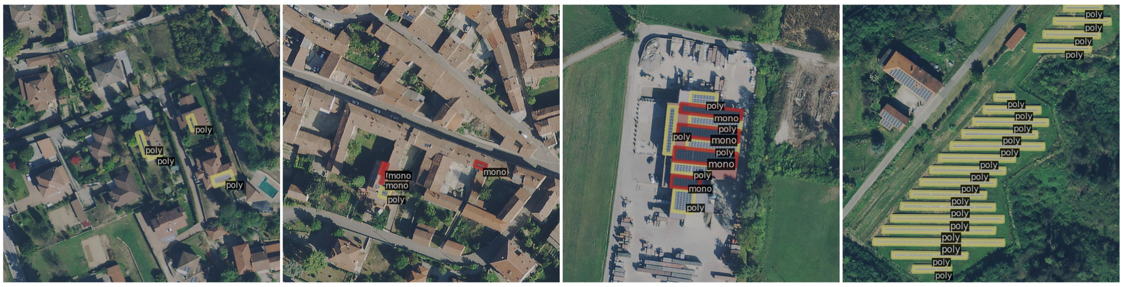


Figure 4.5: Examples tiles extracted from the dataset, together with their annotations. Starting from the left, the first two images show the smaller domestic installations, while the last two illustrate larger industrial plants.

We construct our dataset focusing on the Piedmont region, in Italy. The study area for our dataset comprises the provinces of Asti and Alessandria in Piedmont, which have the highest concentration of PV installations in the region. Piedmont itself ranks fourth among Italian regions in terms of the total number of PV panels and first in terms of energy production from photovoltaic sources as of 2022 [62]. By focusing on this area, we ensure that our dataset is representative of a wide range of panel types and installation scenarios, including both urban and rural settings, as well as large-scale industrial plants and smaller agricultural and residential installations. The aerial images used in our dataset were sourced from the Terraitaly

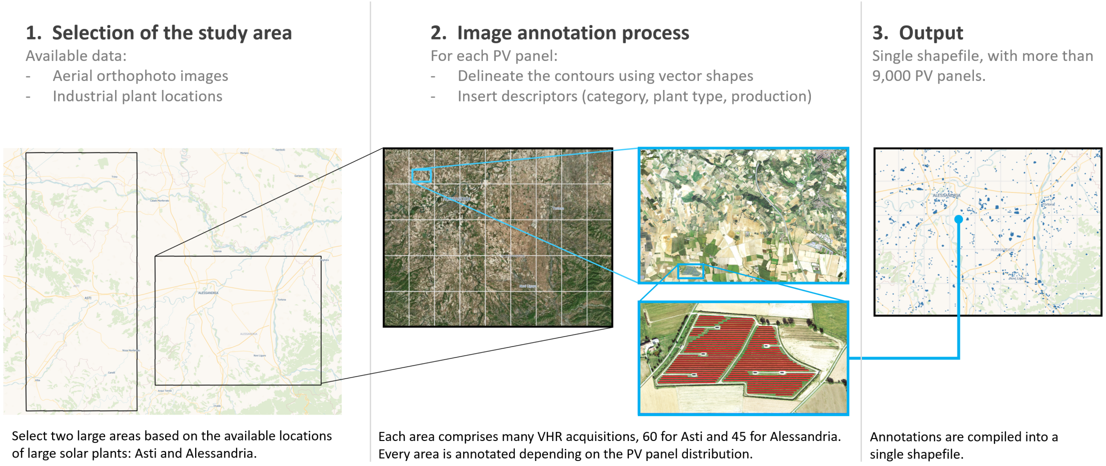

1. Selection of the study area
Available data:
- Aerial orthophoto images
- Industrial plant locations
2. Image annotation process
For each PV panel:
- Delineate the contours using vector shapes
- Insert descriptors (category, plant type, production)
3. Output
Single shapefile, with more than 9,000 PV panels.
Select two large areas based on the available locations of large solar plants: Asti and Alessandria.
Each area comprises many VHR acquisitions, 60 for Asti and 45 for Alessandria. Every area is annotated depending on the PV panel distribution.
Annotations are compiled into a single shapefile.

Figure 4.6: Schematic describing the annotation pipeline, from the selection of the areas, to the acquisition of the VHR imagery, to the manual annotations, stored in a single vector file.

catalog provided by the Compagnia Generale di Riprese aeree (CGR) s.p.a, an Italian company specializing in high-resolution aerial imagery [1]. We selected the most recent orthorectified photo archive from 2018, which offers a spatial resolution of 30cm/pixel, consistent with the resolution of other aerial datasets such as the California dataset [23]. This level of detail is sufficient for accurately detecting and delineating individual PV panels, including smaller domestic installations, while also providing a manageable data volume for processing and analysis. In total, our dataset includes 105 very high resolution (VHR) images in RGB-NIR format (i.e., visible spectrum and Near-infrared (NIR)), with 60 images covering the Asti province and 45 images covering the Alessandria province. Each acquisition covers an average area of 18 km^2, with dimensions of approximately 20,000 × 16,000 pixels. The images are provided as georeferenced TIFF files, using the Universal Transverse Mercator (UTM) zone 32N Coordinate Reference System (CRS), corresponding to the EPSG code 32632, to minimize distortion from the map projection. The annotation process, displayed in Fig. 4.6, involved not only identifying and outlining the boundaries of PV panels but also assigning metadata information to each annotated instance. This task was performed by a team of domain experts, who exhaustively labeled all registered large-scale industrial plants using their known addresses or approximate coordinates, while agricultural and domestic installations were annotated by manual lookup with a uniform spatial distribution to avoid geographical bias. In this last case, the presence of every possible annotation is not guaranteed on every tile. In addition to the geometric outline of each panel, the annotations include

[1] https://www.cgrspa.com/

several attributes, such as a unique identifier for the panel and the associated plant (in the case of industrial or agricultural installations), the main area they reside into (either *Asti* or *Alessandria*), orientation, estimated power output, installation type (*industrial*, *agricultural*, or *domestic*), and PV technology (*monocrystalline* or *polycrystalline*). The inclusion of PV technology type is particularly relevant for estimating the energy production of a plant, as monocrystalline panels offer higher efficiency and durability compared to polycrystalline panels, despite their higher cost [154]. The final dataset contains a total of 9,462 manually annotated PV panels, with 8,967 polycrystalline and 495 monocrystalline installations. The annotations were compiled into a single shapefile to facilitate integration with the aerial imagery and other geospatial data sources. Fig. 4.5 presents a sample of the dataset, illustrating the diversity of installation types and settings covered, from large industrial plants to small domestic systems.

4.3.3 Method

From a machine learning point of view, the problem of delineating PV panels represents a standard semantic segmentation task, where each pixel is classified as either monocrystalline, polycristalline, or background. The challenge arises from the wide range of scales at which PV installations can be found, from large industrial plants, to domestic systems covering a few square meters. To address this issue, we propose a multiscale framework that leverages both global and local context to accurately segment PV panels at different scales while maintaining a manageable computational complexity. Our approach incorporates a multiscale regularization technique that encourages consistency between the predictions at different resolutions, ensuring that the segmentation is coherent across scales. This is achieved by introducing a loss term that penalizes discrepancies between the downsampled local predictions and the global predictions. To make the segmentation output more useful for practical applications, we also introduce a polygonization post-processing step. This involves converting the pixel-wise segmentation map into a set of vectorized polygons representing individual PV panels. The polygonization process consists of instance identification, minimum bounding rectangle fitting, and boundary refinement stages, providing a compact and meaningful representation of the detected PV installations.

Multi-Scale Regularization

To address this challenge of scale variability, we draw inspiration from GLNet [42] and introduce a multiscale training approach with a consistency regularization across scales. While the prohibitive size of our aerial images precludes the direct application of a complete global-to-local or local-to-global regularization [42], we

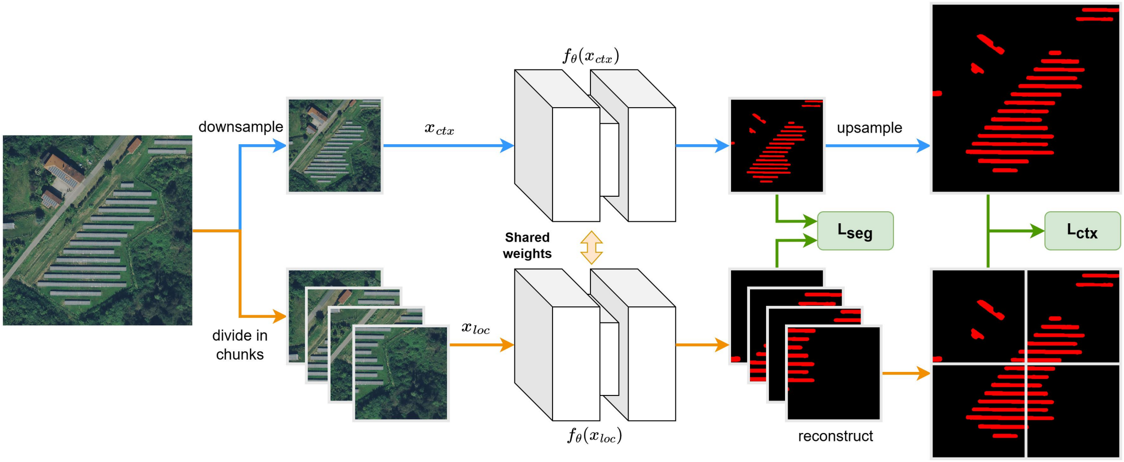


Figure 4.7: Semantic segmentation framework with multiscale regularization. The input is processed at global and local scales, and the extracted features are fused to produce pixel-wise predictions. A consistency loss term regularizes the multiscale learning process.

propose a context loss that compares larger image portions with smaller crops extracted from the same region to carry out a similar task. We recursively subdivide each training tile into quadrants by splitting vertically and horizontally for a specified number of times S. This results in a hierarchy of 4^S local tiles at the finest level. We process these local tiles independently to extract features at different scales. To introduce contextual information, we define a downscaled version of the input image, which we refer to as the contextual image x_c. We extract contextual features from this downscaled image using a shared encoder. The contextual features are then compared with the local features extracted from the patches to enforce consistency. Formally, let $\phi_\theta(x_l)$ be the reconstructed local features, obtained by providing the model each i-th local full-resolution patch x_l^i obtained from the input tile, and let x_c be the downscaled contextual image. We define the multiscale context loss as:

$$\mathcal{L}_{ctx} = |\phi_\theta(x_l) - \phi_\theta(x_c)|_2^2 \tag{4.5}$$

where ϕ_θ represents the model output before the final Softmax activation. This loss penalizes the discrepancy between local and contextual features, encouraging the model to learn scale-invariant representations. During training, the multiscale context loss is added to the standard semantic segmentation loss as a regularization term. By enforcing consistency between local and contextual features, we aim to improve the model's robustness to scale variations commonly present in real-world applications. A visual representation of this framework is shown in Fig. 4.7

Polygonization

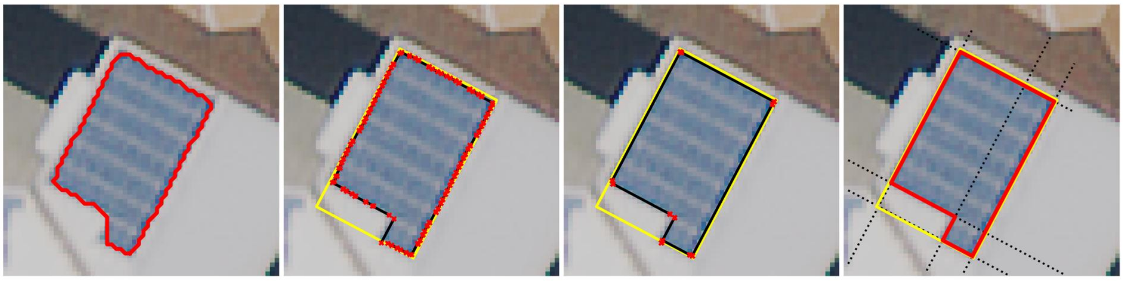


Figure 4.8: Key steps of our polygonization algorithm for regularizing the segmentation output of photovoltaic panels, from left to right: (a) conversion of the raw mask into a coarse polygon, (b) alignment of edges with the dominant orientations of the Minimum Bounding Rectangle, (c) merging of close parallel edges, and (d) computation of the final regularized polygon. Our approach leverages the typical rectangular shape of PV panels to generate clean, rectilinear polygons that more accurately delineate the panels while removing noise and artifacts.

In order to provide a more practical and usable output for further analysis and downstream tasks, the raw predictions generated by semantic segmentation models often require additional processing. Two key aspects need to be addressed: first, the raster-based segmentation output should be converted into a vectorized polygon representation, ensuring that the generated shapes maintain a regular structure without noisy or overlapping boundaries. Second, in the case of semantic segmentation, an additional step is needed to transform the pixel-wise classifications into a more meaningful instance-level representation. We tackle the latter challenge by applying Connected Components Labeling (CCL) [94] to extract individual PV panel instances from the semantic segmentation output. CCL is an algorithm that assigns distinct numerical labels to each connected component in a binary image. To prevent the panel category from interfering with the instance extraction process, we first apply CCL on a binarized version of the segmentation output. We then iterate over each connected component, discarding instances with a surface area below a minimum pixel threshold t and assigning a single class label to each panel based on majority voting over the pixels within the component. Once we have obtained an instance-level representation of the PV panels, we focus on converting the raster output into a more compact and regularized polygon format. Our polygonization procedure draws inspiration from previous work [247], adapting it to the specific characteristics of PV panels. The visual representation of each step is visible in Fig. 4.8. We start by assuming that PV panels have a rectangular shape or, in the case of more complex installations, that they are composed of rectangular subcomponents. The goal of our algorithm is to regularize the edges of each polygon such that the internal angles are always 90 degrees, while minimizing deviations from the original shape. The first step in our polygonization pipeline is to convert the raster instance mask into an initial polygon using the Douglas-Peucker algorithm [202], with a low tolerance value to avoid oversimplification. From this coarse polygon, we extract the oriented Minimum Bounding Rectangle (MBR), defined as the rectangle with the smallest area that fully encloses the original shape and is not constrained to be axis-aligned. Given the typically rectangular structure of PV panels, we consider the MBR to be the best approximation of the panel's orientation. We leverage this information to further regularize the edges of the coarse polygon. Each edge is aligned with the direction of the MBR edge that is closest to its current orientation. This step ensures that all edges are either parallel or perpendicular to each other, while preserving the dominant orientations of the original shape. However, rotating the edges independently introduces discontinuities along the polygon's perimeter. To address this, we perform a simplification step before reconnecting the edges. We remove consecutive edges that are parallel (i.e., lie on the same side of the polygon) if the length of the edge connecting their endpoints is below a threshold T. This threshold is defined as $T = \alpha \cdot L$, where α is a scalar factor between 0 and 1, and L is the length of the polygon's longest edge. Essentially, this step discards edges that contribute the least to defining the object's boundary. Finally, we reconstruct the regularized polygon by computing the intersection points between the lines defined by the remaining edges. The complete polygonization procedure is summarized in Appendix A.

4.3.4 Experiments

In this section, we first provide a detailed description of our experimental setup, including the dataset preparation, model configurations, and training procedures. We then discuss the quantitative results obtained by our method, comparing its performance against baseline models and highlighting the impact of our key contributions, such as the multiscale context loss and the polygonization algorithm.

Implementation Details

To assess the performance of our semantic segmentation approach, we employ the custom PV panel dataset introduced in Section 4.3.2. The dataset provides accurate pixel-wise annotations for over 9,000 PV panels, and we focus on a multi-class segmentation problem. distinguishing between *monocrystalline* or *polycrystalline* panels. The original images, each measuring approximately 20,000 × 16,000 pixels, are divided into smaller tiles of 512 × 512 pixels with an overlap of 256 pixels to ensure that all relevant context is preserved, and border pixels eventually become center pixels in the subsequent block. Due to the sparsity of the problem at hand, this process results in a total of 696 tiles containing PV panels, which we further split into subsets for training (404 tiles), validation (123 tiles), and testing (169 tiles). For a more robust evaluation and testing procedure, first we split training and validation based on the solar plant, so that different installations belong to different splits, then we further select all the tiles belonging to the Asti province as test set. As an attempt to introduce additional variation, we also train on an extended version of the dataset, obtained by applying a copy-paste augmentation strategy [83]. We randomly sample PV panels from the training set and paste them onto background tiles, ensuring that the pasted panels do not overlap with existing ones. This process allows us to generate an additional 10,000 training tiles, effectively increasing the dataset size and diversity. We refer to this augmented dataset as the *extended* dataset in our experiments. Additionally, we perform the experiments on both RGB-only and RGB-NIR inputs, to assess the performance improvements introducing the extra band.

Given the focus on lightweight and deployable solutions, we experiment with U-Net based models [197]. Specifically, we adopt the residual variant (ResUNet) [59] as our decoder, which incorporates residual skip connections to aid convergence. For the encoder, we experiment with standard ResNet backbones [92] as well as the recently proposed ConvNext [133], which introduces architectural modifications to narrow the gap with transformer-based models while preserving high throughput. Because of the aerial setting, we adopt weights pretrained on DOTA [252] for the encoders. Our semantic segmentation models are implemented using the PyTorch deep learning framework [178]. All models are trained using the AdamW optimizer with an initial learning rate of 2.5×10^{-3}, a momentum of 0.9, and a weight decay of 1×10^{-4}. We employ a batch size of 4 and train the models for 80 epochs. During training, we apply a combination of photometric and geometric data augmentations, including random flipping, rotation, scaling, brightness and contrast adjustments, and Gaussian noise and blur, to improve the models' robustness and generalization capacity.

Results

To evaluate the performance of our semantic segmentation models, we employ the Intersection over Union (IoU) metric. Considering the class imbalance present in the dataset, we report both the macro-averaged IoU ($MIoU$) as in Eq. (2.3) and micro-averaged IoU ($mIoU$) as in Eq. (2.4) to provide a comprehensive overview of the model's performance. The results, summarized in Table 4.7, highlight the challenges associated with detecting monocrystalline PV panels. When training on the *base* dataset, which only includes the original annotated tiles, we observe that the semantic segmentation approach already surpasses the delineation capabilities of the instance-based model in two out of three categories. The available input information is sufficient to distinguish between background and polycrystalline pixels, which constitute the majority of installations. However, the model struggles or completely fails to delineate monocrystalline plants. The inclusion of

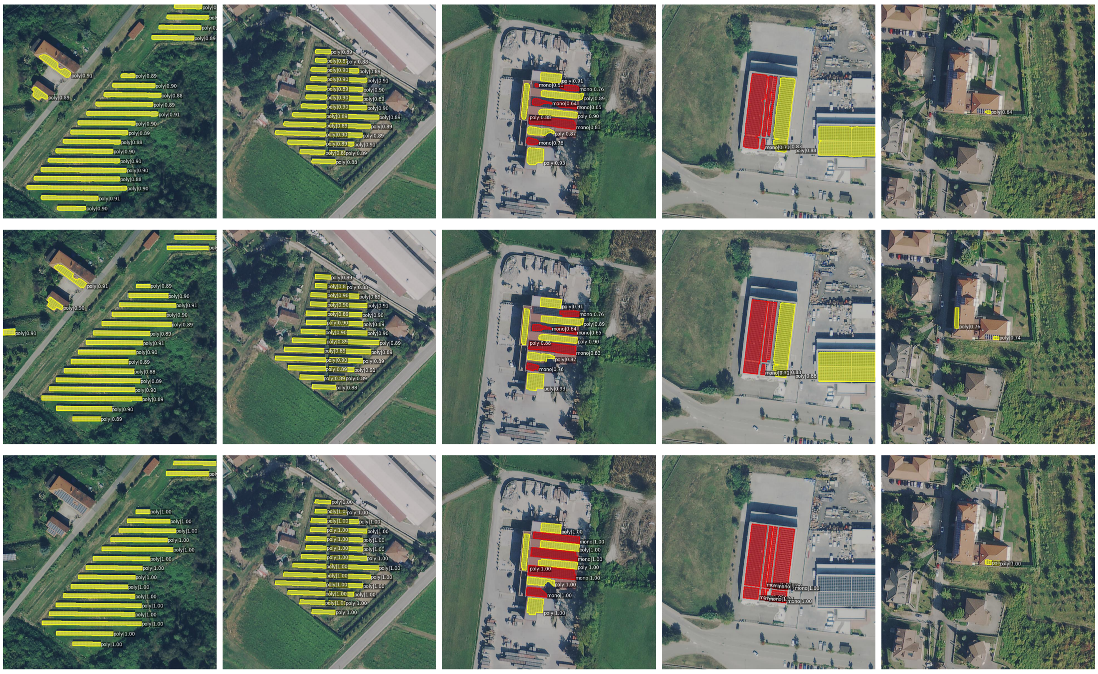
Encoder	Bands	Dataset	Backgr.	Mono.	Poly.	MIoU	mIoU
ResNet50	RGB	ext.	93.19	0.00	57.59	50.26	87.24
ConvNext	RGB	ext.	94.12	5.14	58.34	52.53	88.76
ResNet50	RGB-NIR	ext.	94.78	0.00	68.74	54.51	90.85
ConvNext	RGB-NIR	ext.	95.21	5.95	70.05	57.07	91.05
ResNet50	RGB	base	94.97	0.00	82.71	59.23	91.25
ConvNext	RGB	base	95.62	0.00	85.08	60.24	92.17
ResNet50	RGB-NIR	base	96.32	36.46	86.01	72.93	93.41
ConvNext	RGB-NIR	base	97.24	45.90	86.83	76.66	94.37
ResNet50	RGB	base-ms	97.34	37.50	81.46	59.48	93.84
ConvNext	RGB	base-ms	97.32	24.74	83.03	53.88	94.06
ResNet50	RGB-NIR	base-ms	98.25	58.88	85.69	80.94	96.57
ConvNext	RGB-NIR	base-ms	98.17	62.49	85.53	82.86	96.53
ConvNext	RGB-NIR	ms+post	**98.63**	**74.12**	**86.91**	**86.55**	**97.34**

Table 4.7: Semantic segmentation results using ResUnet with various combinations of encoders, bands, and datasets. The last row provides an assessment of the best segmentation model with additional post-processing.

an additional infrared band during training proves to be highly beneficial for the monocrystalline category, improving its class IoU to 45.90, as well as for the other categories, with improvements of at least one point each. Contrary to our expectations, dataset extension through copy-paste augmentation is detrimental for the semantic segmentation task. The model achieves good results only on the easier polycrystalline class, reaching a maximum $MIoU$ of 57.07, which does not improve upon the baselines. We hypothesize that this is due to label contamination [171] introduced by the copy-paste mechanism. While pixels entirely inside or outside the PV panels correctly represent their respective classes, the copy-paste mechanism introduces strong discontinuities in the visual patterns along these borders, which can adversely affect the pixel-wise prediction mechanism of semantic segmentation approaches. Our best results, and the overall best performance across all experiments, are obtained using multiscale training, denoted by the *ms* entries in Table 4.7. The introduction of this consistency regularization across scales leads to a +6.2 $MIoU$ improvement over the best baseline, with the most significant contribution coming from the monocrystalline class, which reaches an IoU of 62.49 without further post-processing. The additional infrared band once again proves crucial for accurately defining this category, while the polycrystalline and background classes exhibit comparable results. As a final test, we assess the effectiveness of our post-processing algorithm by extracting regularized polygons from the raw predictions and subsequently converting them back to raster tiles for evaluation. With the application of this procedure alone, we observe a substantial improvement in the monocrystalline category, achieving an IoU of 74.34, an increase of +11.63 from the highest raw value obtained. The improvements over the raw predictions can be visually observed in Fig. 4.9, where the regularized polygons are compared with the raw output and the original ground truth. When examining large-scale installations (columns (a) and (b) in Fig. 4.9), we note that the segmentation approaches provide excellent results. However, for smaller industrial and agricultural plants, they may experience a performance decline: while models provide precise localization, they may lose parts of the panel surface, or misclassify the panel type due to brightness and contrast differences. This is especially visible the case in smaller domestic installations, where noise and variety are much higher. Interestingly, both approaches are able to identify additional panels that were not present in the ground truth annotations, due to installations not yet registered or mapped, as evident in columns (d) and (e) in Fig. 4.9. In conclusion, our experiments demonstrate the effectiveness of the semantic segmentation task in PV panel detection, particularly when utilizing ad-hoc measure such as multiscale training and post-processing techniques. The inclusion of an infrared band proves highly beneficial, especially for the challenging monocrystalline category. While the copy-paste augmentation did not yield the expected improvements, it provided valuable insights into the challenges of dataset extension for semantic segmentation tasks.

4.4 Summary

In this chapter, we explored two significant challenges in remote sensing: class imbalance and scale imbalance. To address class imbalance in flood delineation from satellite imagery, we constructed the MMFlood dataset comprising SAR imagery, DEM data, and hydrography maps. We evaluated various deep learning models and techniques on this dataset, including Entropy-Weighted Sampling (EWS) to focus training on more informative samples, and multi-encoder architectures to effectively fuse the different data modalities. The results demonstrated the benefits of these approaches in improving flood delineation accuracy, especially when dealing with the inherent class imbalance between flooded and non-flooded areas. For tackling scale imbalance, we focused on the task of photovoltaic (PV) panel segmentation from VHR aerial imagery. We constructed a custom dataset covering the Piedmont region in Italy with over 9,000 annotated PV panels. We introduced a multiscale regularization approach to encourage consistency between local and global features during training. Experiments showed that this technique, using of the multispectral input, significantly improved performance on the PV panel dataset, especially for the challenging monocrystalline category. Moreover, we developed a post-processing algorithm to refine the segmentation output and generate cleaner polygonal representations of the detected panels. The main findings and



(a) (b) (c) (d) (e)

Figure 4.9: Qualitative results on different scenarios: large-scale installations (a, b), industrial and agricultural plants (c, d), and domestic installations (e). The rows present the raw semantic segmentation output, the post-processed regularized polygons, and the ground truth annotations. The model accurately delineates PV panels in large-scale installations, while providing precise localization in smaller scenarios. The post-processing step improves the delineation of monocrystalline panels, resulting in more visually appealing polygons. Notably, the model identifies additional panels not present in the ground truth annotations (fourth and fifth columns).

conclusions of this chapter can be summarized as follows. First, class and scale imbalance are common challenges in remote sensing applications that necessitate the development of specialized methodologies. The experimental results demonstrate that techniques such as EWS and multi-encoder architectures are valuable tools for mitigating the effects of class imbalance. Furthermore, the adoption of multiscale training paradigms and the integration of domain-specific post-processing algorithms can yield substantial improvements in segmentation performance, particularly for objects exhibiting significant scale variability, such as photovoltaic panels. Future works could be aimed at addressing some shortcomings observed in these works. For instance, the MMFlood dataset could be further enhanced by incorporating additional modalities and time series data, potentially leading to more comprehensive and accurate flood delineation models. Similarly, the PV panel dataset currently covers a relatively limited geographical area, and expanding its scope to include a wider range of regions and landscapes could improve the generalization capabilities of the trained models. Moreover, exploring more advanced training techniques, such as self-supervised learning, may be promising for future research in these domains, given the scarcity of ad-hoc labels in specific tasks.

Chapter 5

Domain Robustness and Weak Supervision

5.1 Introduction

In every application scenario, the effectiveness of deep learning models heavily relies on the availability of large-scale, high-quality annotated datasets, which are often time-consuming and expensive to obtain. This is especially true in the context of semantic segmentation in remote sensing, where two of the major challenges that hinder the performance and generalization of models are the domain shift problem and the scarcity of annotated data. Domain shift occurs when models trained on one dataset fail to generalize well to another dataset with different characteristics [71], such as different sensors, resolutions, or geographical locations. This issue is particularly prevalent in remote sensing applications, where the data can exhibit significant variations in terms of spectral, spatial, and temporal properties. On the other hand, the limited availability of high-quality annotated data can limit the performance of supervised learning methods, as obtaining such annotations is often time-consuming and expensive. To address the domain shift problem and improve the overall robustness of the models to real-world scenarios, in this chapter we investigate four different techniques: Unsupervised Domain Adaptation (UDA), semantic segmentation from sparse labels, multitask learning, and applying recent Vision Foundation Models (VMFs) for automated annotations. In the UDA task, we have access to labeled data from a source domain and unlabeled data from a target domain, and the goal is to adapt the model trained on the source domain to perform well on the target domain without using any labels from the target domain. Recent techniques typically carry out this objective using self-training techniques with a *teacher-student* paradigm, where the former produces stable and confident pseudo-labels for the latter to learn from [96, 223]. In this section, we propose a novel framework called HIUDA (Hierarchical Instance Mixing for Unsupervised Domain Adaptation) that introduces two key components: (1) a new mixing strategy called

HIMix (Hierarchical Instance Mixing) that extracts connected components from the semantic masks and mixes them according to a semantic hierarchy, and (2) a twin-head architecture that produces finer pseudo-labels for the target domain. By addressing the shortcomings of existing domain mixing strategies in aerial imagery, HIUDA aims to improve the domain robustness of semantic segmentation models.

To tackle the limitations of annotated data, both in terms of extent and reliability, we explore techniques for learning from weak labels, such as scribbles or sparse annotations. In this context, our objective is to expand the sparse labels to a dense map, effectively assigning a semantic category to every pixel in the image. We reformulate this task in a manner that closely resembles UDA), by treating the labeled pixels as our source domain and the unlabeled pixels as our target domain. We propose a framework called SPADA (SParse Annotations with DAformer) that leverages sparse annotations and self-training to improve semantic segmentation performance. SPADA utilizes a teacher model to generate pseudo-labels on the target domain, which are then mixed with the sparse ground truth labels to train the student model in a self-supervised manner. By effectively exploiting both labeled and unlabeled pixels during training, SPADA aims to enhance the performance of semantic segmentation models in scenarios with limited annotated data.

In addition to domain adaptation and weak supervision, we also investigate the potential of multitask learning to improve the robustness and performance of models in remote sensing applications. In Section 5.4 we focus on the task of burned area delineation, which aims to identify and delineate areas affected by wildfires from satellite imagery. The main challenges in this task arise from two primary sources. First, we observe an inherent imbalance in the pixel distribution between burned and unburned regions, as wildfires typically affect a relatively small portion of the total area. Second, the geographical locations of wildfires introduce an unavoidable bias in the data, as these events occur more frequently in temperate forested regions. These challenges, similar to those encountered in Section 4.2, can hinder the performance and generalization of models trained for burned area delineation. To mitigate these issues, we propose RoBAD (Robust Burned Area Delineation), a multitask learning framework that incorporates land cover classification as an auxiliary task to guide the training of the burned area delineation model towards more robust and generalizable features. By learning shared representations between the two tasks, we demonstrate that the multitask approach yields more stable and robust performance compared to single-task learning, especially in the absence of pretrained solutions. This highlights the potential of multitask learning to improve the generalization and robustness of models in remote sensing applications.

Finally, we investigate recent machine learning advances in addressing the challenges of large-scale annotation in remote sensing. The emergence of foundation models in computer vision [112, 267] has opened new possibilities for transfer learning and generalization across diverse tasks. These models, trained on extensive and varied datasets, have shown remarkable adaptability to different domains with minimal additional training [267]. Building on this progress, we develop FMARS (Foundation Model Annotations in Remote Sensing), an approach that exploits state-of-the-art vision models to automate the annotation process for high-resolution satellite imagery. FMARS combines VFMs and existing large-scale datasets to generate detailed labels for key elements in disaster-prone areas, such as buildings, transportation networks, and vegetation. By applying this method to imagery from numerous disaster events worldwide, we demonstrate its potential to rapidly produce large datasets for training specialized models. Furthermore, we show how techniques like UDA can be adopted to effectively train smaller models on these automatically generated annotations, allowing for more efficient and scalable solutions.

In summary, our work addresses the challenges of domain robustness and weak supervision in semantic segmentation for remote sensing. We propose novel techniques for unsupervised domain adaptation (HIUDA), learning from sparse annotations (SPADA), multitask learning (RoBAD), and automated annotations (FMARS) which aim to improve the generalization and performance of models in real-world scenarios. The contents presented in this chapter derived in the following publications:

- E. Arnaudo, A. Tavera, C. Masone, F. Dominici and B. Caputo, “Hierarchical Instance Mixing Across Domains in Aerial Segmentation”, in IEEE Access, vol. 11, pp. 13324-13333, 2023.

- M. Galatola, E. Arnaudo, L. Barco, C. Rossi and F. Dominici, “Land Cover Segmentation with Sparse Annotations from Sentinel-2 Imagery”, IGARSS 2023–2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 2023, pp. 6952-6955.

- E. Arnaudo, L. Barco, M. Merlo and C. Rossi. “Robust Burned Area Delineation through Multitask Learning”, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2023, MACLEAN workshop (Best paper award).

- E. Arnaudo, J. L. Vaschetti, L. Innocenti, L. Barco, D. Lisi, V. Fissore, C. Rossi, “FMARS: Annotating Remote Sensing Images for Disaster Management Using Foundation Models”, IGARSS 2024–2024 IEEE International Geoscience and Remote Sensing Symposium, 2024.

5.2 Unsupervised Domain Adaptation in Aerial Settings

While significant advances have been made in semantic segmentation, particularly with the advent of deep learning models [197, 39, 270], most of these approaches rely on the availability of large amounts of annotated data. However, collecting such pixel-level annotations is an extremely time-consuming and costly process [52]. This limitation becomes even more pronounced in the context of aerial and remote sensing imagery, where the scenes can be vast and complex, and the cost of acquiring labeled data is prohibitively high. To address this challenge, we focus on the problem of unsupervised domain adaptation (UDA) for aerial semantic segmentation, a task that can be seen as a form of transfer learning. The aim is to adapt a model trained on a labeled source domain to perform well on an unlabeled target domain, by leveraging the unlabeled data from the target domain during training. In our setting, we assume access to a set of labeled aerial images from a source domain, as well as a collection of unlabeled aerial images from a target domain that we wish to segment. The goal is to learn a segmentation model that can accurately predict the semantic labels of the target domain images, without requiring any manual annotations from this domain. However, applying UDA techniques to aerial semantic segmentation poses several unique challenges. First, there can be a significant disparity in the pixel-level class distributions between the source and target domains. For instance, the source domain may contain mostly urban scenes with buildings and roads, while the target domain may be predominantly rural with large expanses of vegetation and water bodies. This class imbalance can lead to a domain shift that is difficult for the model to overcome. Secondly, unlike in other applications such as autonomous driving [52], aerial scenes often lack a consistent structural layout. In driving scenarios, for example, the images typically have a regular arrangement with the road at the bottom, buildings on the sides, and sky at the top. This consistent structure is often preserved across different domains. In contrast, aerial scenes can have objects and regions arranged in arbitrary configurations, which makes it harder to transfer knowledge from the source to the target domain. To tackle these challenges, we propose a novel framework for UDA in aerial semantic segmentation, called HIUDA (Hierarchical Instance Mixing for Unsupervised Domain Adaptation), introducing two technical improvements over standard solutions. First, we propose a new domain mixing strategy called HIMix (Hierarchical Instance Mixing). HIMix operates at the instance level by extracting connected components from the semantic labels of both the source and target images. It then mixes these instances in a hierarchical manner based on their pixel counts, placing the smaller objects on top of the larger regions. This approach preserves the semantic structure of the objects and mitigates the class imbalance between the domains, generating more realistic and coherent mixed images for training. Second, we introduce a twin-head architecture for the segmentation network, where two separate decoder heads are trained in a contrastive manner. The two heads share the same encoder backbone but receive differently augmented versions of the input images. By enforcing consistency between the outputs of the two heads, the model learns more robust and invariant features that can better generalize to the target domain. Additionally, the agreement between the two heads is used to generate reliable pseudo-labels for the target domain images, providing additional supervision during training.

We evaluate our HIUDA framework on the LoveDA dataset [242], one of the few benchmarks for UDA in aerial semantic segmentation. Our experiments demonstrate that HIUDA outperforms existing state-of-the-art UDA methods on this dataset, achieving significant improvements in segmentation accuracy on the target domain. We also provide detailed ablation studies to analyze the impact of our proposed components. In summary, in this section we provide the following contributions: (i) we identify and analyze the key challenges in applying UDA techniques to aerial semantic segmentation, namely class imbalance and lack of structural consistency, (ii) we present a novel UDA framework, HIUDA, which introduces a hierarchical instance mixing strategy (HIMix) and a twin-head architecture to address these challenges, and (iii) we carry out an extensive set of experiments on the LoveDA dataset, including an ablation study, and demonstrate the improvements of HIUDA over existing UDA methods for aerial semantic segmentation.

5.2.1 Related Works

Segmenting Aerial Images Semantic segmentation in aerial imagery presents unique challenges that require tailored solutions, despite sharing some common features with other domains. While encoder-decoder architectures [135, 39, 270, 197] and recent vision Transformers [61, 131, 215, 255] have shown promising results, aerial scenes often contain a large number of entities with complex spatial relationships. These can be modeled, for instance, through the use of attention mechanisms [164] or relation networks [163] to capture long-range dependencies. In the context of unsupervised domain adaptation (UDA), ensuring feature consistency across domains is crucial for effective training. Aerial images pose additional challenges due to their top-down perspective and lack of reference points common in natural images. As described in Chapter 3 and Chapter 4, some works have already exploited this peculiarity by learning rotation-invariant features [91, 218] or employing regularization techniques [8]. Here, we argue that invariance to a broad range of geometric and photometric transformations is also beneficial for generalization in the UDA setting. Similar to the previously described methods, with the twin-head architecture we aim to extract comparable features from the same image, although augmented in different ways, following a contrastive-like approach. Furthermore, aerial images often exhibit significant class imbalance, with small objects like cars positioned on top of large portions of land. Although techniques such as class-balanced sampling and weighting [96] or specialized loss functions [110] can help mitigate this issue, this is not enough in the context of domain-adaptive segmentation. To address this challenge, we propose HIMix, a hierarchical mixing approach that aims to balance the class distribution and respect the relative depth ordering of objects. By ensuring that smaller objects appear on top of larger ones during mixing, HIMix prevents large surfaces from overwhelming less frequent categories, enabling fairer training.

Domain-Adaptive Segmentation. Domain Adaptation (DA) is a challenging problem in semantic segmentation, where models trained on one domain often fail to generalize well to new domains with different data distributions. The main objective of DA is to bridge the *domain shift* between the source and target domains, enabling effective knowledge transfer and improved performance on the target task. Early DA approaches focused on minimizing statistical measures of divergence between the source and target feature distributions. For example, DAML [82] proposed a metric learning approach to align the domains in a shared feature space, while similar work [136, 227] utilized Maximum Mean Discrepancy (MMD) to measure and reduce the distribution mismatch. In past years, adversarial training emerged as a prominent approach for domain adaptation in semantic segmentation [224, 139, 258, 219]. These techniques involve training a discriminator to differentiate between source and target features, while the segmentation network aims to deceive the discriminator by aligning the feature distributions. However, aligning features at a global level can occasionally result in mismatches between classes, causing semantically distinct samples to be mixed in the feature space. Another line of work explores image-to-image translation for domain adaptation, represented by methods like CyCADA [95], DCAN [251], and FDA [258]. The goal of these approaches is to generate source-like images from the target domain or vice versa, therefore minimizing the discrepancies in low-level visual characteristics between domains. However, they do not often directly tackle the potential mismatches in the texture and semantic content of classes across the datasets. More recently, self-training techniques have demonstrated promising results for unsupervised domain adaptation. These methods exploit the source model's predictions on unlabeled target data to generate pseudo-labels for fine-tuning. For example, PyCDA [125], CBST [274], and IAST [150] employ highly confident pseudo-labels to gradually adapt the model to the target domain. However, a potential drawback of these approaches is confirmation bias, where the model becomes overconfident in its predictions for easily adaptable classes while struggling to learn harder or less frequent classes. Current state-of-the-art methods, such as DACS [223] and DAFormer [96], integrate self-training with image mixing strategies to mitigate the impact of noisy pseudo-labels and enhance class balance during adaptation. These techniques have proven highly successful on datasets with consistent scene structure, like driving scenarios [52]. By randomly copying objects from source to target images guided by the predicted segmentation masks, these methods drive models towards learning domain-invariant features and prioritize semantic content over low-level domain discrepancies. Despite their success in driving scenes, we argue that these mixing-based UDA methods face several challenges when applied to aerial imagery. Firstly, as discussed in Chapter 4, aerial scenes typically comprise objects with significantly varied scales. This peculiarity can in turn lead to severe class imbalance in the mixed images if not properly addressed. Secondly, aerial images lack the strong structural regularities found in driving scenes, meaning that naive mixing can produce semantically inconsistent or unrealistic object arrangements. HIMix aims to mitigate these issues, by overlaying objects on top of each other based on their size, maintaining consistency and pixel balance across categories.

5.2.2 Method

In this section, we present our proposed framework HIUDA, which addresses the limitations of existing approaches, further narrowing the domain gap between source and target datasets. Our methodology combines two key components: (i) a novel hierarchical instance mixing strategy (HIMix) that balances the class distribution and preserves the semantic consistency of mixed images, and (ii) a twin-head architecture that enhances the quality and consistency of pseudo-labels for self-training. We first provide details on the HIMix strategy, describing how it carries out the instance-based mixing from the source and target segmentation masks, balances the class distribution, and composes the mixed images while respecting the relative depth ordering of objects. Next, we describe the twin-head architecture and its role in improving the pseudo-label quality and consistency through contrastive learning. Last, we present the overall training procedure, which integrates the HIMix strategy and the twin-head architecture into a unified UDA framework.

Problem Statement

We focus on Unsupervised domain adaptation (UDA) in semantic segmentation applied to aerial imagery. UDA is a subfield of transfer learning that aims to adapt a model trained on a source domain, where labeled data is available, to a target domain, where only unlabeled data is accessible. By addressing this problem, we try to enhance the performance of aerial semantic segmentation models in novel scenarios with limited annotated data. Formally, let $\mathcal{X}$ denote the set of images (in this case RGB), each composed of a set of pixels $\mathcal{I}$, and let $\mathcal{Y}$ represent the corresponding set of semantic masks, which assign a class label from the set of semantic classes $\mathcal{C}$ to each pixel $i \in \mathcal{I}$. In the UDA setting, we have access to two distinct datasets during training: (i) a labeled source domain dataset $\mathcal{X}_S = (x_S, y_S)$, where $x_S \in \mathcal{X}$ and $y_S \in \mathcal{Y}$, and (ii) an unlabeled target domain dataset $\mathcal{X}_T = (x_T)$

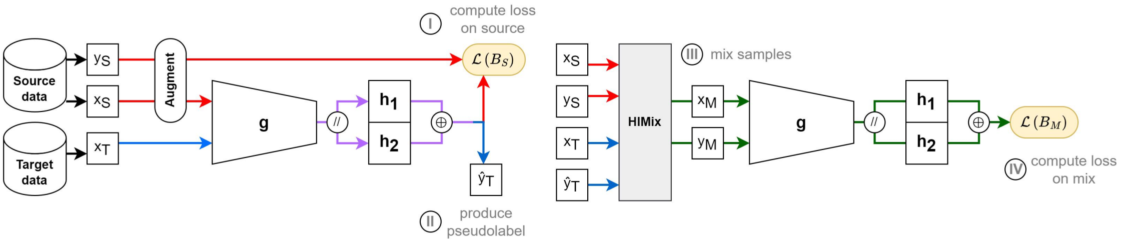


Figure 5.1: Overview of our proposed training framework: (i) The model undergoes standard supervised training on the source domain using the source image (x_S) and its corresponding ground truth label (y_S), optimizing the segmentation loss $\mathcal{L}_{seg}(B_S)$. (ii) For the target domain, pseudo-labels $\hat{y}_T$ are generated from the target image (x_T) by encoding it with the shared backbone g and performing majority voting between the outputs of the two segmentation heads (h_1 and h_2). (iii) The source and target samples are then combined using the Hierarchical Instance Mixing (HIMix) strategy, resulting in a new pair of mixed samples x_M and y_M. (iv) Finally, the segmentation loss $\mathcal{L}_{seg}(B_M)$ is computed on the mixed pairs to optimize the model for domain adaptation.

(i.e., without labels y_T) with N_T images, where $x_T \in \mathcal{X}$. Our objective is to learn a parametric function f_θ that maps an input image to a pixel-wise probability distribution over the semantic classes, i.e., $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^{|\mathcal{I}|\times|\mathcal{C}|}$. Once obtained, the goal is to apply this function on unseen images from the target domain and achieve high segmentation accuracy. We denote the model's output probability for pixel i and class c as $p_i^c(x) = f_\theta(x)[i, c]$. In line with common semantic segmentation approaches [135, 39, 96], we train the model parameters θ by minimizing a standard categorical Cross-Entropy (CE) loss:

$$\mathcal{L}(x, y) = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{c\in\mathcal{C}} y_i^c \log(p_i^c(x)), \tag{5.1}$$

where y_i^c is the ground truth label for pixel i and class c. Although other loss functions, such as the combination of CE and Dice loss [242], could potentially improve performance in aerial settings, our primary focus in this work remains on the UDA problem itself. Therefore, we adopt the CE loss as it provides a fair and consistent comparison with other state-of-the-art methods [96].

Hierarchical Instance Mixing

One main component of our framework is the Hierarchical Instance Mixing strategy (HIMix), which addresses the limitations of existing mixing techniques like ClassMix [171] in the aerial image domain. We observe that directly superimposing source domain instances onto the target domain without considering their semantic

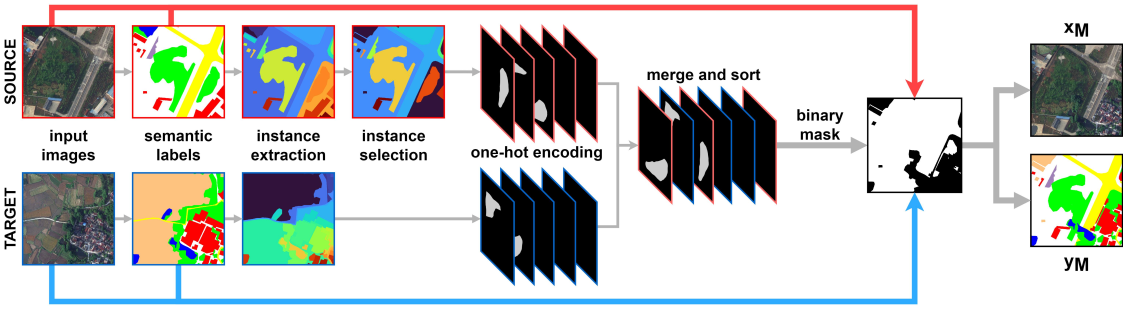


Figure 5.2: Visual representation of the Hierarchical Instance Mixing (HIMix) strategy. HIMix involves four steps: (i) extraction of the connected components (instances) from the source ground truth and target pseudo-label, (ii) random selection of a subset of source instances to be mixed, (iii) hierarchical merging of the selected source instances with the target instances based on their size, placing smaller instances on top of larger ones, and (iv) generation of a binary mask M to create the final mixed image x_M and its corresponding label y_M.

hierarchy can lead to unrealistic compositions, such as large fields appearing on top of buildings, or roads superimposed on houses. In this context, given pairs (x_S, y_S) and $(x_T, \hat{y}_T)$, where $\hat{y}_T = f_\theta(x_T)$ are pseudo-labels derived from model predictions on the target domain, our goal is to create a new pair (x_M, y_M). This pair integrates content from both the source and target domains using a binary mask M. HIMix carries this out in two key steps: (i) instance extraction, and (ii) hierarchical mixing.

For the instance extraction phase, we note that aerial images often provides uniform land cover, with multiple instances of the same category within a single image. In the absence of explicit instance labels, we can leverage this characteristic to segment semantic annotations into connected components. By definition, a *connected component* consists of pixels with the same semantic label, with paths between any two pixels entirely within this set [85]. This is visible in Fig. 5.2, where, for instance, two semantically equivalent areas displaying a forest are segmented into two separate components by a road. This method increases the number of selectable regions for mixing, helping balance the pixel distribution between source and target domains in the final mixed sample. This procedure is applied to both source and target labels, so that the different layers can be shuffled with more variation.

For the mixing phase, we argue that, in aerial imagery, objects exhibit an inherent hierarchy dictated by their semantic categories. For instance, land cover types like *barren* or *agricultural* generally form the background relative to instances like *roads* or *buildings*. In addition, the former usually cover a larger surface than the entities of top of them. In our mixing process, we attempt to maintain this hierarchy when combining source and target instances, as illustrated in Fig. 5.2, simplifying the sorting phase using the raw per-entity pixel count. First, we encode both sets of instance labels into a one-hot representation, producing separate mask layers for each component. We then merge and sort these layers by the amount of pixels equal to 1, positioning larger layers at the bottom. Finally, a reduction from top to bottom projects the 3D tensor into a 2D binary mask M, where positive values indicate *source* pixels, and null values indicate *target* pixels. In summary, HIMix constructs mixed images that maintain the semantic hierarchy of the visual elements, thus improving the efficacy of UDA training in aerial segmentation tasks.

Twin-Head Architecture

To complement our HIMix procedure, we introduce a twin-head segmentation framework aimed at overcoming the limitations of existing self-training UDA strategies. While teacher-student approaches like DAFormer [96] aim to improve the temporal consistency of pseudo-labels, they do not directly address inconsistencies in geometry or style. The twin-head architecture directly addresses these shortcomings by generating more stable pseudo-labels. As illustrated in Fig. 5.1, our framework comprises a shared encoder g followed by two parallel lightweight segmentation decoders, h_1 and h_2. The model is trained end-to-end, leveraging labeled source data and online pseudo-labels computed on the target images.

For source training, we feed the two heads with contrastive variations of each source image to encourage the learning of augmentation-invariant representations. Similar to the Augmentation Invariance approach presented in Section 3.2, at each iteration, given a source image x_S and its ground truth y_S, we apply a random sequence of geometric transformations $\mathcal{T}_g$ (horizontal flipping, rotation) and photometric augmentations $\mathcal{T}_p$ (color jitter) to obtain an augmented pair $\tilde{x}_S = T_p(T_g(x_S))$ and $\tilde{y}_S = T_g(y_S)$. The concatenated augmented batch $B_S = (x_S \bigoplus \tilde{x}_S, y_S \bigoplus \tilde{y}_S)$ is passed through the shared encoder g to extract features, which are then split and fed into the two heads to obtain outputs $h_1(g(x_S))$ and $h_2(g(\tilde{x}_S))$. A standard cross-entropy loss is computed on both outputs, as reported in Eq. (5.1). By operating independently on different image variations, the two heads can evolve differently while optimizing the same objective, while sharing the encoder yields a contrastive-like feature extraction that is more robust to perturbations, which is key for generating stable and accurate pseudo-labels.

For the mixed training, the twin-head architecture is explicitly designed to generate refined pseudo-labels, however the two heads produce independent outputs that need to be merged consistently. Given an unlabeled target image x_T, we compare the probabilities $\sigma(h_1(g(x_T)))$ and $\sigma(h_2(g(x_T)))$ obtained by forwarding the image to both heads and passing them through a Softmax function σ to normalize them, and select the more confident value between the two. Once the probability p_i^c is derived for every pixel i and class c, the pseudo-label $\hat{y}_T$ necessary for the mixing strategy is generated for each target input x_T through the following formula:

$$\hat{y}_T^{(i,c)} = [c = argmax_c \, p_i^c(x_T)] \tag{5.2}$$

Once a pseudo-label is obtained, it can be mixed with the source label using HIMix. The mixed pairs (x_M, y_M), respectively composed of source and target samples, are computed by extracting the connected components, sorting and rearranging the labels, as described in previous section. To further improve the generalization abilities of the model, a batch composed of the original mixed images and their transformed versions $B_M = (x_M \bigoplus \tilde{x}_M, y_M \bigoplus \tilde{y}_M)$ is generated using comparable geometric and photometric transformations $\mathcal{T}_g$ and $\mathcal{T}_p$, then fed to the model to compute $\mathcal{L}(B_M)$. To reduce the impact of low-confidence areas, a pixel-wise weight map w_M is generated following the approach proposed in previous works [96]. For each pixel i, w_M is computed as the percentage of valid points above a threshold:

$$w_M^i = \begin{cases} 1, & i \in y_S \\ \\ \frac{m_\tau}{|\mathcal{I}|}, & i \in \hat{y}_T \end{cases} \tag{5.3}$$

where m_τ represents the Max Probability Threshold [123] computed over pixels belonging to the pseudo-label:

$$m_\tau^i = \mathbb{1}_{[\operatorname{argmax}_c \; p_i^c(x_T) > \tau]} \tag{5.4}$$

In practical terms, each pixel in the mixed label receives a weight of 1 if it originates from the source domain. For pixels derived from the target domain, the weight is calculated based on the proportion of pixels exceeding the confidence threshold, normalized by the total number of pixels. Once this map has been produced, it can be directly applied to the loss computation by pixel-wise multiplication. The full training process described here is documented in Appendix B. This is not applied in multiple steps, but rather in one forward pass block, with a single backward pass on the accumulated losses.

5.2.3 Experiments

In this section, we present a comprehensive evaluation of our proposed HIUDA framework for unsupervised domain adaptation in aerial image semantic segmentation. We first describe the dataset, evaluation metric, and implementation details used in our experiments. We then compare the performance of HIUDA against various state-of-the-art UDA methods on the LoveDA benchmark [242] in every provided settings. Additionally, we conduct an ablation study to analyze the contribution of each component in our framework.

Implementation Details

We evaluate the performance of our proposed HIUDA framework on the LoveDA benchmark, designed for domain adaptation methods in remote sensing semantic segmentation. Following the protocol outlined in the original paper, we conduct experiments in two settings: *rural*→*urban* and *urban*→*rural* adaptation. Bot the *rural* and *urban* splits provide labels and an individual subdivision into training and test sets. However, in this case, we have access to labeled training data from the source domain and training data without labels from the target domain. The model's performance is then assessed on the separate test set from the target domain, considering the available ground truth for the evaluation. The LoveDA dataset is, to the best of our knowledge, the only publicly available large-scale remote sensing dataset designed for evaluating unsupervised domain adaptation (UDA) methods in land cover semantic segmentation. It comprises both urban and rural scenes from 18 different regions in China, with 1156 urban and 1366 rural training images. Each image is provided as a 1024×1024 pixel tile with annotations for seven land cover categories, namely *background*, *building*, *road*, *water*, *barren*, *forest*, and *agriculture*. LoveDA poses challenges for UDA due to the presence of multiscale objects, complex background samples, and inconsistent class distributions between the domains, as clearly visible in Fig. 5.3.

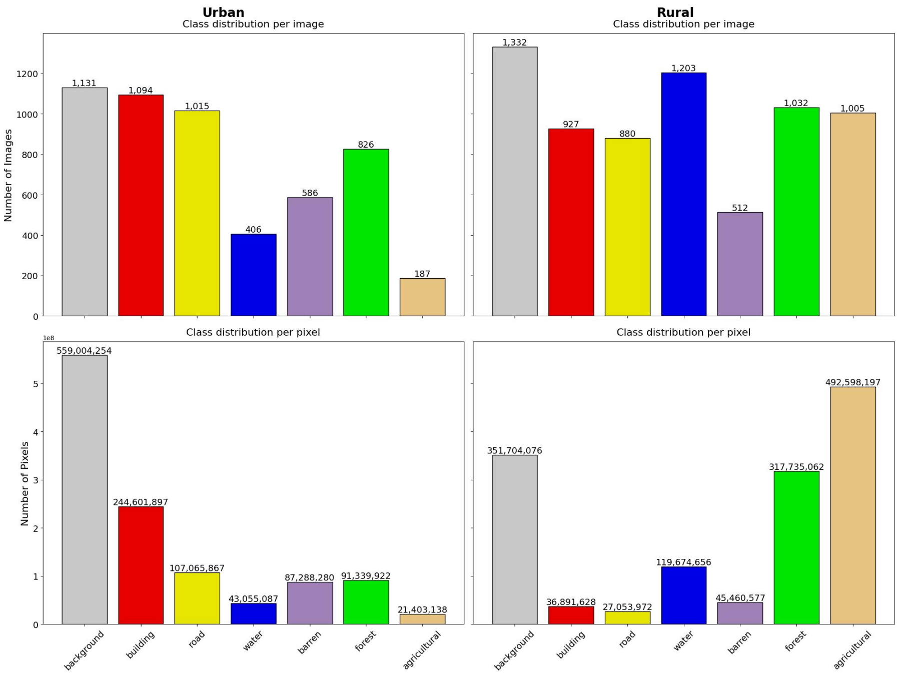


Figure 5.3: Class-wise and pixel-wise distributions across the *urban* and *rural* domains.

Following standard procedure [242], we adopt the mean Intersection over Union (mIoU) metric reported in Eq. (2.2) to evaluate the segmentation accuracy across all experiments. We compare our HIUDA approach to various state-of-the-art UDA methods. First, we include a baseline trained solely on the source domain: this should provide a lower bound and set the limit for what's achievable without specific countermeasures. In terms of domain-adaptive solutions, we compare our approach to MMD metric-based method [227], adversarial training approaches such as AdaptSegNet [224], FADA [237], CLAN [139], and TransNorm [244], as well as self-training techniques like CBST [274], PyCDA [125], IAST [150], DACS [223], and DAFormer [96]. Our HIUDA framework is implemented using the PyTorch-based *mmsegmentation* library [158]. We adopt the MiT-B5 [255] backbone, pre-trained on ImageNet, as our shared encoder, and the SegFormer head [255] as the segmentation decoder. The hyperparameters concerning the standard *teacher-student* self-training pipeline are set following the current state of the art, identified with DAFormer [96]. Specifically, we train for 40k iterations using the AdamW optimizer with a learning rate of 6×10^{-5}, weight decay of 0.01, and betas of $(0.9, 0.99)$. We employ a polynomial learning rate decay with a factor of 1.0 and a warm-up phase of 1500 iterations. Given the possible fluctuations in score in the UDA setting, we repeat each experiment three different times with three different seeds, then report the average results for better comparison. During training, we perform additional data augmentation by applying random resizing (scale range $[0.5, 2.0]$), horizontal and vertical flipping, 90-degree rotations (with probability $p = 0.5$), and photometric distortions (brightness, saturation, contrast, and hue). The confidence threshold τ for pseudo-labeling is set to 0.968 [223, 96]. All experiments are conducted on an NVIDIA V100 GPU with 32 GB of memory. Concerning the evaluation on the available test set, we do not apply any additional test-time augmentation.

Results

The results of our experiments are reported in Table 5.1 for the *Rural*→*Urban* setting, while in Table 5.2 for the *Urban*→*Rural* scenario.

Considering *Rural*→*Urban* scenarios first, the source domain primarily consists of large-scale natural objects with a limited presence of man-made structures. Nevertheless, the UDA methods under consideration effectively transfer the acquired knowledge to the target urban domain, even for underrepresented categories. Similar to previous settings, self-training approaches exhibit better performance than adversarial methods, with an average improvement of +9.1 over the source-only baseline, while adversarial techniques achieve comparable results. The two top-performing self-training models and the closest competitor outperform the source-only model by +6.0 and +15.2 in terms of mIoU, respectively. HIUDA attains a remarkable gain of +17.4 over the lower bound baseline, surpassing DACS and DAFormer by +11.4 and +2.2, respectively. The qualitative results, displayed in the bottom row of Fig. 5.4 showcase the ability of our framework to differentiate between rural and urban classes. While DACS fails to recognize *buildings* and DAFormer partially misclassifies them as *agricultural* land, our model effectively minimizes bias towards categories with larger spatial extents, yielding results that

Table 5.1: Experimental results for the UDA task in the Rural→Urban setting. The table reports the per-class IoU and the mean IoU (mIoU) for each method. Methods marked with an asterisk (*) were replicated using the original implementation to ensure a fair comparison. The best result for each class and the overall mIoU are highlighted in bold.

Method	**Backg.**	**Building**	**Road**	**Water**	**Barren**	**Forest**	**Agric.**	**mIoU**
Source Only	43.3	25.6	12.7	76.2	12.5	23.3	25.1	31.3
MCD [227]	43.6	15.4	12.0	79.1	14.3	33.1	23.5	31.5
AdaptSeg [224]	42.4	23.7	15.6	82.0	13.6	28.7	22.1	32.6
FADA [237]	43.9	12.6	12.8	80.4	12.7	32.8	24.8	31.4
CLAN [139]	43.4	25.4	13.8	79.3	13.7	30.4	25.8	33.1
TransNorm [244]	33.4	05.0	03.8	80.8	14.2	34.0	17.9	27.7
PyCDA [125]	38.0	35.9	45.5	74.9	07.7	40.4	11.4	36.3
CBST [274]	48.4	46.1	35.8	80.1	19.2	29.7	30.1	41.3
IAST [150]	48.6	31.5	28.7	86.0	**20.3**	31.8	36.5	40.5
DACS* [223]	46.0	31.6	33.8	76.4	16.4	29.3	27.7	37.3
DaFormer* [96]	49.2	47.7	55.2	**86.6**	16.5	39.5	30.8	46.5
HIUDA	**49.3**	**55.0**	**55.4**	86.0	17.1	**41.2**	**36.9**	**48.7**

Table 5.2: Experimental results for the UDA task in the Urban→Rural setting. The table presents the IoU for each class and the mean IoU (mIoU) achieved by the different methods. Approaches marked with an asterisk (*) were reproduced using the original implementation to ensure a fair comparison. The highest IoU for each class and the best overall mIoU are denoted in bold.

Method	**Backg.**	**Building**	**Road**	**Water**	**Barren**	**Forest**	**Agric.**	**mIoU**
Source Only	24.2	37.0	32.6	49.4	14.0	29.3	35.7	31.7
MCD [227]	25.6	44.3	31.3	44.8	13.7	33.8	26.0	31.4
AdaptSeg [224]	26.9	40.5	30.7	50.1	17.1	32.5	28.3	32.3
FADA [237]	24.4	33.0	25.6	47.6	15.3	34.4	20.3	28.7
CLAN [139]	22.9	44.8	26.0	46.8	10.5	37.2	24.5	30.4
TransNorm [244]	19.4	36.3	22.0	36.7	14.0	**40.6**	03.3	24.6
PyCDA [125]	12.4	38.1	20.5	57.2	**18.3**	36.7	41.9	32.1
CBST [274]	25.1	44.0	23.8	50.5	08.3	39.7	49.7	34.4
IAST [150]	30.0	49.5	28.3	64.5	02.1	33.4	61.4	38.4
DACS* [223]	20.1	50.5	35.9	60.6	09.9	35.4	17.5	32.9
DAFormer* [96]	29.5	57.9	41.8	67.1	07.6	35.3	48.1	41.0
HIUDA	**31.5**	**59.6**	**51.5**	**68.1**	08.2	37.4	**53.9**	**44.3**

closely resemble the ground truth.

Focusing instead on *Urban→Rural*, the scores highlight the challenging nature of the task. The source domain is predominantly composed of urban scenes with a high density of man-made structures like buildings and roads, but few natural elements. This inconsistent class distribution leads to negative transfer when adapting to the target rural domain, as evidenced by the performance of adversarial training and self-training methods, which is similar to or even worse than the Source Only baseline. The best-performing adversarial method, CLAN [139], achieves only a limited improvement of +1.8 over the source-only model. This limited improvement can be attributed to the difficulty in aligning features between domains with drastically different class distributions using adversarial training. Self-training approaches prove to be more effective, with DACS [223] and its class mixing strategy improving upon the baseline by +1.2 points. By generating pseudo-labels on the target domain and mixing them with source samples, DACS can partially mitigate the class distribution mismatch. However, the mixing strategy in DACS does not explicitly account for the large disparities in sizes between categories in remote sensing, limiting its effectiveness. DAFormer [96], which employs a Transformer backbone alongside the same mixing technique as DACS, surpasses the baseline by +9.3 mIoU. Its ability to capture long-range dependencies and the attention mechanism provided by Transformers contribute to its improved performance in this challenging scenario. Nevertheless, DAFormer still struggles with underrepresented classes and fails to fully exploit the hierarchical relationships between objects. Our proposed HIUDA framework, combining the twin-head architecture with the novel hierarchical instance mixing (HIMix), significantly outperforms the source-only model by a substantial margin of +12.6 and exceeds the performance of its closest competitor, DAFormer, by +3.3 points. As visible from the qualitative results in Fig. 5.4, HIUDA excels at boosting performance on underrepresented rural classes such as *agriculture*. By explicitly considering the hierarchical structure of objects during mixing, HIUDA ensures that smaller, less frequent objects like buildings are not overwhelmed by larger, more dominant classes like *agriculture* or *forest*. This enables our model to better capture the patterns and boundaries of these underrepresented categories. Compared to DACS and DAFormer, our method more accurately delineates object contours and classifies challenging categories like *water*, despite their scarcity in the source domain. The hierarchical mixing in HIUDA allows for the preservation of fine-grained details and the correct placement of objects within the scene, leading to more precise segmentation maps. HIUDA also demonstrates superior performance on classes with diverse visual appearances across domains, such as *road*, which can be paved or unpaved.

To evaluate the effectiveness of the twin-head architecture, we compare it against the traditional single-head structure, which relies on a secondary teacher network derived from the student model using an exponential moving average for generating pseudo-labels. Additionally, we investigate the potential benefits of the HIMix strategy when combined with both single-head and twin-head training. In Table 5.3, we report the results of the ablation study using the MiT-B5 [255] as main backbone for every experiment. Our findings indicate that the twin-head design, even when paired with the standard class mixing strategy (Table 5.3, ID 3), outperforms the single-head architecture (ID 1). This suggests that the twin-head approach is

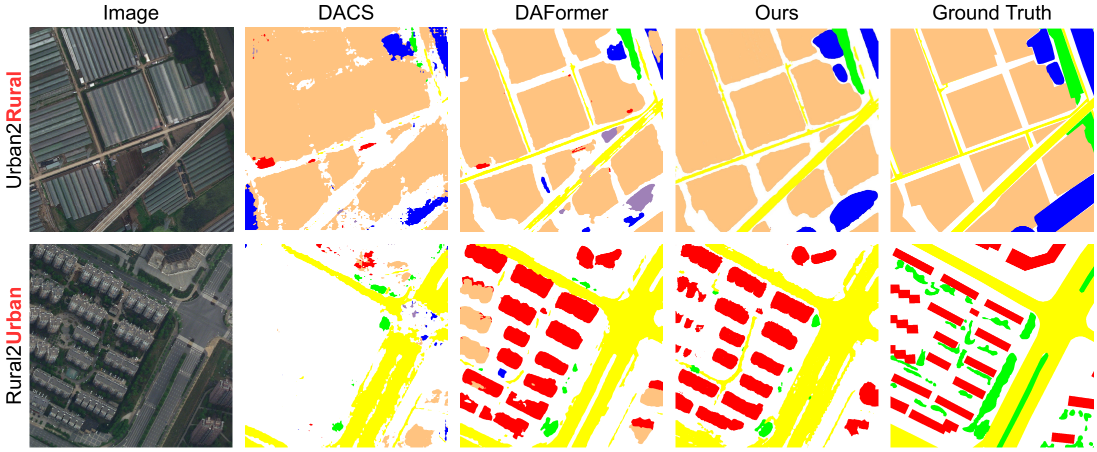


Figure 5.4: Qualitative results obtained in the *Urban→Rural* setting (top) and *Rural→Urban* scenario (bottom), with a sample image from the target domain.

Row ID	Twin-Head	ClassMix	Inst. Mix	H. Sort.	mIoU (U2R)	mIoU (R2U)
1		✓			41.0 ± 0.33	46.5 ± 0.41
2			✓	✓	43.4 ± 0.76	47.6 ± 0.10
3	✓	✓			42.9 ± 0.35	47.1 ± 0.34
4	✓		✓		43.2 ± 0.35	47.4 ± 0.16
5	✓		✓	✓	**44.3** ± 0.39	**48.7** ± 0.06

Table 5.3: Ablation study on the components of our HIUDA framework, namely the twin-head architecture and the HIMix strategy, demonstrating their individual and combined contributions to the overall performance.

more effective at generating accurate pseudo-labels with precise class boundaries, as also highlighted by the qualitative results in the first row of Fig. 5.5. The HIMix strategy enhances segmentation performance even when coupled with a single-head architecture (ID 2). This improvement is particularly evident for categories with smaller spatial extents, which are typically obstructed by larger classes when using the standard class mixing approach. As a result, in the top-right image of Fig. 5.5, the single-head model with standard mixing fails to capture the semantics of these underrepresented classes, incorrectly classifying *buildings* as *agricultural* land. In contrast, HIMix with a single segmentation head allows the model to accurately distinguish *buildings* (Fig. 5.5, bottom-right), despite the lack of precision in delineating the boundaries. The optimal results are achieved when the twin-head architecture's capacity to generate refined segmentation maps is combined with the ability of HIMix to preserve the hierarchical semantic structure of the scene (ID 5). This combination yields the highest accuracy and produces segmentation maps with the finest level of detail, as illustrated in the bottom-left quadrant on Fig. 5.5.

To further understand the individual contributions of the components in HIMix, we incrementally activate the instance extraction and hierarchical sorting phases. Our tests reveal that the connected components extraction effectively provides more variety during training, albeit with minimal improvements, while the hierarchical mixing step consistently improves upon the instance extraction stage, resulting in mIoU gains of +1.1 and +1.3 in the Urban→Rural and Rural→Urban scenarios, respectively. This demonstrates the importance of considering the hierarchical relationships between objects when mixing instances from different domains.

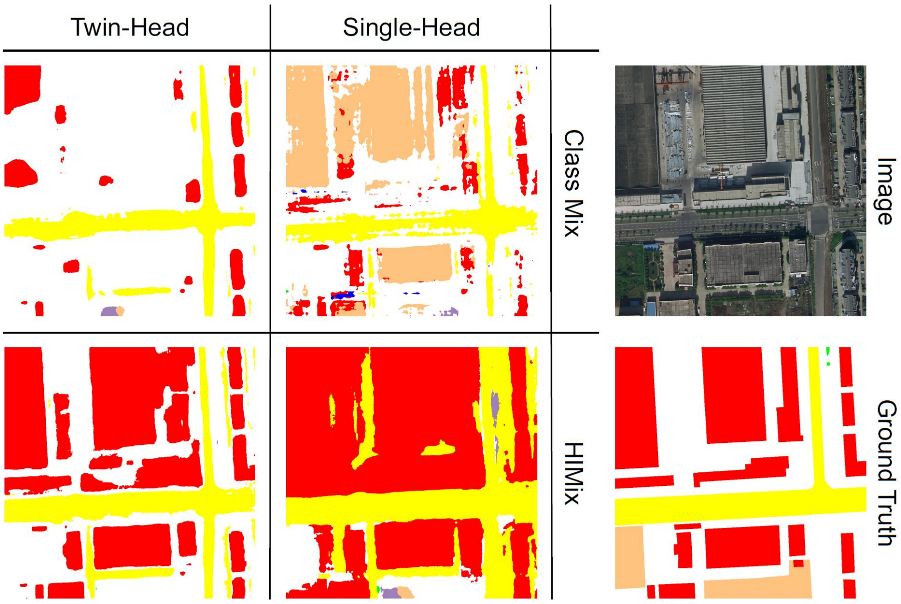


Figure 5.5: Qualitative comparison of single-head and twin-head architectures using standard class mixing or our proposed HIMix strategy.

5.3 Learning from Sparse Annotations — Land Cover Mapping

Land cover (LC) segmentation is a crucial task in remote sensing, with a wide range of applications. Accurate and up-to-date LC maps provide valuable information for many fields of application, from urban planning [199] to disaster management [64], studying the propagation and impact of calamities such as wildfires and floods. For instance, these maps can enable the distinction and delineation of highly flammable areas, like forests and shrubs, from urban borders, like buildings and roads. However, generating high-quality delineations is complex and time-consuming, requiring the expertise of multiple expert annotators and frequent updates to reflect environmental changes.

One major challenge in producing efficient and reliable maps is the trade-off between spatial resolution and coverage. Large open datasets at the European scale, such as Corine Land Cover (CLC) [53], typically offer lower spatial resolution than necessary for certain applications. On the other hand, higher resolution products like Urban Atlas (UA) are only available on a subset of regions and limited in their classification taxonomy, due to the significant effort needed for their production. The more effort and resources invested in increasing accuracy and validation, the sparser the ground truth data becomes. This issue is evident in data sources like the Land Use/Cover Area frame Survey (LUCAS), which, despite offering consistent and reliable land use and cover data for the entire EU through manual *in situ* surveys, results in observations from a very limited number of points.

To exploit the characteristics of every source at our disposal, we propose a selective combination of all these Copernicus datasets, namely CLC, UA, and LUCAS, so that the composition can leverage the strengths of each data source. By merging these datasets, we obtain a sparse ground truth that benefits from the high resolution and accuracy of UA in urban areas, the detailed land cover information of CLC, and the reliable point-wise observations of LUCAS. However, the sparsity of the resulting annotations presents a significant challenge for training semantic segmentation models effectively. Despite the improved quality and diversity of the merged dataset solves a key limitation in existing sources, the presence of sparse ground truth labels introduces a new challenge for an optimal model performance. Similar to Domain Adaptation techniques described in Section 5.2, this issue requires the application of specific approaches that can efficiently learn from limited labeled data while exploiting the available unlabeled information.

In this section, we introduce SParse Annotations with DAformer (SPADA), a novel framework for semantic segmentation that addresses the problem of land cover mapping with sparse annotations. Given the similarities, our approach employs Unsupervised Domain Adaptation (UDA) techniques within a *teacher-student* paradigm, while the main difference resides into the concept of *source* and *target* datasets. In this case, the teacher model generates robust pseudo-labels to augment the sparse ground truth annotations across the full input space. These pseudo-labels are then mixed with the processed sparse labels through a series of steps. First, the former are filtered based on a confidence threshold to ensure that only high-quality predictions are considered. Next, the filtered pseudo-labels are weighted according to their prediction confidence, giving more importance to labels with higher certainty. This weighting scheme helps to mitigate the impact of potential errors in the self-generated ground truth. Finally, the weighted pseudo-labels are combined with the sparse ground truth labels, effectively expanding the labeled data available for training the student model. This approach is comparable to the techniques used in DACS and DAFormer [96], which instead leverage two different sources altogether and adapt the model to a target domain. In practice, in this case we consider the labeled *pixels* as our source, and the unlabeled *pixels* as the label, effectively applying the UDA pipeline on the same image with a different purpose at each step.

To assess the effectiveness of SPADA, we conduct a thorough evaluation using the only two reliable ground truth datasets at our disposal: LUCAS and Urban Atlas, assessing the performance of our proposed technique against state-of-the-art semantic segmentation approaches and third-party products Our results demonstrate the effectiveness of the proposed technique, with SPADA outperforming even robust third-party products such as S2GLC [142]. In summary, the main contributions of this section are twofold: first, we propose and evaluate SPADA, a novel framework for generating land cover maps from Sentinel-2 imagery, combining vision transformers with all the information at our disposal, including labeled and unlabeled pixels during training. Second, we publicly release the dataset and code used in this work at https://github.com/links-ads/igarss-spada, including the input data and sparse annotations used to train the segmentation model, to foster further research in this area.

5.3.1 Related Works

Supervised aerial semantic segmentation presents several unique challenges compared to traditional computer vision tasks. These challenges include the high dimensionality of input data, with images often containing multiple spectral bands, the large size of the images, and the top-down viewpoint, which differs from the perspective of most computer vision datasets. Additionally, the availability of ground truth annotations is often limited, making it difficult to train models effectively. To address the challenge of high-dimensional input data, researchers have proposed various approaches. One common strategy is to include additional spectral bands by introducing multiple encoders or expanding the input layers of the model [177]. This allows the model to exploit the rich spectral information present in remote sensing imagery, with sources such as Sentinel-2. Another approach is to leverage the large input dimensions and top-down viewpoint to implement additional regularization techniques, as mentioned in previous sections. For example, GLNet [42] proposed a multiscale regularization approach that takes advantage of the spatial structure of the imagery, while ReDet [91] introduced invariance to rotation. When facing the challenge of limited ground truth annotations, weakly-supervised learning approaches are often proposed. These methods aim to reduce the reliance on precise, pixel-level labels by leveraging less accurate but readily available ground truths. A typical technique is the simple use of Class Activation Maps (CAM) or attention mechanisms to identify discriminative regions within an image. By focusing on these salient areas, labels can be propagated from the most informative pixels to the entire object, effectively expanding the annotated regions [4]. Other approaches explore the concept of semantic affinities between neighboring image regions. Methods like AffinityNet [4] and its extensions [165] train models to predict the similarity between adjacent pixels or patches based on their semantic content. By encouraging consistency among neighboring regions and penalizing those with divergent semantics, these approaches can propagate labels more accurately and produce more coherent segmentations. Sparse annotations, such as scribbles, present a unique set of challenges in semantic segmentation. Unlike dense pixel-wise labels, scribbles provide only a partial and approximate representation of the objects' boundaries. The goal in this setting is to expand the sparse ground truth to cover the entire object while maintaining semantic consistency. Various techniques have been proposed to tackle this problem, each with its own strengths and limitations. ScribbleSup [127] employs a graph-based propagation approach, where the scribbles serve as initial seeds, and labels are iteratively diffused to neighboring pixels based on their similarity. On the other hand, Tree Energy Loss [126] takes a more global perspective by constructing a minimum spanning tree among pixels based on their pairwise affinities. By enforcing consistency along the tree edges, this approach can propagate labels to distant regions while preserving semantic boundaries. FESTA [99] introduces a novel unsupervised neighborhood loss that encourages consistent predictions among nearby pixels. By leveraging the inherent spatial structure of the image, FESTA can expand the sparse annotations to cover the entire object without relying on additional labeled data. Building upon these sparse annotation approaches, we propose to carry out the label expansion process adjusting self-training UDA methods, such as DAFormer [96], to address the challenge of limited labeled data. We treat the sparsely annotated pixels as the source domain and the unlabeled pixels as the target domain, effectively casting the problem as a domain adaptation task. This formulation allows us to leverage both the available sparse annotations and the abundant unlabeled data during training. The model learns to expand the sparse labels to cover the entire image while benefiting from the pseudo-labels to improve its generalization capability.

5.3.2 Dataset

To train and validate our framework, we construct a comprehensive dataset by combining multiple data sources from the Copernicus program. The core of our dataset consists of Sentinel-2 Level-2A (L2A) cloud-free mosaics, which provide high-resolution, multi-spectral imagery covering the entire study area. These mosaics serve as the primary input for the models, offering spectral information across 12 bands at various spatial resolutions. Given our application scenario focusing on fuel map generation for wildfire management, we concentrate our study on the Mediterranean region, where wildfires are more frequent and intense. A visual representation of our training and testing areas is provided in Fig. 5.6. The ground truth for each training region consists of two types of annotations: scribble labels and point-wise labels.

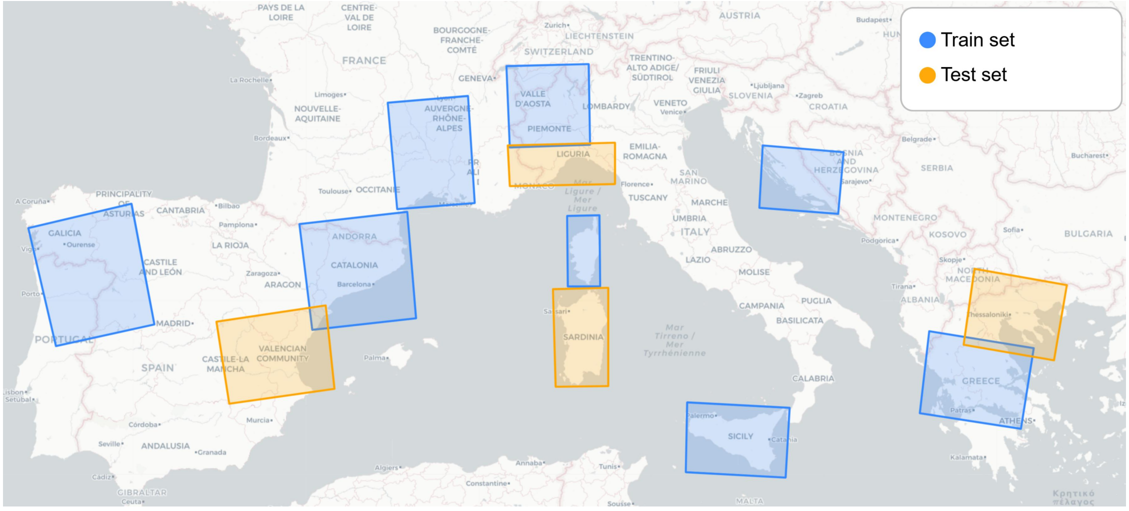


Figure 5.6: Geographical distribution of the train and test areas across Europe. Focus was given to regions with high wildfire occurrence during the years. Overlapping areas are removed during the offline tiling process to guarantee disjoint sets.

Category	CLC ID	LUCAS ID	Color
Artificial surfaces	111 112 121 122 123 124 131 132 133 142	7	🟥
Bare surfaces	331 332 335	6	⬜
Wetlands	411 412 421 422 423	–	🟪
Water	511 512 521 522 523	8 9	🟦
Grasslands	211 231 321	3	🟧
Agricultural fields	212 213 221 222 223 241 242 243 244	1 2	🟨
Broadleaved veg.	311	4*	🟩
Coniferous veg.	312	4*	🟩
Shrubs	322 323 324 333	5	🟫
Ignored	141 313 334 999	–	⬛

Table 5.4: Mapping of CLC and LUCAS class IDs to the shared land cover taxonomy. The table shows the correspondence between the original class IDs from CLC and LUCAS datasets and the aggregated classes used in the SPADA dataset. The color column indicates the RGB values assigned to each category for visualization purposes. Asterisks indicate labels that underwent disambiguation.

To build a reliable ground truth for training and evaluation, we integrate several existing land cover and land use sources. First, we include the Corine Land Cover (CLC) dataset, a pan-European land cover classification dataset that provides consistent information on land cover and land use across 39 countries. CLC offers a comprehensive, dense hierarchical classification system with 44 land cover classes, making it a valuable resource for our task thanks to its detailed taxonomy. However, the relatively poor precision of the labels (e.g., each polygon has a minimum mapping unit of 25 hectares) may limit its usability for certain applications requiring finer spatial resolution. To complement CLC and provide a more detailed representation of land cover, we incorporate the Land Use and Coverage Area frame Survey (LUCAS) dataset. LUCAS is a European-wide survey that collects harmonized data on land use and land cover through direct field observations. While LUCAS provides only point-wise annotations, the available data has been manually validated, ensuring high accuracy and reliability. Given its precise albeit local information, we leverage LUCAS for both training and validation purposes, especially considering the soil classification performance.

In addition to CLC and LUCAS, we include the Urban Atlas (UA) dataset, which focuses specifically on land cover and land use within large urban areas across Europe. UA offers a higher spatial resolution compared to CLC, with a minimum mapping unit of 0.25 hectares for urban classes and 1 hectare for other classes. However, UA covers a limited number of classes and is available only for selected urban areas. Despite these limitations, we also exploit the detailed delineations provided by UA for training and validation purposes, focusing on the segmentation performance in this case.

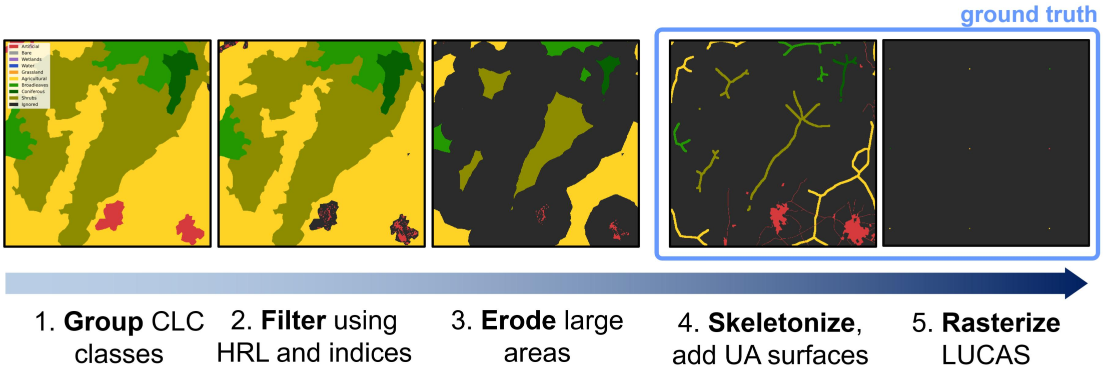


Figure 5.7: Overview of the sparse label generation process. (1) CLC classes are grouped into the target classes. (2) The grouped CLC classes are filtered using the High Resolution Layer (HRL) and spectral indices to remove potentially mislabeled pixels. (3) Large contiguous areas are eroded to create sparse labels. (4) Urban areas from the CLC map are replaced with more precise annotations from Urban Atlas (UA). (5) LUCAS points are disambiguated, rasterized and combined with the sparse labels from CLC to produce the final ground truth.

Starting from the aforementioned sources, the sparse label generation pipeline constructs two separate ground truth labels: a set of scribble annotations that exploit CLC as main source, referred to as *scribble labels* from now on, and the rasterized LUCAS annotations, named *point-wise annotations* for clarity. The overall process is displayed in Fig. 5.7 and simplified to convey the main steps and the final results.

The pipeline for the scribble labels starts at the CLC dataset, which undergoes a series of processing functions to convert the original classes into the required taxonomy. The workflow starts with the adoption of a custom simplified taxonomy to cope with the differences among datasets. To accommodate the diverse sources and maintain a focus on vegetation, we consolidate the detailed CLC hierarchy into a consistent taxonomy consisting of 9 categories:, namely: *artificial surfaces*, *bare surfaces*, *wetlands*, *water*, *grassland*, *agricultural fields*, *broadleaved vegetation*, *coniferous vegetation*, and *shrubs*. It is important to note that the *wetlands* category is absent from the LUCAS dataset, while UA does not differentiate between broadleaved and coniferous vegetation in forested areas. To address these discrepancies during the evaluation phase, we exclude the *wetlands* category when testing on LUCAS data and combine the broadleaved and coniferous vegetation classes into a single forest type when assessing performance on UA. The final aggregation is detailed in Table 5.4.

Second, following previous work [142], we filter vegetation-based classes using the Normalized Difference Vegetation Index (NDVI), the Normalized Difference Water Index (NDWI), and HRL for Impervious Surfaces, defining independent thresholds for each one of them to eliminate potentially mislabeled pixels. In practical terms, when NDVI or NDWI are lower than nominal values, or the pixel is confidently labeled as impervious surface, we remove the point from the ground truth, assigning the *ignored* label.

Third, we acknowledge that the boundaries in CLC are approximate, but we assume that the labels themselves provide a reasonable indication of the overall land cover within each delineated area. Building upon this assumption, we process the filtered CLC labels using Connected Components Labeling (CCL) [94] to identify and separate individual contiguous regions. Subsequently, these connected components are eroded by a relative percentage to create a buffer zone around the boundaries. The eroded regions are then subjected to morphological skeletonization, which reduces the thick, approximate semantic labels to thin, simplified lines. This transformation process effectively converts the precise contours and well-defined boundaries of the original CLC labels into sparse, indicative scribbles. These scribbles serve as a coarse guide, suggesting the likely presence of specific land cover categories in each area without providing exact delineations. As a last step, we further apply morphological dilation with a fixed-size kernel to each component in order to increase their thickness and transform the skeletons into actual scribbles. To enhance the precision of urban area delineations, we additionally overlay the artificial surfaces from UA on top of the scribble annotations, where available. Although UA does not provide comprehensive coverage, it offers more detailed and accurate annotations for urban regions, as illustrated in Fig. 5.7 (step n.4).

The point-wise annotations, on the other hand, are generated by rasterizing the LUCAS manual observation within the study area and assigning the closest class from the unified taxonomy to each point. This process results in a set of precise point-level labels that provide valuable information for training and validation. Since LUCAS data only provides latitude and longitude coordinates without specifying the spatial extent of each observation, we empirically determine an appropriate point size of 5×5 pixels to represent each LUCAS point in the rasterized dataset.

To further enhance the land cover information in our dataset, we incorporate the Dominant Leaf Type HRL product as a disambiguation source for forest locations in LUCAS. This layer provides a distinction between broadleaved and coniferous tree cover, enabling a more detailed characterization of forested areas. However, it is important to note that the HRL product itself is derived through an automated machine learning process. Consequently, we refrain from using the HRL data directly as ground truth. Instead, we employ it as a reliable auxiliary source for disambiguation and filtering purposes, leveraging its information to refine and improve the accuracy of our land cover labels.

To construct the final dataset, we acquire Sentinel-2 Level-2A cloudless mosaics spanning from April to August 2018, aligning with the most recent release of the datasets employed in our study. These mosaics are obtained through the Microsoft Planetary Computer platform [214] and encompass all 12 available spectral bands at 10m/pixel resolution (i.e., the maximum resolution for this source). To handle processing and training more easily, we divide each area of interest from Fig. 5.6 into non-overlapping macro-tiles of 2048×2048 pixels, utilizing the EPSG 3035 CRS to ensure consistent spatial representation across the dataset. From the macro-tiles designated for training, we further allocate 20% of the tiles for validation purposes. This results in a dataset comprising a total of 538 images for training, 135 images for validation, and 394 images for testing. Each image in the dataset is accompanied by two distinct sets of labels: scribble annotations and point-wise labels. The scribble annotations, derived from the processed CLC dataset, provide coarse indications of land cover categories, while the point-wise labels, obtained from the rasterized LUCAS observations, offer precise ground truth information at specific locations.

5.3.3 Method

In this section, we present our methodology by first formally defining the problem statement and the associated challenges. Next, we introduce the DAFormer baseline, the state-of-the-art UDA framework that serves as the foundation for our proposed approach. Finally, we describe SPADA, our novel framework that extends DAFormer to effectively leverage sparse annotations for accurate and dense predictions.

Problem Statement

In this work, we address the task of semantic segmentation for fuel mapping using sparsely annotated data. The challenge arises when only a small subset of pixels in an image are labeled with their corresponding class, while the majority of the pixels remain unmarked. Let $\mathcal{X}$ denote the set of multi-spectral input images, where each image x comprises a set of pixels $\mathcal{I}$. We define $\mathcal{Y}$ as the set of semantic annotations that assign a class label from the set $\mathcal{C}$ to each pixel $j \in \mathcal{J}$, where the number of labeled pixels is significantly smaller than the total number of pixels, i.e., $|\mathcal{J}| \ll |\mathcal{I}|$. As described in Section 5.3.2, our dataset consists of two types of sparsely annotated maps: (i) a set of *scribble* annotations, denoted as Y_S, and (ii) a set of *point-wise* annotations, referred to as Y_P. The scribble annotations provide coarse, incomplete labels for a portion of the image, while the point-wise annotations offer precise, albeit even sparser labels at specific locations. The goal is to learn a parametric function f_θ that maps a multi-spectral image to pixel-wise probabilities, i.e., $f_\theta : \mathcal{X} \rightarrow \mathcal{R}^{|\mathcal{I}| \times |\mathcal{C}|}$, and evaluate its performance on unseen images. In line with previous work [96], the model parameters θ are optimized using a standard categorical cross-entropy loss, as detailed in Eq. (5.1). The main challenges in this problem setting include: (i) dealing with the limited and sparse nature of the annotations, which cover only a small portion of the image pixels, (ii) leveraging both the scribble and point-wise annotations effectively to guide the learning process, and (iii) generating accurate and dense predictions for the entire image, including the unmarked regions, based on the sparse annotations. To address these challenges, we propose a framework that combines the strengths of UDA and self-training techniques, as described in the following sections.

DAFormer Baseline

DAFormer [96] is a state-of-the-art UDA framework that combines a powerful network architecture with effective training strategies. Starting from SegFormer [255] as reference point, the architecture of DAFormer consists of a Transformer-based encoder and a context-aware fusion decoder, specifically designed to improve domain adaptation performance. For the encoder, DAFormer employs the Mix Transformer (MiT) [255], a recently proposed variant of Vision Transformer (ViT) backbone [61] tailored for semantic segmentation. The choice of using a Transformer-based encoder is motivated by their demonstrated superior performance and robustness compared to traditional CNN-based architectures in various computer vision tasks, including semantic segmentation. The MiT encoder is designed to generate multi-level feature maps at different resolutions, allowing to capture both high-level semantic information and fine-grained details. The feature maps are progressively downsampled using overlapping patch merging, which preserves local continuity and reduces computational complexity. On the decoder side, DAFormer drops the simple MLP head from SegFormer, and introduces a context-aware fusion module that aggregates multi-level features from the encoder. Unlike previous works that only consider context information at the bottleneck features, DAFormer leverages context across features from different encoder levels. This approach is motivated by the observation that earlier features provide valuable low-level concepts at higher resolutions, which can also offer important context information for semantic segmentation. The context-aware fusion module employs depth-wise separable convolutions with varying dilation rates, inspired by the ASPP module [39], to capture multiscale context while maintaining a low number of parameters. The MiT encoder brings robustness and the ability to capture global context, while the context-aware fusion decoder effectively aggregates multi-level features and incorporates multiscale context information. The DAFormer training strategy is based upon DACS (Domain Adaptation via Cross-Domain Sampling) [223], a self-training framework based on a *teacher-student* paradigm, where the teacher network provides robust and consistent pseudo-labels on the target domain. DACS also introduces ClassMix, a mixing strategy to combine source and target information into a single image and label pair. ClassMix can be seen as a data augmentation technique that mixes two images and their corresponding labels based on the semantic classes present in the images, creating a new augmented sample that preserves the semantic context of the original images. In this case, one pair is composed of source image and source ground truth, while the second pair comprises the target image and a pseudo-label, generated by the teacher network.

To further enhance the domain adaptation performance, DAFormer introduces three key training strategies: (i) Rare Class Sampling (RCS) to address the class imbalance in the source dataset and improve the quality of pseudo-labels for rare classes; (ii) Thing-Class ImageNet Feature Distance (FD) to regularize the model by distilling knowledge from ImageNet features, focusing on thing-classes; and (iii) learning rate warm-up to gradually increase the learning rate, facilitating better feature transfer from ImageNet pretraining. In our framework, we drop the ImageNet feature distance given the remote sensing setting, while we maintain the sampling strategy and the warm-up scheduler for the experiments.

SPADA Framework

Our proposed framework, SPADA (SParse Annotations with DAformer), exploits the DAFormer architecture and training strategy to effectively leverage sparse annotations for semantic segmentation. Similarly, SPADA also combines the MiT encoder with a context-aware fusion decoder to take advantage of short and long range dependencies in remote sensing images. Drawing inspiration from self-training UDA strategies, the SPADA framework is composed of three main components. First, a student network f_θ is trained on a combination of ground truth scribbles, dense pseudo-labels, and point-wise annotations. Second, a teacher network g_ϕ is obtained as an exponential moving average (EMA) of the student model, tasked

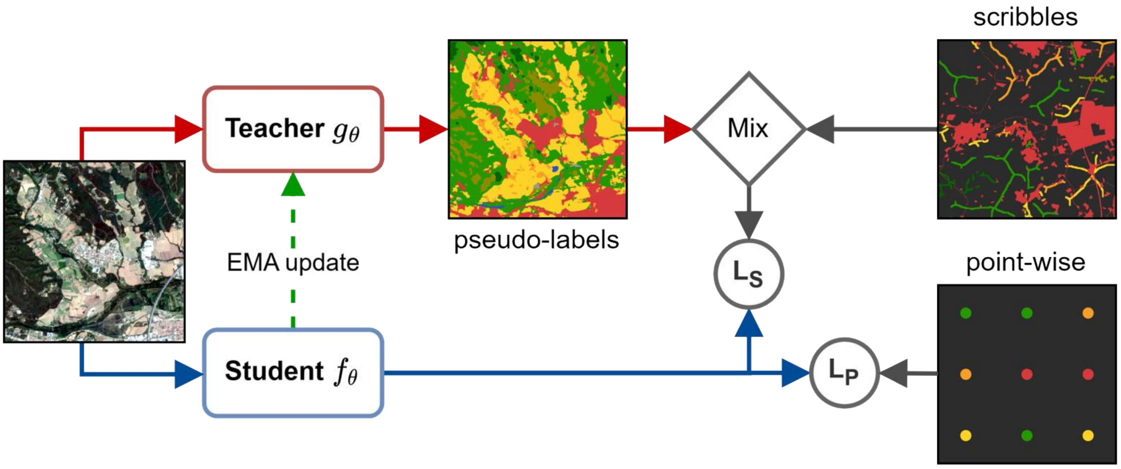


Figure 5.8: Visual representation of the SPADA framework. (i) the teacher model g_θ generates the pseudo-labels, (ii) the sparse ground truth is mixed with the pseudo-labels to obtain a dense mask to learn from, (iii) the total loss is computed as sum of L_S and L_P, (iv) the teacher weights are updated with the student weights through EMA.

with generating robust and consistent pseudo-labels from the target inputs. Last, a label mixing strategy comparable to ClassMix is employed to combine the scribble annotations with the generated pseudo-labels, leveraging the strengths of both types of annotations.

The training process of SPADA, as illustrated in Fig. 5.8, involves several key steps. First, the teacher network g_ϕ generates pseudo-labels from the target domain inputs, providing an initial set of annotations for the unlabeled target data. Following UDA standards [223, 96], these pseudo-labels are then filtered based on a fixed confidence threshold to obtain a set of reliable pseudo-labels, ensuring that only high-quality annotations are used in the subsequent steps. Next, the scribble annotations, which provide sparse but accurate labels, are fused with the filtered pseudo-labels to obtain the mixed labels y_M. This fusion process is formally defined as a composition of the pseudo-labels $\hat{y}_T$ and the scribble annotations y_S, represented by the equation $y_M = \hat{y}_T \odot y_S$. In contrast with UDA techniques where the source labels are randomly chosen with varying probability, here we always select all the available labels due to their sparsity. By combining the two types of annotations, SPADA leverages both the dense coverage of the pseudo-labels and the enhanced precision of the scribble annotations. The student network f_θ is then trained using a combination of the mixed labels y_M and the more reliable point-wise annotations y_P, which serve as additional regularization. The training objective is defined as a weighted sum of two categorical cross-entropy losses: $L_S = L_S(\hat{y}, y_M) + \lambda L_P(\hat{y}, y_P)$.

Here, $\hat{y}$ represents the predicted label, λ is a weighting factor that balances the contribution of the two losses, and L_S and L_P are the losses computed for the mixed labels and point-wise annotations, respectively. This training objective ensures that the student network learns from both the dense mixed labels and the precise point-wise annotations. Finally, to maintain stability and consistency in the pseudo-label generation process, the teacher network g_ϕ is updated using an Exponential Moving Average (EMA) of the student network's weights. This EMA update allows the teacher network to gradually incorporate the knowledge learned by the student network while mitigating the impact of noisy or inconsistent pseudo-labels. By iterating through these steps, SPADA progressively refines the pseudo-labels and adapts the student network to the target domain, effectively leveraging the sparse annotations to guide the domain adaptation process.

5.3.4 Experiments

We evaluate our approach on four distinct test areas, distributed across South Europe, as shown in Fig. 5.6, namely Catalonia in Spain, Liguria, and Sardinia in Italy, and Macedonia in Greece. In the following paragraphs, we provide the implementation details and experimental configurations tested, as well as the results obtained on the available test data.

Implementation Details

As introduced in Section 5.3.2, we utilize the Sentinel-2 L2A product as input for every experiment, obtained as cloudless mosaic from spring 2018, effectively aligning with the available ground truth Copernicus datasets employed in this study (i.e., CLC, UA, LUCAS, HRL). This temporal alignment also enables a fair comparison with state-of-the-art products like the Sentinel-2 Global Land Cover (S2GLC) [142]. S2GLC is itself an automated land cover classification using a combination of complex data processing, with rule-based and Random Forest models at its core. It was obtained through a fully automated workflow using multi-temporal Sentinel-2 imagery, with an overall accuracy of 86.1% for the pan-European land cover map. For fairness, we acknowledge that the most recent S2GLC map dates back to 2017, while our solution and baselines are trained and compared against 2018 data sources. At the same time, this third-party product benefits from the use of a time series of Sentinel-2 images and fully supervised training to enhance its accuracy, while our approach exploits sparse annotations from a single image. These factors certainly reduce the performance on categories such as the vegetation types, where the seasonality is a key factor to consider. Nevertheless, we adopt S2GLC as our current state of the art for lack of better alternatives in terms of taxonomy and geographical extension. The comparison against S2GLC highlights the robustness of our approach in handling sparse annotations and leveraging domain adaptation techniques to achieve competitive performance in land cover segmentation tasks.

Among the 12 available regions, the selected 8 training areas are further partitioned into training and validation sections, each measuring 2,048 × 2,048 pixels, with an 80%/20% split. Subsequently, all data is further tiled into 512 × 512 overlapping chips, resulting in a final set of 20,398 tiles for training, 5,100 for validation, and 394 full sections for testing, which are divided into tiles at runtime.

We compare SPADA against various semantic segmentation baselines, specifically UNet [197], OCRNet [262], PSPNet [270], DeepLabV3+ [39] and third-party products, including S2GLC [142] and the original CLC. The baseline models are trained on the available ground truth without additional measures, while CLC and S2GLC are remapped to match the uniformed land cover classes. Due to the absence of dense, manually validated annotations, we assess the performance using the only two reliable ground truth sources at our disposal: LUCAS and UA. First, we use the LUCAS observations available in the test areas to assess the classification accuracy of the model, in terms of F1 score (Eq. (2.5)). Second, we exploit the finer segmentation labels provided by UA to focus the assessment on the segmentation quality, using the standard IoU metric (Eq. (2.2)). Each model undergoes training for 160,000 iterations. Similar to DAFormer, we use an AdamW optimizer with linear warm-up of 1,500 iterations, learning rate of 1×10^{-4}, and weight decay of 0.01. Considering the UDA-based components, we maintain the EMA update parameter $\alpha = 0.99$ and the pseudo-label threshold $\tau = 0.968$ [223]. Concerning the weighting factor λ, we employ a fixed value of 1 throughout our experiments. Preliminary tests with higher values (i.e., up to 10) highlighted that the loss contribution did not bring additional performance gains. However, the critical aspect was maintaining the point-wise loss at least on par with the scribble loss. For both comparison and practical deployment consideration reasons, we adopt a MiT B3 [255] backbone for both the SegFormer baselines and our SPADA framework. During training, we further augment the available chips with transformations such as horizontal and vertical flips, or affine transforms, and color jitter or Gaussian blur to obtain more diverse inputs.

Results

The results on the LUCAS and UA test sets are presented in Table 5.5 and Table 5.6 respectively. Here, we compare against each baseline mentioned above, and S2GLC as strongest competitor. Additionally, we include the performance of the raw CLC layers, the reference land cover product at European scale, against the two selected ground truths as a lower bound. While, on one hand, these additional evaluations provide context on the agreement between the employed sources, they also give further insights on the low accuracy of CLC labels as ground truth for fully supervised training.

Table 5.5 presents the F1 scores computed on the LUCAS dataset. In this case,

Method	Agric.	Grassl.	Broadl.	Conif.	Shrubs	Bare	Artif.	Water	mF1	Acc.
CLC	55.72	24.71	66.49	68.51	33.55	60.63	52.71	63.40	55.53	52.54
UNet	57.94	28.40	72.30	68.33	34.72	34.47	55.98	73.01	58.00	55.62
OCRNet	59.39	26.25	70.71	65.17	33.77	33.81	53.88	68.73	56.87	54.38
PSPNet	58.51	18.69	66.38	64.56	31.66	52.87	52.88	68.18	55.04	52.10
DeepLabV3+	60.64	17.30	67.56	63.62	32.25	55.45	54.39	67.06	55.84	53.42
SegFormer	75.71	22.62	69.91	69.47	34.75	59.30	54.43	68.58	62.19	60.72
S2GLC	69.39	36.06	75.22	70.60	34.15	61.62	55.71	**75.99**	64.07	62.14
SPADA (Ours)	**77.72**	**39.36**	**76.78**	**74.19**	**38.08**	**63.64**	**58.54**	70.49	**67.93**	**66.99**

Table 5.5: Results obtained on the LUCAS test set in terms of classwise F1 scores, macro-averaged F1 (mF1) and macro-averaged Accuracy (Acc.).

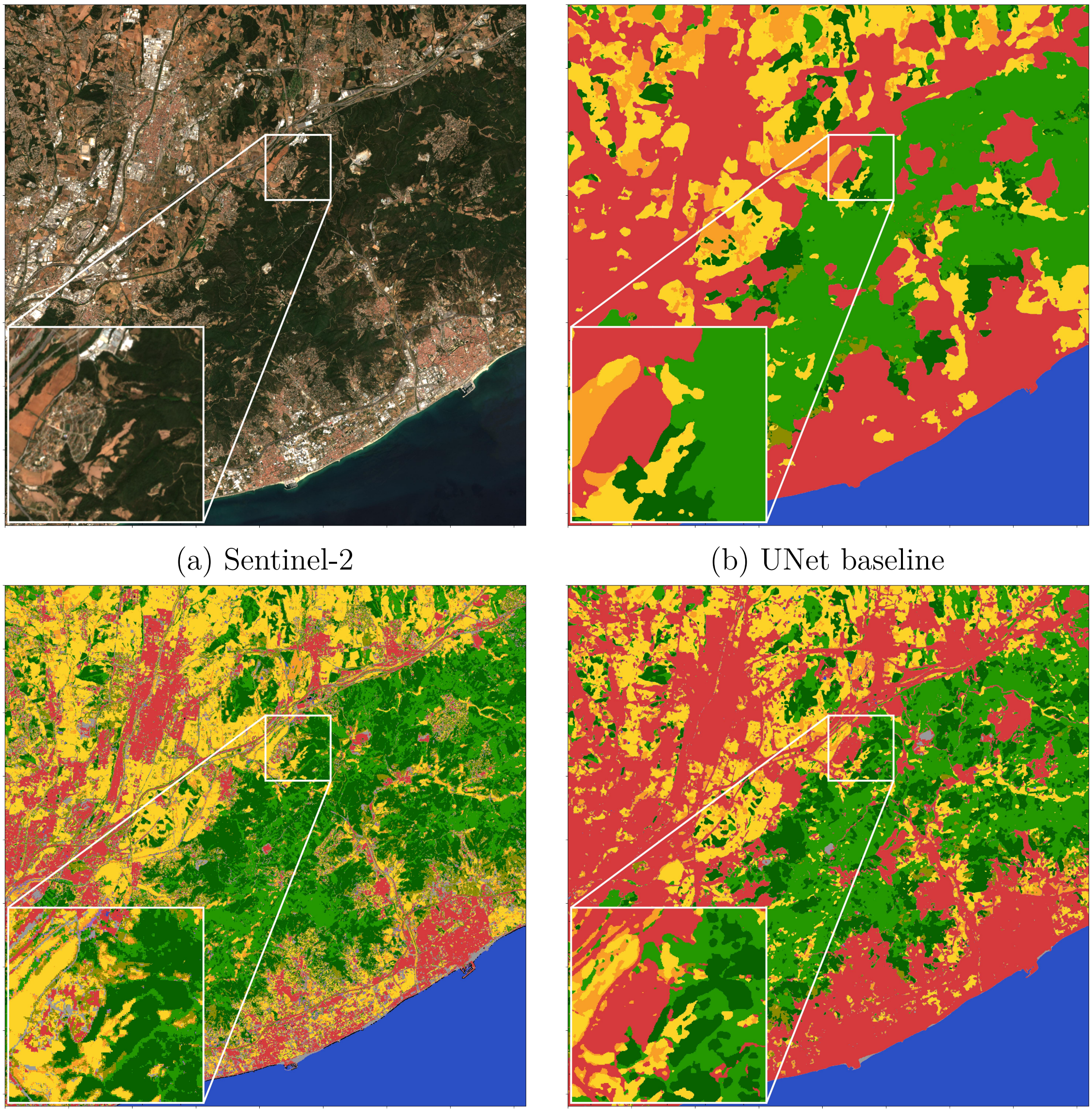
Method	Artif.	Bare	Wetl.	Water	Grassl.	Agric.	Forest	mIoU	Acc.
CLC	58.86	14.78	32.44	58.68	**14.62**	36.56	45.69	37.37	49.81
UNet	51.06	1.83	25.10	32.00	11.77	39.88	53.67	30.76	49.04
OCRNet	53.15	2.39	**34.36**	30.79	12.57	39.34	50.59	31.88	48.15
PSPNet	55.12	3.85	34.12	30.68	11.81	38.72	46.57	31.55	47.64
DeepLabV3+	54.09	5.60	30.06	31.25	10.74	41.31	48.45	31.64	47.55
SegFormer	56.77	**15.69**	23.01	55.41	7.19	52.60	49.89	37.22	56.06
S2GLC	40.55	7.12	4.00	**66.97**	14.19	46.09	**65.95**	34.98	52.49
SPADA (Ours)	**64.36**	13.56	27.27	65.95	10.01	**54.3**	64.56	**42.86**	**58.11**

Table 5.6: Results obtained on the Urban Atlas test set in terms of classwise IoU scores, macro-averaged IoU (mIoU) and macro-averaged Accuracy (Acc.).

given the sparse ground truth, we assign as prediction for each ground truth observation the majority class in a window $N \times N$ centered on the LUCAS point. We empirically set $N = 1$ (i.e., one-shot), as in practice bigger windows did improve on the final result. The UNet and SegFormer baselines achieve good results, with SegFormer being the stronger of the two. S2GLC proves to be a competitive baseline, proving the effectiveness of its fully supervised training. However, SPADA manages to outperform all tested baselines, including S2GLC, achieving a +3.86 increase in average F1 score. This improvement is highlighted also in the class-wise results, where SPADA consistently outperforms every other solution, except for the *water* category.

Notably, SPADA obtains high scores both on artificial surfaces, with a +2.83 points with respect to S2GLC, and more challenging vegetation categories such as *shrubs*, with a +3.93 increment. The raw CLC layers, as expected, have significantly lower accuracy compared to the other methods, highlighting the limitations of using coarse-resolution land cover products for detailed analysis.

Table 5.6 reports instead the IoU scores computed on the Urban Atlas dataset. In this case, given the similarities between these two products, the CLC baseline outperforms most of the baseline in terms of average IoU. S2GLC again proves to



(a) Sentinel-2

(b) UNet baseline

(c) S2GLC

(d) SPADA

Figure 5.9: Qualitative comparison of the outputs from (a) Sentinel-2 input, (b) UNet baseline, (c) S2GLC, and (d) SPADA. Best viewed zoomed in.

be a competitive baseline, demonstrating its ability to generate accurate segmentation maps. However, SPADA obtains comparable or substantially higher results on average for most classes, with an IoU increment of +7.88 compared to S2GLC. This significant improvement in IoU scores demonstrates the framework's ability to generate more precise and coherent segmentation maps, especially for classes such as *artificial surfaces*, *agricultural fields*. It is worth noting that the performance

Row ID	CLC	LUCAS	Mix	DAF. Head	F1 (L)	mIoU (UA)
1	D				58.63	36.24
2		✓			53.22	36.90
3	✓	✓			62.19	37.22
4	✓	✓	✓		65.13	40.44
5	✓	✓	✓	✓	**67.93**	**42.86**

Table 5.7: Ablation results with different combinations of datasets, strategies and architectures. *D* represents the dense version of the CLC ground truth (i.e., before scribble processing).

on some categories, such as *bare* and *grassland*, may be affected by the inherent challenges in distinguishing these classes from others with similar spectral characteristics, as well as classification errors in the original ground truth. However, the overall improvement across most categories highlights the effectiveness of our approach in capturing the subtle differences between land cover types.

Fig. 5.9 provides a qualitative comparison between the outputs of the Sentinel-2 input, a UNet baseline, which remains the best performing model among non-Transformer models, S2GLC, and SPADA. The UNet baseline (top right) produces uniform results, but the boundaries between different land cover types appear too approximate and lack precision. S2GLC, while visually appealing at first glance, exhibits significant noise and inconsistencies when zoomed in. This limitation is probably due to the pixel-based classification, that completely disregards contextual information. In contrast, SPADA demonstrates the ability to maintain a good level of detail for the boundaries while effectively reducing noise, thanks to the Transformer-based architecture and context-aware segmentation head. The maps generated by SPADA exhibit smoother and more coherent transitions between different land cover types, preserving the intricate patterns.

The ablation study, presented in Table 5.7, provides insights into the contributions of each component in our framework. In the first row, we observe that training with the dense version of the CLC ground truth (denoted as D) leads to suboptimal performance, with an F1 score of 58.63 on LUCAS. This highlights the importance of using sparse annotations, as the dense CLC labels may introduce noise and imprecise boundaries. When using only the LUCAS dataset for training (row 2), the performance drops slightly compared to the dense CLC training, achieving an F1 score of 53.22. This suggests that while the LUCAS dataset provides valuable information, its sparsity may not be sufficient to carry out a segmentation task. Combining the sparse CLC labels with the LUCAS dataset (row 3) leads to a significant improvement, with an F1 of 62.19. This suggests that the sparse CLC labels provide a broader context, while the LUCAS dataset offers more precise information on the underlying class. Introducing the DACS mixing strategy (row 4) further enhances the performance, achieving an F1 score of 65.13. Finally, replacing the simple SegFormer head with the more advanced DAFormer head (row 5) yields the best results, with an F1 score of 67.93 and a mean IoU of 42.86.

5.4 Learning from Multiple Tasks — Burned Area Mapping

Multitask learning is emerging as a powerful tool to improve the robustness and generalization capabilities of deep learning models [268]. By training models to simultaneously solve multiple related tasks, this paradigm allows them to capture more comprehensive and transferable features that can better adapt to variations in input data. In the context of domain robustness, multitask learning can provide several key benefits. In this particular case, the shared representations learned through multitask learning can act as indirect regularization, preventing overfitting to individual tasks or domains and promoting better generalization.

In this section, we address the problem of burned area delineation through a multitask learning approach. Wildfires have become an increasingly pressing issue in recent times, with their growing frequency and intensity posing significant threats to the environment, wildlife, and population. To effectively manage and respond to these events, it is crucial to accurately map the extent of the affected areas. However, due to the many challenges of this specific domain, traditional approaches based on binary segmentation often fall short in terms of robustness and generalization, especially when trained from scratch. In fact, burned area delineation inherently presents challenges of limited labels and domain shifts. The scarcity of large-scale, diverse datasets for this specific task, as mentioned earlier, naturally creates a low-resource scenario. Moreover, the varying landscapes and geographical locations where wildfires occur introduce significant domain shifts, making it difficult for models to generalize across different environments. Existing datasets[51, 184], often have limitations in terms of geographic coverage or variability, which can negatively affect the performance of models when applied to different scenarios. Moreover, the inherent bias and class imbalance in this task, with wildfires typically happening in forested areas and cover a relatively small portion of the input image, further exacerbate this problem and hinders the model generalization abilities.

To overcome these limitations, we propose a two-fold approach. First, we construct a dataset specifically tailored for burned area delineation, leveraging information from the Copernicus Emergency Management System (EMS), Sentinel-2 satellite imagery, and the ESA WorldCover 2020 land cover on global scale [264]. Our dataset encompasses a comprehensive set of samples, focusing primarily on European landscapes, and includes annotations for both burned area delineation and a more general purpose land cover segmentation tasks. By incorporating this last information as an auxiliary learning objective, we aim to improve the contextual understanding of the models and enhance their robustness in handling the burned area delineation task, which remains our primary target.

Second, we develop a multitask learning framework for Robust Burned Area Delineation (RoBAD) that integrates the land cover classification task alongside the primary objective. By training the model to simultaneously predict both the extent of the burned regions and the underlying land cover categories, we enable it to learn shared representations and exploit the complementary nature of these tasks. This approach not only enriches the semantic understanding of the models but also helps to mitigate the challenges posed by limited data and class imbalance.

To evaluate the effectiveness of our multitask learning framework, we conduct extensive experiments with state-of-the-art segmentation models, including UPerNet [254] concerning convolutional-based networks, and SegFormer [255] for Transformer-based solutions. We compare the performance of these models in both single-task and multitask settings, considering various configurations such as training from scratch and leveraging pretrained weights. Our results demonstrate the superior performance and robustness of the multitask approach, highlighting its potential to enhance the accuracy and generalization capabilities even in the presence of pretrained backbones. The code and dataset associated with this work are publicly available at `https://github.com/links-ads/burned-area-segmentation`.

5.4.1 Related Works

The majority of current semantic segmentation approaches rely on convolutional encoder-decoder architectures (CNNs), which employ various strategies to capture both global context and fine-grained details within the scene. Works such as Fully Convolutional Networks (FCN) [135] and U-Net [197] introduced the concept of using bottleneck components to encode pixel information into semantically meaningful representations, combined with skip connections to integrate lower-level features with higher resolution. Subsequent approaches, including DeepLab [39] and PSPNet [270], expanded upon these ideas by incorporating multiscale feature extraction and fusion techniques, where inputs are processed using varying kernel sizes and dilations to simultaneously capture local and global context. More recent architectures often combine these concepts to generate more robust and informative features [254]. However, the application of semantic segmentation to aerial and remote sensing imagery presents several unique challenges that often require domain-specific solutions. Unlike traditional image segmentation tasks, satellite data usually provides additional bands beyond the visible spectrum, such as Near-Infrared (NIR) or Shortwave Infrared (SWIR), which can be integrated in various ways, such as additional input channels [177] or through the use of specialized encoders for feature fusion [228]. Moreover, aerial images are typically denser, containing numerous entities against complex backgrounds and exhibiting wider spatial relationships. To effectively capture these long-range dependencies, attention mechanisms have been widely adopted to model pixel-level similarities across larger distances [164]. Consequently, transformer-based architectures and their segmentation variants [255] have become increasingly popular in this domain, leveraging their inherent ability to extract long-range relationships. In the specific context of burned area delineation, a wide range of techniques have been proposed to accurately map the extent of affected regions from remote sensing data. Traditional approaches often rely on spectral indices to distinguish between burned and unburned areas by combining information from multiple spectral bands. The most widely used indices include the Normalized Burn Ratio (NBR) [50] and the differenced Normalized Burn Ratio (dNBR) [153], which are frequently employed in conjunction with other indices, such as the NDVI, that provides information about the underlying vegetation [249]. Additionally, specialized indices have been developed to better adapt to specific satellite sensors, such as the Burned Area Index for Sentinel-2 (BAIS) [76]. While index-based methods are still widely used to this day, they often produce noisy results when not applied correctly and require further manual refinement. Moreover, certain indices, such as dNBR, necessitate the availability of pre-wildfire images to compare the same regions before and after the event, limiting their applicability in certain scenarios. In the last decades, machine learning and deep learning techniques have shown remarkable results on this task, reducing the need for manual intervention while achieving good performance. Standard supervised classification algorithms, such as Support Vector Machines (SVM) and Random Forests (RF), have been extensively utilized for burned area mapping [191, 114]. These approaches operate on a per-pixel basis and remain effective on lower-resolution imagery, such as MODIS. However, their lack of contextual information may lead to suboptimal results when applied to higher-resolution data, such as Sentinel-2 [114]. More recently, convolutional neural networks (CNNs) have been extensively applied to burned area mapping with success, particularly when considering post-wildfire images alone. U-Net-based segmentation architectures [160, 114] have become the standard approach in this domain. Furthermore, transformer-based architectures have demonstrated their effectiveness in various remote sensing tasks [218], including burned area segmentation [27]. These architectures leverage self-attention mechanisms to capture long-range dependencies and global context, which are crucial for segmentation in complex remote sensing scenes. Despite the progress made in burned area delineation using semantic segmentation techniques, the limited availability of large-scale, diverse datasets specifically tailored for this task remains a significant challenge. Existing datasets often suffer from limited geographic coverage or lack variability in terms of land cover types and characteristics [51, 184]. This scarcity of comprehensive training data can hinder the generalization capabilities of segmentation models and their ability to adapt to different scenarios encountered in real-world applications. Among other tools, Multi-Task Learning (MTL) has emerged as a promising approach that aims to leverage commonalities and differences among related tasks to improve generalization performance on all the tasks [231, 201]. Several works have explored MTL for dense prediction tasks like segmentation or depth estimation, in applications such as autonomous driving [228], or even remote sensing [120, 246, 86]. Architecturally, most MTL approaches for segmentation employ some form of feature sharing. Encoder-focused architectures typically share backbones across tasks and use task-specific decoders or heads to generate the output [156, 144, 109]. This allows for efficient parameter sharing while still providing flexibility to adapt to each task, however the gain from the shared learning remains limited. On the other hand, decoder-focused architectures introduce information sharing at multiple scales within the decoder [256, 269, 230]. In terms of optimization, a key challenge in MTL is balancing the influence of each task during training. Uncertainty weighting [109] and gradient normalization [43] are popular strategies to adaptively adjust task weightings based on their homoscedastic uncertainty or gradient magnitudes. Other approaches formulate MTL as a multi-objective optimization problem [206] or explicitly learn the task weightings and branching structure [87]. In this work, we adopt a decoder-focused MTL architecture, exploiting the same feature maps to carry out both the burned area delineation as a primary task and an auxiliary land cover mapping. Given the complementary nature of these objectives (i.e., burned areas effectively change the underlying land cover) and our focus on burned area mapping, we simply combine their losses with a weighted sum, without introducing further complexity.

5.4.2 Dataset

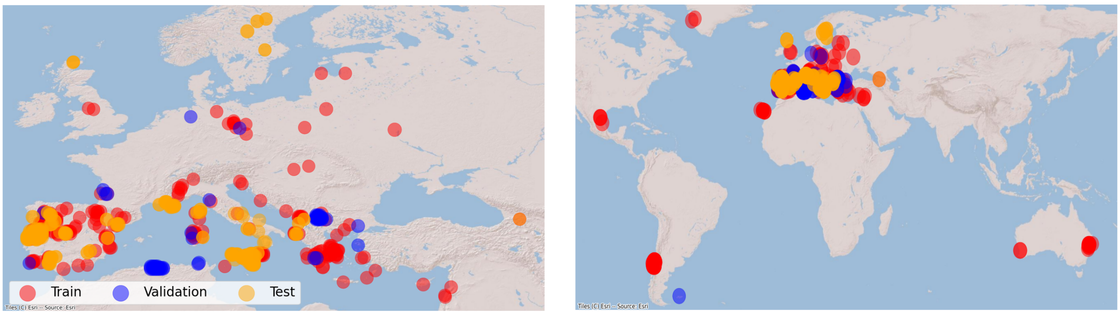


Figure 5.10: Geographic coverage of wildfire events in our dataset, categorized into training (red), validation (blue), and testing (yellow) subsets, focused on Europe (left), and displayed on a global scale (right).

To first tackle the problem at its root, we construct a comprehensive dataset specifically tailored to burned area delineation, with additional land cover information, attempting to address the limitations of existing resources [51, 184, 26]. The core of our dataset consists of a collection of 171 wildfire events, sourced from the Copernicus Emergency Management Service (EMS) [64], providing valuable information and resources for disaster management and emergency response. Within the EMS, the Rapid Mapping module plays a vital role by offering a curated set of Areas of Interest (AoI) associated with each event. These AoIs have undergone a precise and manual analysis by a team of experts, resulting in the generation of precise burned area delineations for the affected regions. To ensure a comprehensive representation of each wildfire event, we gather and harmonize data from multiple sources. For each AoI, we retrieve the corresponding Sentinel-2 satellite imagery, which serves as the primary input for our deep learning models. Sentinel-2 captures high-resolution multispectral data across 12 spectral bands, with spatial resolutions ranging from 10 to 60 meters. In this study, we focus on the Level-2A (L2A) product, which provides Bottom-of-Atmosphere (BoA) reflectance values, derived through atmospheric correction. In addition to the Sentinel-2 imagery, we incorporate the EMS burned area delineation maps, to serve as the ground truth labels for the primary task. These maps are generated by field experts using a combination of manual interpretation and semi-automated techniques, ensuring a high level of accuracy and reliability[1]. The EMS provides three different incremental products, which provide a gradually refined map of the event, in vector format. First Estimate Products (FEP) are produced as quickly and as soon as possible therefore, as the name suggests, the resulting delineation usually provides approximate boundaries. On the other hand, the Delineation (DEL) and Grading (GRA) products offer a much more refined output, comprising the precise borders of the event for the former case, and an additional subdivision by damage estimate for the latter. For these reasons, we keep and rasterize the best product available, starting from the grading as primary source, and falling back to delineation first, and FEP after, if no other option is available. To enable multitask learning, we integrate land cover information from the ESA WorldCover dataset [264], a global land cover map that provides detailed information on land cover classes at a spatial resolution of 10 meters. It covers the entire Earth's landmass and offers a comprehensive set of 11 land cover classes, including natural vegetation (i.e., tree cover, shrubland, grassland), cultivated areas, built-up areas, bare soil, and water bodies. Through this additional data, we aim to provide models with additional context and enable them to learn the relationships between burned areas and different land cover types, as well as mitigate recurring errors such as dark water pixels classified as burned.

We retrieve Sentinel data through the SentinelHub services [212], which provide efficient access and processing capabilities. Considering the input requirements of the deep learning models, we impose a minimum dimension of 512 pixels on each side of the image, expanding the region until the requirement is satisfied for every AoI for areas that do not meet this criterion. Conversely, we divide larger areas into multiple subsections with a maximum size of 2,500 × 2,500 pixels to facilitate practical usage and computational processing costs. In order to obtain

[1] https://emergency.copernicus.eu/mapping/ems/quality-control

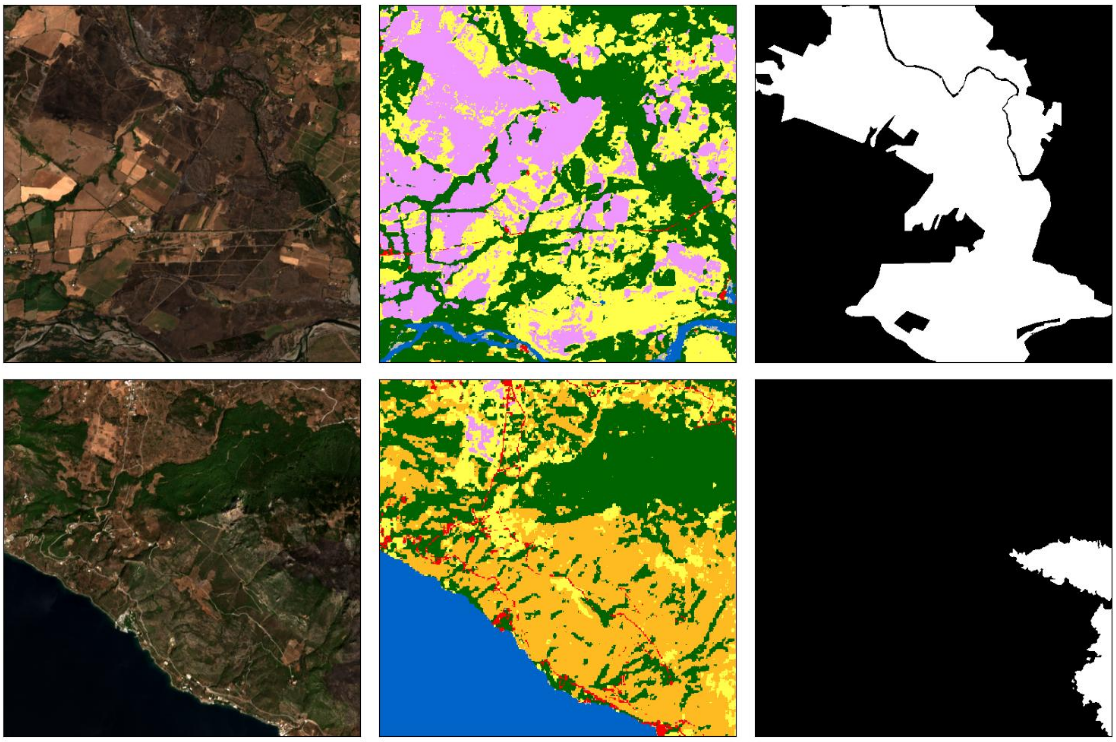


Figure 5.11: Dataset samples, each consisting of a Sentinel-2 satellite image (left), its corresponding land cover map (middle), and the associated burned area delineation (right).

the clearest possible images, minimizing the presence of smoke or large cloud cover, we consider a time frame of up to 30 days following the reported event date for the catalog queries. Within this window, we select the Sentinel-2 acquisition with the least cloud coverage, ensuring the highest quality input data for our models. Despite these precautions, the presence of clouds in the final image samples cannot be completely avoided. To address this, we employ a cloud segmentation model derived from CloudSen12 [14] to generate a validity map that identifies cloudy pixels in the imagery. During the training process, we apply this additional mask to exclude cloud-covered pixels from the loss computation, preventing them from affecting the learning process. We retrieve instead the corresponding raster layers for ESA WorldCover through the Microsoft Planetary Computer platform [214]. The land cover maps undergo minimal processing, primarily involving a direct remapping of the original ESA taxonomy to a contiguous list of categories indexed from 0. Pixels that lack a specific land cover category are assigned a value of 255 to distinguish them from valid classes. To ensure that all the layers in our dataset are aligned, we resample each band and rasterize vector layers at a resolution of 10m/pixel, matching both Sentinel-2 most resolved bands, and ESA WorldCover.

An example of the final results obtained after this process is visible in Fig. 5.11.

Fig. 5.10 illustrates the geographic distribution of the wildfire events included in our dataset, categorized into training (red), validation (blue), and testing (yellow) subsets. The map on the left depicts the global distribution of events, while the map on the right focuses specifically on Europe, where the majority of the events are concentrated. This visual representation highlights the diversity and comprehensive coverage of our dataset, encompassing wildfire incidents across different regions and continents. The resulting dataset comprises a total of 433 samples with varied sizes, spanning from 2017 to the early months of 2023. As previously mentioned, the events are primarily concentrated in Europe, with additional samples from Australia and the American continent. It is worth noting that our dataset effectively extends the coverage of previous EMS-based sources [51]. To facilitate direct comparisons with prior work and assess the generalization and performance of our proposed approach, we maintain all the activations present in previous sources as a testing subset, while the remaining events are utilized for training and validation purposes.

5.4.3 Method

In this work, we aim to develop a MTL framework to improve burned area delineation through the help of a more generic land cover segmentation task at training time only. The method applied here is quite straightforward, foreseeing a single encoder and decoder, with the addition of two separate shallow linear layers to map the feature representations into either burned or unburned on one side, and a specific land category on the other. In the following paragraphs, we formalize the problem at hand, and we provide details on the underlying architecture.

Problem Statement

We model the problems of burned area and land cover mapping as two separate tasks, a binary segmentation and a multi-class segmentation objective, respectively. Our dataset provides a delineation label (y_D) and a land cover label (y_{LC}) as ground truth for each sample. The proposed framework consists of a single encoder and a single decoder, with two classification heads: h_D for burned area delineation and h_{LC} for land cover classification. The primary objective is to train the model f_θ, parameterized by θ, to accurately predict burned area delineations ($\hat{y}D$) while simultaneously learning from the auxiliary task of land cover classification ($\hat{y}LC$). The shared representations ϕ_θ, learned by the decoder, enable the model to capture common patterns and features relevant to both tasks, potentially improving the overall segmentation performance.

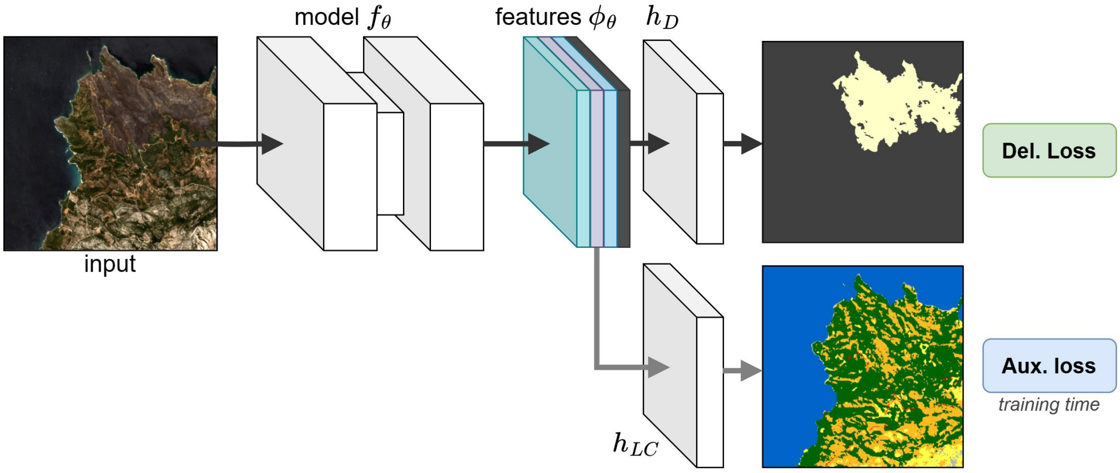


Figure 5.12: Visualization of the RoBAD multitask learning framework, where the model f_θ learns shared features ϕ_θ from the decoder, which are used by both the primary head h_D for burned area delineation and the auxiliary head h_{LC} for land cover classification during joint training. At test time, the auxiliary head is dropped, and only the primary head is used for the final burned area prediction

RoBAD Framework

Fig. 5.12 illustrates RoBAD, our proposed MTL framework for burned area delineation. The core idea is to train the full model f_θ by simultaneously predicting burned area delineation and land cover segmentation using the shared representations ϕ_θ learned by the decoder. By learning a single set of representations from both tasks, the model can identify and exploit informative features and patterns that generalize across the burned area delineation and land cover classification objectives. These shared features serve as a common foundation, allowing the model to develop a more comprehensive view of the input data and potentially leading to an improved segmentation. Throughout the training phase, we optimize the model using a standard cross-entropy loss for each task, employing the binary variant for burned area delineation, while utilizing the multi-class formulation for land cover classification. At training time, we combine these losses with a weighted sum, and we derive the gradients in a single backward pass, optimizing the network parameters in a single step. During inference, we concentrate exclusively on the burned area delineation task by removing the auxiliary head h_{LC} and applying standard binary segmentation using the learned features from the decoder. To investigate the impact of different architectural designs on the burned area delineation task using this framework, we consider three different segmentation models: we adopt a standard UPerNet decoder [254] with a Residual Network (ResNet) encoder, UPerNet with a Vision Transformer (ViT) backbone, and SegFormer as Transformer-only solution [255]. The UPerNet decoder stands out as a versatile and robust architecture, employing a unified perceptual parsing structure that accommodates both traditional convolutional neural networks (CNNs) and state-of-the-art transformer-based approaches as encoders. This architectural flexibility allows us to perform an unbiased comparison between CNN-based and transformer-based backbones within a consistent framework. On the other hand, SegFormer represents an alternative end-to-end solution that has demonstrated strong performance on various aerial image segmentation tasks, including burned area delineation [27, 218]. SegFormer introduces a novel hierarchical transformer encoder that captures long-range dependencies and multiscale features, coupled with a lightweight MLP decoder. By comparing the performance of these three model families within our MTL configuration, we aim to provide a comprehensive evaluation of different architectural designs and the effectiveness of the RoBAD framework in this setting.

5.4.4 Experiments

In this work, we test the three combinations of models in two different settings, considering a training with and without pretrained backbones, evaluating them on the activations dedicated for testing. In the following paragraphs, we provide additional details on the implementation and the experimental configurations, as well as a discussion on the obtained results.

Implementation Details

We split the available activations from our dataset for training and testing, leaving out those that are already present in existing EMS burned area datasets [51]. These remaining activations are reserved for testing purposes. This allows us to establish a fair comparison with previous works and assess the generalization capabilities of our proposed approach. We further allocate 10% of the training activations for validation purposes, resulting in a total of 129 wildfire events for training, 15 for validation, and 27 for testing. To accommodate the variability in image dimensions across the dataset, we employ different sampling strategies during training and evaluation. During the training phase, we adopt a standard random sampling approach that extracts square crops of 512×512 pixels from random locations within the input images. In our experiments, we do not employ any specific weighting or guiding techniques to prioritize the sampling of burned area regions during training. Instead, we aim to assess the model's ability to generalize and leverage the complementary information provided by the dense land cover labels. For validation and testing, we adopt a standard sequential sampling strategy with a sliding window with overlap to ensure comprehensive coverage of the entire input image. The predictions for the overlapping tiles are seamlessly merged using a smooth blending technique based on splines, allowing for the reconstruction of the original input dimensions. We conduct two sets of experiments to evaluate the effectiveness of our multitask learning framework. In the first set, we focus solely on the task of burned area delineation, training the models in a single-task setting. The second set of experiments involves multitask training, where we simultaneously learn to predict both burned area delineation and land cover segmentation. In the multitask setting, we apply an additional masking operation to the land cover annotations, excluding the burned pixels, to avoid any inconsistencies in the labels. To provide a comprehensive evaluation, we consider three state-of-the-art segmentation architectures: UPerNet [254] with two widely used encoders, namely ResNet-50 and ViT-S, and SegFormer [255] with the MiT-B3 encoder, thus obtaining a CNN-based, hybrid, and Transformer-based architectures. These models have demonstrated strong performance in various semantic segmentation tasks and offer a good balance between accuracy and computational efficiency. In addition to the architectural choices, we investigate the impact of using pretrained weights for the backbone networks in both single-task and multitask settings. For UPerNet with ResNet-50 and ViT-S encoders, we utilize weights pretrained on the large-scale SSL4EO-S12 dataset [245], which has shown to be effective for remote sensing tasks. However, due to the lack of suitable pretrained weights for SegFormer, we resort to using weights pretrained on the ImageNet dataset [255]. Our implementation is based on the *mmsegmentation* library [158], which provides a modular and extensible framework for semantic segmentation. We adapt the library to support the additional input channels required for our multispectral satellite imagery and make the necessary modifications to accommodate the multitask learning setup. For all experiments, we train the models on a single NVIDIA A100 GPU for 30 epochs, using a batch size of 32 tiles. We employ the AdamW optimizer with a learning rate of 1e-4 and utilize a standard cross-entropy loss function as in Eq. (5.1) for both tasks. These hyperparameters align with settings commonly used in similar works [27]. To assess the performance of our models, we adopt two widely used evaluation metrics, specifically binary F1 score as in Eq. (2.5) and binary Intersection over Union (IoU) Eq. (2.2).

Results

We conduct experiments on single-task learning (STL) and multitask learning (MTL) approaches using both pretrained and non-pretrained weights for the backbone networks, and for each configuration, we perform three separate runs with different random seeds and average the results. Table 5.8 summarizes the experimental results, reporting the average scores and standard deviations for each model and setting. Focusing on the experiments conducted without pretrained weights, we observe that the MTL setting consistently achieves higher scores and exhibits much lower standard deviation compared to the single-task setting. Except for SegFormer, which demonstrates the good scores in both STL and MTL variants,

		From scratch		Pretrained	
Setting	Model	F1	IoU	F1	IoU
STL	SegFormer (B3)	**89.01± 1.39**	**80.22± 2.25**	90.79± 0.46	83.13± 0.78
	UPerNet (RN50)	82.33± 9.17	70.94± 12.63	**91.27± 0.08**	**83.95± 0.13**
	UPerNet (ViT-S)	87.65± 2.01	78.08± 3.17	89.20± 1.29	80.53± 2.09
MTL	SegFormer (B3)	**90.94± 0.17**	**83.38± 0.29**	90.91± 0.28	83.34± 0.47
	UPerNet (RN50)	89.82± 1.76	81.57± 2.87	**91.86± 0.30**	**84.94± 0.51**
	UPerNet (ViT-S)	89.76± 0.15	81.43± 0.25	90.69± 0.58	82.98± 0.97

Table 5.8: Experimental results comparing single-task (STL) and multitask (MTL) training, with models trained from scratch or using pretrained encoders. Results show that MTL generally improves performance, with pretrained models outperforming those trained from scratch. UPerNet (RN50) demonstrates the highest F1 and IoU scores when pretrained, while SegFormer (MiT-B3) performs best when trained from scratch.

Setting	Model	Time per Ep.	Param. (M)
STL	**SegFormer (B3)**	3h28m	44,6
	UPerNet (RN50)	3h20m	64,1
	UPerNet (ViT-S)	3h20m	57,9
MTL	**SegFormer (B3)**	3h40m (+12m)	44,6
	UPerNet (RN50)	3h50m (+30m)	64,1
	UPerNet (ViT-S)	3h30m (+10m)	57,9

Table 5.9: Analysis of the computational costs in terms of training time over one epoch, as average of three epochs, and total network parameters. While the training time increases by a small margin, the parameter increase is effectively negligible given the shared encoder-decoder structure.

the multitask approach yields a non-negligible improvement of +3.85 in terms of F1 score and +5.71 in terms of IoU on average, across all models. Moreover, the results obtained in the multitask configuration are considerably less variable, with a decrease in standard deviation of −3.51 for F1 score and −4.88 for IoU. This robustness is also evident in the qualitative results presented in Fig. 5.13, where the multitask models produce more reliable and accurate segmentation maps. As expected, when considering the experiments that utilize pretrained weights, the performance gap between STL and MTL settings becomes less pronounced. The pretrained models achieve higher and more stable scores even in the STL setup, which is in line with the effectiveness of large-scale pretraining demonstrated in various contexts [245]. Nevertheless, multitask training still provides an average overall improvement of +0.73 in F1 score and +1.21 in IoU, regardless of the underlying architecture. Comparing the results obtained with and without pretraining, we observe that the pretrained models consistently outperform their non-pretrained counterparts, particularly in the multitask configuration. When examining the top-performing models from both settings (i.e., SegFormer in the non-pretrained case and UPerNet-RN50 in the pretrained case), the single-task setting achieves an improvement of +2.24 in F1 score and +3.72 in IoU, while the multitask setting achieves an average improvement of +0.92 in F1 score and +1.56 in IoU. It is worth noting that, while SegFormer consistently outperforms UPerNet variants when training from scratch, in the pretrained settings it falls behind. We argue that this might depend on the suboptimal pretrained weights derived from ImageNet, while the SSL4EO weights greatly help.

In addition to the performance evaluation, we also provide an overview of the computational costs associated with the STL and MTL approaches. Table 5.9 presents a comparison of the training speed and memory requirements for both settings. Despite the inclusion of an additional segmentation head in the multitask models, we observe only a modest increase in training time compared to the STL counterparts, with a marginal difference of approximately 20 minutes per training epoch. While the multitask models do incur a slight increase in memory usage due to the additional parameters, this increment remains negligible and does not significantly impact the feasibility of implementation. It is worth noting that the multitask setting effectively adds only the parameters of a single pixel classification head, which can be realized as a 1×1 convolutional layer with $|\phi_\theta|$ feature channels as input and 11 categories as output, with a number of additional parameters by the thousands. Moreover, during inference, the auxiliary head is omitted, eliminating any computational overhead associated with the auxiliary task.

5.5 Learning from Large Vision Models — Automating the Annotation Process

In recent years, the field of machine learning has undergone a significant paradigm shift, moving away from the traditional approach of training models from scratch on specific downstream tasks. Instead, as briefly observed in Section 5.4, the focus has shifted towards large-scale pretraining and the development of vision foundation models, also known as Large Vision Models (LVMs) [239]. These models are trained on vast amounts of diverse data, often using self-supervised or weakly-supervised learning techniques, to capture general-purpose representations that can be adapted to various tasks using minimal finetuning, few-shot or even zero-shot learning [239]. The emergence of foundation models has revolutionized the way we approach machine learning problems, particularly in the domain of computer vision. Models such as CLIP [188], GroundingDINO [130], and Segment Anything (SAM) [112]

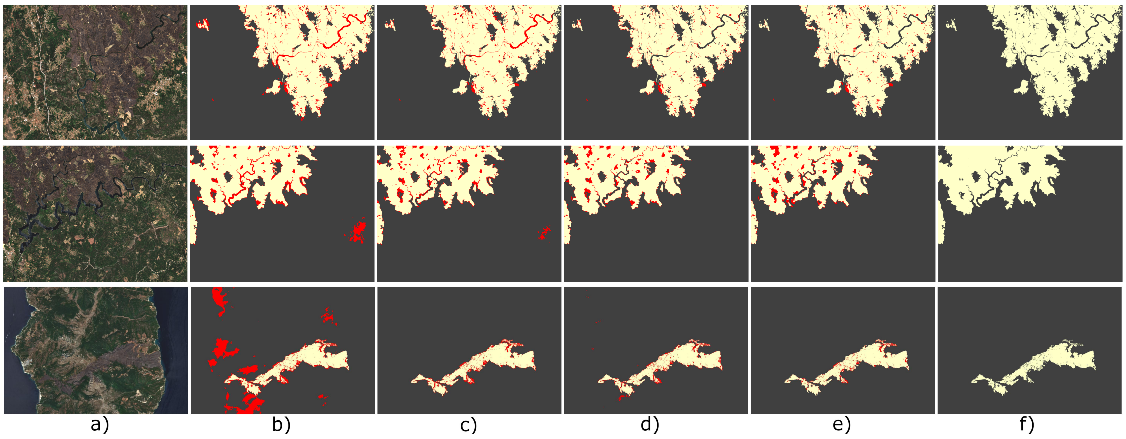


Figure 5.13: Qualitative samples of delineation results obtained from UPerNet with ResNet-50 (our best performing model). We compare the predictions of models trained from scratch in STL (b) and MTL (c) settings, and with pretrained weights, again with STL (d) and MTL (e) variants, with the ground truth (f). The red pixels in the prediction maps indicate misclassified areas compared to the ground truth delineation

have demonstrated remarkable performance across a wide range of tasks, including image classification, object detection, and segmentation. These models leverage the power of large-scale pretraining or billion-samples datasets [112] to learn rich and transferable representations that can be easily adapted to new domains and tasks with minimal additional training.

The paradigm shift towards foundation models has significant implications for various domains, including remote sensing. Remote sensing data, particularly Very-High Resolution (VHR) imagery, offers a huge amount of information content. However, the effective utilization of such data with supervised machine learning techniques often faces challenges due to the limited availability of high-quality annotations. In this context, LVMs have the potential to serve as few-shot learners or robust annotators, enabling the automatic generation of labels for large-scale remote sensing datasets. Despite their impressive performance on natural images, the application of foundation models to remote sensing data has just begun, with only a few studies exploring their potential at scale [235]. This work addresses this gap by proposing an automated pipeline, named FMARS (Foundation Model Annotations in Remote Sensing), that combines open data sources with foundation models to generate either instance or semantic segmentation labels, starting from robust box prompts. In this section, we first introduce the FMARS annotation pipeline, which leverages LVMs to automatically generate annotations for VHR remote sensing imagery. As an application example, we employ this pipeline to generate the FMARS dataset, a large-scale dataset with labels obtained from pre-event imagery over 19 disaster events worldwide, derived from the Maxar Open Data initiative [148]. To validate the effectiveness of the annotation approach, we train state-of-the-art models on the generated labels, employing UDA techniques [96] [97] to enhance stability over the final results. Finally, we discuss the implications of our findings and highlight the potential of leveraging LVMs in remote sensing applications.

5.5.1 Related Works

Computer vision has witnessed significant advancements in a very short span, especially with the advent of LVMs). These models, pre-trained on vast amounts of diverse data, have demonstrated remarkable adaptability to various downstream tasks with minimal to no fine-tuning. Notable examples include CLIP [188], which leverages natural language supervision to learn transferable visual representations, and self-supervised models like DINO, or DINOv2 [172], which capture rich visual features without relying on explicit annotations. Building upon these advancements, recent works have introduced models that excel in specific tasks, such as object detection and segmentation. GroundingDINO [130] combines the strengths of DINO with grounded pre-training to enable open-set object detection, while Segment Anything (SAM) [112] offers a flexible and promptable framework for segmenting objects in images. Despite the impressive capabilities of foundation models, their application to remote sensing data has been relatively limited compared to natural images. Remote sensing datasets, such as DOTA or its segmentation equivalent iSAID [252] and DIOR [266], provide valuable resources for object detection and visual grounding tasks, but their scale and scope remain modest compared to their natural image counterparts [112], as also highlighted in Table 5.11. Moreover, the unique characteristics of remote sensing imagery, such as varying resolutions, spectral bands, and viewpoints, pose additional challenges for adapting foundation models to this domain [104]. Nevertheless, recent works have begun assessing the potential of LVMs in the context of remote sensing. For instance, similar to our approach, SAMRS [235] leverages SAM to automatically generate annotations for well-known datasets like DOTA and DIOR, demonstrating the feasibility of scaling up remote sensing segmentation datasets using LVMs. Other works have assessed the applicability of foundation models to various remote sensing tasks, such as building extraction [173] and semantic segmentation [267]. In the context of disaster management, the xBD dataset [89] represents a notable effort to provide annotations for building damage assessment from satellite imagery. However, its focus is limited to the single building category, making it challenging to adapt to other scenarios. To address this limitation, our work aims to leverage the growing availability of Very-High Resolution (VHR) remote sensing imagery and the potential of foundation models as robust annotators to automatically generate labels for a broader range of objects of interest.

5.5.2 Method

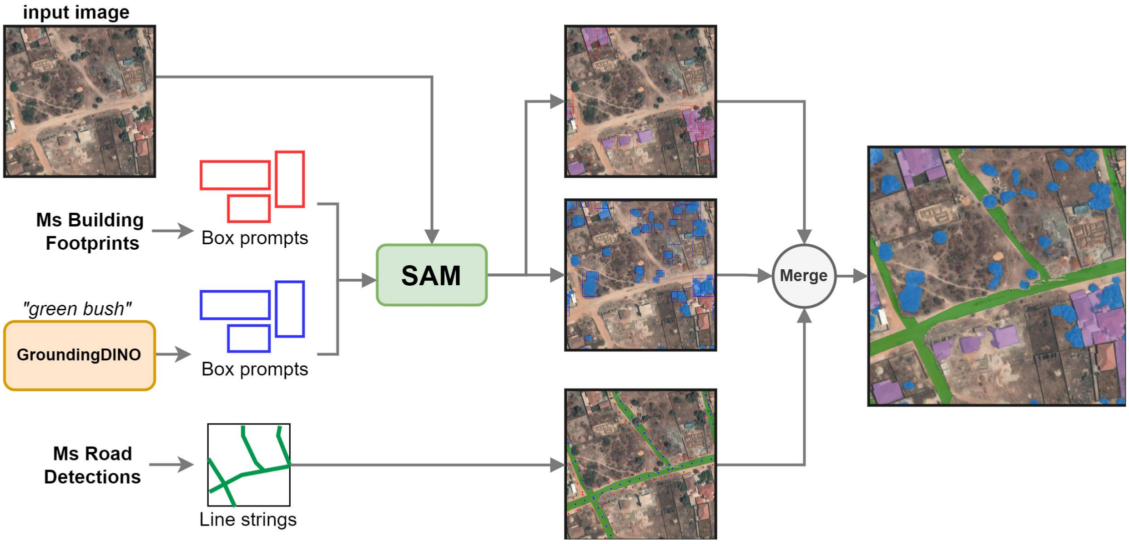


Figure 5.14: FMARS annotation workflow: for buildings and roads, prompts are generated from existing open datasets, while for high vegetation, GroundingDINO is employed to generate bounding box proposals. The prompts are then fed into SAM to obtain segmentation masks. The resulting masks are processed and merged to compose the ground truth.

In this work, we employ a combination of two state-of-the-art VFMs, namely Segment Anything (SAM) [173] and GroundingDINO [130], for the annotation process. SAM is a transformer-based segmentation network that has been trained using the so-called *promptable segmentation*. Unlike traditional segmentation objectives, SAM receives two inputs: an image and a prompt. The image is processed using a robust image encoder, while the prompt is embedded into the decoder using a prompt encoder and exploited as a query by a lightweight mask decoder to produce segmentation masks. To handle ambiguities, SAM can predict multiple outputs with associated confidence scores for the same inputs. The prompts can be highly flexible, ranging from sparse inputs such as points, bounding boxes, or text, to dense arrays like binary masks. Points and boxes are encoded as simple positional embeddings, text can be processed using off-the-shelf models like CLIP [188], and masks are combined with the encoded image using a series of convolutions and element-wise sums. In our annotation pipeline, we adopt box prompts for their robustness and flexibility, which aligns well with the available inputs, including open data sources for buildings and roads, and box object detections derived from GroundingDINO. GroundingDINO introduces cross-modal fusion between a text prompt and an image to provide open-set object detection capabilities. It utilizes BERT [58] as a text processor and a Swin Transformer [131] as an image encoder.

While the outputs may be approximate compared to human annotations, GroundingDINO offers substantial flexibility in generating bounding boxes for potentially any known object, given a text prompt. This allows us to obtain initial estimates for objects lacking ground truth, such as high vegetation, and leverage GroundingDINO as a prompt generator for the subsequent SAM masking phase [194].

Event name	**Year**	**Area (km^2)**	**Event name**	**Year**	**Area (km^2)**
Cyclone Mocha	2023	3,446.4	Morocco earthquake	2023	49,901.9
Italy (Emilia) flooding	2023	1,519.1	Canada (NWT) wildfires	2023	468.6
Gambia flooding	2022	391.2	Sudan flooding	2022	249.3
Hurricane Fiona	2022	1,341.8	Afghanistan earthquake	2022	4,180.6
Hurricane Ian	2022	30,743.2	Cyclone Emnati	2022	8,506.0
Hurricane Idalia	2023	12,156.4	Kentucky flooding	2022	1,641.6
India floods	2023	496.3	Pakistan flooding	2022	7,528.7
Indonesia earthquake	2022	1,011.3	Georgia landslide	2023	157.4
Turkey earthquake	2023	2,745.7	South Africa flooding	2022	559.7
Kalehe flooding	2022	89.9			

Table 5.10: List of events from Maxar Open Data included in the FMARS dataset, including the covered area (km^2) derived from VHR imagery, and the year it happened in.

Data Sources. Considering the importance of fine-grained segmentation in disaster management and damage assessment contexts, Very-High Resolution (VHR) imagery is essential, as lower-resolution satellite sources like Copernicus Sentinel-2 [65] do not provide sufficient image content to characterize objects of interest, such as buildings or roads. Currently, the largest source of disaster-related VHR imagery remains the Maxar Open Data Program [148]. This initiative offers pre- and post-event images from more than 100 major crisis events worldwide since 2017, covering a total surface area of over 2.6M km^2. We select a subset of resources containing RGB imagery, comprising 19 events spanning from 2022 to 2023, as shown in Table 5.10, summing up to an area of 127,134 km^2. As use case, inspired by current disaster management datasets [89], we focus our construction process on infrastructures, namely buildings and roads, which are often the primary focus in post-event damage assessment, and high vegetation, which typically occludes the underlying surface. Among open resources providing infrastructure information, we select Microsoft's Building Footprints and Road Detection datasets [151], which contain building footprint polygons and road graphs on a global scale generated by applying deep learning models to VHR satellite imagery. For buildings, we do not directly adopt them as ground truth labels, but rather exploit them as trustworthy yet approximate prompts for the SAM model. In the absence of a reliable source to derive high vegetation prompts, we employ GroundingDINO as our bounding box generator for this category [194].

Dataset	# Images	Image size	Res. (cm)	Bands	# Inst.	# Categ.	Area (km^2)
Vaihingen [199]	33	2,500 × 2,500	9	IRRG	None	6	1.33
Potsdam [199]	38	6,000 × 6,000	5	RGBIR	None	6	11.08
iSAID [252]	2,806	4000 × 4000	≥ 50	RGB	655,451	15	11,224
xBD [89]	9,168	1024 × 1024	≥ 50	RGB	$> 700,000$	4	45,000
SAMRS [235]	105,090	Mixed	≥ 50	RGB	$> 1.6M$	Mixed	Unknown
FMARS	6,896	17,408 × 17,408	≥ 30	RGB	$> 25M$	3	$> 125,000$

Table 5.11: Comparison of our FMARS dataset with well-known general purpose VHR datasets available in literature.

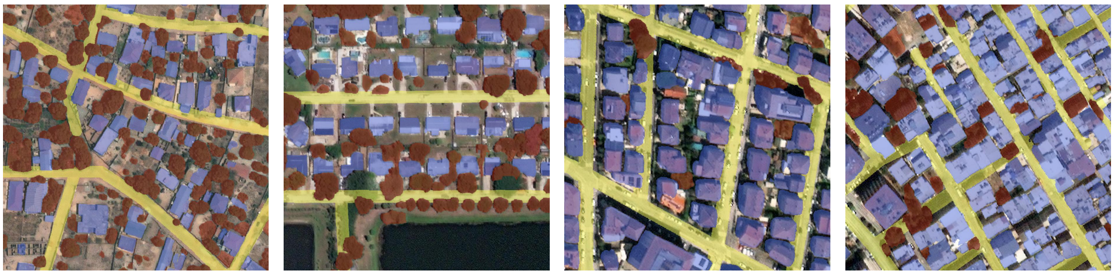


Figure 5.15: Four samples derived from the FMARS dataset in different regions. From left to right: Gambia, USA, Italy, Turkey. Despite the high quality results, it is possible to observe some omissions, especially concerning vegetation and tree crown.

Annotation Workflow. Our objective is to generate segmentation labels for three classes: buildings, roads, and high vegetation. While disaster risk management primarily focuses on damage assessment by comparing pre- and post-event images, we initially concentrate our efforts on delineating infrastructures on pre-event acquisitions only. This approach aligns with typical damage assessment frameworks, which first identify relevant entities in pre-event images and then use the post-event image to determine the sustained damage [208]. For buildings, we generate box prompts by extracting axis-aligned bounding boxes (AABB) from each footprint polygon. On the other hand, road graphs pose a challenge for prompt-based segmentation because their sparse lattice structure does not allow for fine contour generation. In this case, point-based prompts [112] did not yield satisfactory results. In this case, we opt to rasterize the available vector lines with an empirically predefined buffer radius of 5m. For vegetation, we derive boxes using GroundingDINO with simple text queries such as “green trees” or “bushes”. Interestingly, we observe better performance on trees with the latter prompts, likely due to the aerial viewpoint that occludes the tree trunk, which is uncommon in the context of the natural images on which GroundingDINO was trained on. To ensure a certain degree of confidence for the generated outputs, we apply a minimum box threshold of 0.12 and a text threshold of 0.3 [130]. We further filter out noisy outputs by applying non-maxima suppression (NMS) at 0.5, effectively removing boxes with an aspect ratio lower than 1:2, and a maximum area over 7000 m^2. Similar to buildings, we then use the generated boxes as prompts for SAM to extract segmentation labels. Finally, we store the resulting delineation and its class as a single vector polygon to allow for both instance and semantic segmentation tasks. The final FMARS dataset comprises a total of 25,547,189 individual labels across the three target classes. Specifically, it includes 15,104,779 building footprints, 7,868,086 tree instances, and 2,574,324 road segments. While FMARS may cover a smaller geographical area compared to some existing large-scale datasets such as DOTA [252], it provides a higher density of annotations within its scope, as detailed in Table 5.11 and by the visual samples in Fig. 5.15, offering a rich and detailed representation of the covered regions. These labels are stored in GeoParquet format, with each entity represented by its own polygon. We note that, in cluttered environments, the instance-wise separation may not be perfect due to the limitations of the automated annotation process.

5.5.3 Experiments

As plausible application scenario, we carry out our experiments with the obtained FMARS dataset using state-of-the-art semantic segmentation architectures. Our primary objective is to evaluate the knowledge transfer ability from the automatically generated labels to smaller, more deployable models. In the following paragraphs, we describe the experimental setup, as well as the obtained results.

Implementation Details

Using the dataset annotated with our FMARS pipeline, we can train standard semantic segmentation models to evaluate the knowledge transfer ability to more manageable models. In our experiments, we adopt state-of-the-art solutions based on SegFormer [255]. To counteract the inherent inaccuracies in fully automated labeling and the consequently lower recall for categories such as high vegetation, we apply UDA techniques for improved stability during training, as described in Section 5.2. Specifically, we adopt current state-of-the-art frameworks, namely DAFormer [96] and Masked Image Consistency (MIC) [97], both based on self-training in a teacher-student framework paradigm, for benchmarking purposes. We maintain the original configurations of these models with minimal modifications, focusing primarily on evaluating their transfer learning capabilities when trained on our FMARS dataset. This approach allows us to assess the quality and utility of our automatically generated labels in a standardized setting. We select a separate full-size image (i.e., $17k \times 17k$ pixels) from each event as our test set based on the average information content, for a total of 19 images, and we conduct a full training using pretrained ImageNet weights for the backbone components. To address the label noise (i.e., high precision and lower recall of the generated labels, as well as missing instances) we ignore the background pixels of the FMARS labels during training, since some objects may have been excluded from the final labels due to several factors. For instance, GroundingDINO failed to detect the tree crown, or the building was not present in the original dataset to be provided as prompt. As a simple background baseline, we apply a confidence threshold to the Softmax outputs [170], empirically evaluating the optimal cutoff threshold $\tau = 0.9$ for all models. We conduct every experiment using a tile size of 512×512 and random sampling for 30,000 iterations, using AdamW as the optimizer. For the UDA components, we maintain their original hyperparameters, except for the removal of the ImageNet feature distance.

Method	**Background**		**Roads**		**High Veg.**		**Buildings**		**mAcc.**	**mIoU**
	Acc.	**IoU**	**Acc.**	**IoU**	**Acc.**	**IoU**	**Acc.**	**IoU**		
SegFormer (base)	72.91	61.41	0.11	0.10	7.60	1.33	0.00	0.00	20.15	15.71
MIC	44.79	42.47	55.94	29.89	64.45	10.56	82.47	21.33	61.91	26.06
DAFormer	53.06	50.14	55.44	31.79	64.61	16.80	79.91	17.29	**63.26**	**33.07**

Table 5.12: Performance comparison of the FMARS test set (automated labels), in terms of accuracy and IoU score.

Method	**Background**		**Roads**		**High Veg.**		**Buildings**		**mAcc.**	**mIoU**
	Acc.	**IoU**	**Acc.**	**IoU**	**Acc.**	**IoU**	**Acc.**	**IoU**		
FMARS labels	71.34	41.16	68,72	47.03	69.37	58.54	59.47	54.14	67.23	50.22
SegFormer (base)	97.40	27.90	0.06	0.06	8.24	7.68	0.00	0.00	26.44	8.91
MIC	76.59	36.21	44.84	40.15	51.78	48.52	63.54	56.41	59.19	45.32
DAFormer	70.56	38.02	65.97	54.77	56.57	52.64	69.10	60.20	**65.55**	**51.41**

Table 5.13: Performance comparison on a manually labeled subsample of the FMARS test set, including a comparison with the automated labels.

Results

Given the automated pipeline and the low reliability of the obtained labels for performance measurement, we validate results against the left-out FMARS test set, as well as a small sample of 45 manually labeled tiles, derived from crops of each image in the test set. Tables 5.12 and 5.13 present the numerical results in terms of accuracy and IoU, class-wise and averaged (as in Eq. (2.3)), evaluated against the automatically generated and manual labels, respectively. The baseline SegFormer model, trained without domain adaptation, demonstrates poor performance across all classes, highlighting the challenges of training on automatically generated labels without accounting for domain shift or label noise. In contrast, both UDA

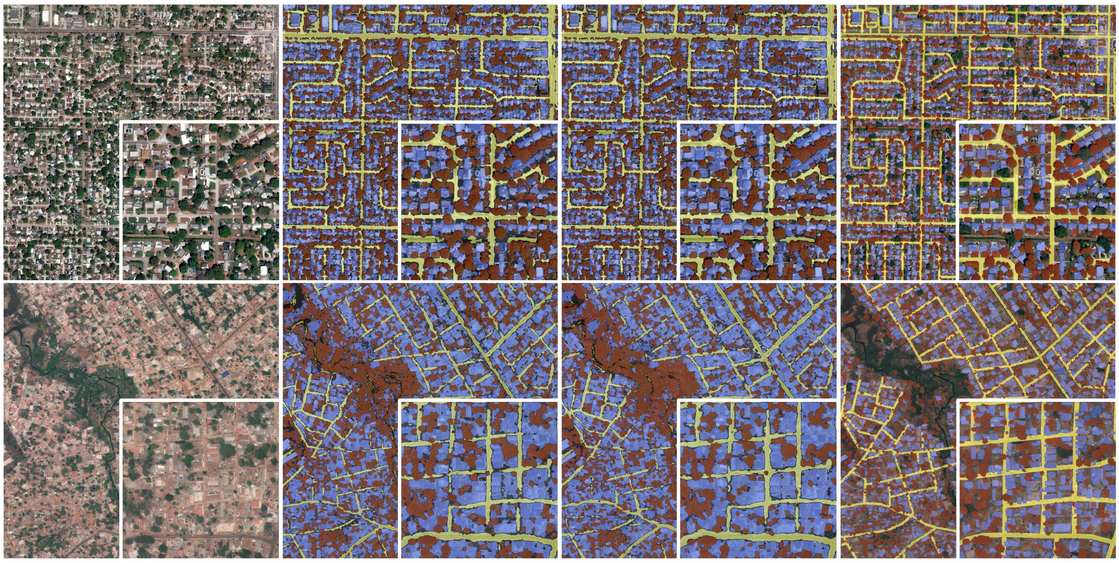


Figure 5.16: Qualitative results on two example areas: USA (top) and Gambia (bottom). From left to right: RGB input image, DAFormer prediction, MIC prediction, and FMARS ground truth.

techniques, DAFormer and MIC, show significant improvements over the baseline. When evaluated against the FMARS labels, MIC achieves a mean IoU of 26.06, while DAFormer slightly outperforms it with 33.07. Interestingly, when evaluated against the manually labeled test set, both UDA models demonstrate even better performance, with DAFormer achieving a mean IoU of 51.41, surpassing the original FMARS labels at 50.22 mIoU. This suggests that the UDA techniques are not only adapting to the domain of the FMARS labels, but also learning to generalize beyond the noise and inaccuracies present in the automated annotations. The performance across different classes provides further insights. Buildings and roads show strong segmentation results, especially with DAFormer, indicating that the FMARS annotations provide a solid foundation for learning these classes. The high vegetation class proves more challenging, likely due to the reliance on GroundingDINO for initial detections, yet both UDA models still show reasonable performance. The background class shows lower IoU scores, which is expected given the background-aware training approach and the potential for missed annotations in the automated process. Qualitative results, displayed in Fig. 5.16, confirm these findings, with both DAFormer and MIC producing accurate segmentation, even in challenging areas with high vegetation or complex urban structures.

Discussion. In this work, we aimed at proving the potential of leveraging VFMs for large-scale annotation of remote sensing imagery, by showing that (i) they can potentially work in a zero-shot paradigm, and (ii) that the resulting annotations can be further applied to downstream tasks effectively. The FMARS pipeline can in fact generate usable annotations in a zero-shot manner, without domain-specific fine-tuning. While the quality of the annotation could be improved, they provide a valuable starting point for training smaller downstream models. Using these labels with appropriate precautions, such as UDA frameworks, it enables smaller models to achieve reasonable performance, even surpassing the original FMARS labels in some cases.

Despite the promising results, there certainly is room for further improvements. The current taxonomy, focusing on three key classes (buildings, roads, and high vegetation), provides a solid foundation, but it does not cover the full range of entities present in VHR images. Expanding this taxonomy in future work would increase the utility of this dataset for a broader range of remote sensing applications. Additionally, the foundation models employed in this study were primarily trained on natural images. Adapting these models specifically for remote sensing data or developing domain-specific foundation models could potentially bridge the domain gap and improve annotation quality, particularly for challenging classes like high vegetation. Future works could also focus on improving upon the scalability of zero-shot approaches, both in terms of accuracy and resources, effectively avoiding further retraining and potentially solving multiple tasks with a more flexible approach.

5.6 Summary

In this chapter, we have explored four key techniques to address the challenges of domain robustness, weak supervision, and large-scale annotation in semantic segmentation for remote sensing applications: Unsupervised Domain Adaptation (UDA), learning from sparse annotations, multitask learning, and leveraging foundation models for automated annotation.

In the context of UDA, we proposed HIUDA, a novel framework that introduces a hierarchical instance mixing strategy (HIMix) and a twin-head architecture to improve domain adaptation performance in aerial imagery. HIMix addresses the shortcomings of existing mixing strategies by preserving the semantic structure of objects and balancing the class distribution between domains. The twin-head architecture enhances the quality and consistency of pseudo-labels for self-training. Extensive experiments on the LoveDA dataset demonstrated the superior performance of HIUDA compared to state-of-the-art UDA methods designed for natural images.

To tackle the issue of limited annotated data, we introduced SPADA, a framework that leverages sparse annotations and self-training to improve semantic segmentation performance. SPADA utilizes a teacher model to generate pseudo-labels on the target domain, which are then mixed with sparse ground truth labels to train the student model in a self-supervised manner. By effectively exploiting both labeled and unlabeled pixels during training, SPADA achieves improved performance in scenarios with limited annotated data.

We investigated the potential of multitask learning to improve the robustness and performance of models in the context of burned area delineation. We proposed RoBAD, a multitask learning framework that incorporates land cover classification as an auxiliary task to guide the training of the burned area delineation model. By learning shared representations between the two tasks, RoBAD demonstrates more stable and robust performance compared to single-task learning, especially in the absence of pretrained solutions.

Lastly, we introduced FMARS, an automated pipeline that leverages foundation models like SAM and GroundingDINO to generate large-scale annotations for VHR remote sensing imagery. FMARS demonstrates the potential of zero-shot annotation and the effectiveness of using these labels for training downstream models with UDA techniques.

While these techniques have shown promising results, there are still limitations to be addressed. The performance of UDA methods like HIUDA may be affected by extreme class imbalances or highly dissimilar domains. SPADA's effectiveness relies on the quality of sparse annotations and the teacher model's ability to generate reliable pseudo-labels. RoBAD's multitask learning approach may be sensitive to the selection of appropriate auxiliary tasks and their balance. FMARS, while effective, could benefit from domain-specific adaptations of foundation models to improve annotation quality.

Future research directions include exploring more advanced mixing strategies and architectures for UDA, investigating the integration of multiple heterogeneous tasks in multitask learning, developing more efficient ways to leverage sparse annotations, and improving the scalability and accuracy of zero-shot annotation approaches. Combining these techniques with domain-specific, large-scale self-supervised pretraining [112, 172] could potentially lead to more robust and generalizable models for remote sensing applications.

Chapter 6

Conclusions

This thesis investigated the unique challenges of semantic segmentation applied to remote sensing imagery and proposed ad-hoc solutions for each encountered scenario. By addressing issues such as orientation invariance, class and scale imbalance, domain robustness, and weak supervision, this work provides a comprehensive overview of grounded problems and practical solutions to tackle them. The main contributions of this thesis can be summarized as follows. First, we addressed the challenges posed by the aerial viewpoint by proposing a framework that integrates Augmentation Invariance (AI) regularization with an Adaptive Sampling (AS) strategy to address orientation invariance and class imbalance in aerial imagery. Additionally, we introduced a Contrastive Regularization and Contrastive Distillation (CRCD) approach for incremental learning, enhancing model robustness to variations in viewpoint and supporting the integration of new classes without forgetting previously learned features. Second, we tackled the issues of class and scale imbalance in remote sensing applications by developing two custom datasets, MMFlood for flood delineation and a dataset for photovoltaic (PV) panel segmentation. We introduced techniques such as Entropy-Weighted Sampling (EWS), multi-encoder architectures, multiscale regularization, and post-processing algorithms, demonstrating improved segmentation accuracy and performance in these specific applications. Third, we explored model robustness through four different works: HIUDA, SPADA, RoBAD, and FMARS, focusing on unsupervised domain adaptation (UDA), segmentation with sparse labels, multitask learning, and foundation models (VFM), respectively. HIUDA provides an ad-hoc framework for UDA that employs hierarchical instance mixing and a twin-head architecture to improve performance on aerial and satellite images. SPADA leverages sparse annotations and self-training to enhance semantic segmentation performance under conditions of limited annotated data. RoBAD, a multitask learning framework, demonstrated the benefits of incorporating auxiliary tasks to improve model robustness in the specific application of burned area delineation. Last, FMARS represents a proof of concept architecture aimed at exploiting large pretrained models for zero-shot learning, or large-scale automated annotation for smaller downstream tasks, where annotations remain scarce.

The approaches presented in this thesis provide insights and practical solutions for semantic segmentation in remote sensing imagery. However, several open challenges and research opportunities remain to be carried out in this field. Thanks to the continuous research efforts and open source contributions, the remote sensing landscape is frequently updated with new and increasingly larger datasets. Future work may first focus on evaluating the generalization capabilities of the proposed techniques on a wider range of remote sensing datasets with different characteristics and scales. Furthermore, while openly available imagery becomes more accessible every year, obtaining reliable and dense segmentation labels remains a major issue. Exploring unsupervised and semi-supervised learning approaches could significantly reduce the reliance on large-scale annotated datasets by leveraging the vast amounts of unlabeled remote sensing data available. Developing methods to effectively fuse and exploit diverse data sources, such as optical imagery, SAR, and LiDAR, could lead to more robust and accurate semantic segmentation models. Finally, investigating the combination of the proposed techniques with large-scale self-supervised pretraining could potentially lead to more robust and generalizable models for remote sensing applications, effectively moving towards the use of large foundation models for multiple and diverse applications.

Appendix A

Panels Polygonization

Algorithm 1: Panel polygonization algorithm. It refines the raw semantic segmentation output by extracting and simplifying contours, converting them to polygons, and applying post-processing steps such as removing small polygons and merging nearby ones.

```
Input:
    $R$, raster prediction to be polygonized
    $t_A$, minimum area threshold
    $t_{DP}$, tolerance for polygonization
    $\alpha$, length factor for edge filtering
Output:
    $P_r$, a set of regularized polygons
Extract polygons from the raster
    // binarize the input prediction
    $B \leftarrow Binarize(R)$
    // extract the set of connected components
    $C \leftarrow CCL(B)$
    // Discard components with area $< t_A$
    $C \leftarrow MinSurface(C, t_A)$
    // Apply Douglas-Peucker to each component
    $P \leftarrow \{\}$
    for $c$ in $C$ do
        $p \leftarrow DouglasPeucker(c, t_{DP})$
        $P \leftarrow P \cup p$
    end
end
Regularize the extracted polygons
    $P_r \leftarrow \{\}$
    for $p$ in $P$ do
        // Extract the MBR
        $mbr \leftarrow MBR(p)$
        // Align edges E with MBR directions
        $E \leftarrow Align(p, mbr)$
        // Remove unnecessary edges
        $E \leftarrow Filter(E, \alpha)$
        // Rebuild the final polygon
        $p_r \leftarrow Link(E)$
        $P_r \leftarrow P_r \cup p_r$
    end
end
```

Appendix B

HIUDA - Pseudocode

Algorithm 2: Pseudocode for the HIUDA training procedure. The model is trained end-to-end on labeled source data and pseudo-labeled target data obtained via the twin-head architecture. Source and target samples are mixed using the HIMix strategy to learn domain-invariant features.

```
Initialize:
Model $f_\theta : \mathcal{X} \rightarrow \mathbf{R}^{|\mathcal{I}| \times |\mathcal{Y}|}$ with encoder $g$ and twin heads $h_1, h_2$;
Input: $\mathcal{X}_S$ source domain with $N_S$ pairs $(x_S, y_S)$, $x_S \in \mathcal{X}, y_S \in \mathcal{Y}$ and
  semantic classes $\mathcal{C}$;
$\mathcal{X}_T$ target domain with $N_T$ images $x_T$, lacking ground truth labels;
Output: $y = \{\text{argmax}_{c \in \mathcal{Y}} p_i^c\}_{i=1}^N$, where $p_i^c$ the model prediction of pixel $i$
  for class $c$ and $\mathcal{Y}$ the label space;
while $epoch\ in\ max\_epochs$ do
    while $x_S, y_S, x_T\ in\ \mathcal{X}_S \times \mathcal{X}_T$ do
        Train on source $\mathcal{X}_S$
            // Compute augmented source batch
              $B_S = (\text{concat}(x_S, \tilde{x}_S), \text{concat}(y_S, \tilde{y}_S))$;
            // Train $f_\theta$ on source labels with $L_{seg}(B_S)$
        end
        Mix source and target pairs
            // Compute pseudo-labels via majority
            // voting $\hat{y}_T = max\,(h_1(g(x_T)), (h_2(g(x_T))))$;
            // Extract source instance labels $i_S = CCL(y_S)$ with instances
              $\in K_S$;
            // Extract target instance pseudo-labels $i_T = CCL(\hat{y}_T)$ with
              instances $\in K_T$;
            // Compute one-hot encoded labels,
            // sorted by pixel size as:
              $1_m = sorted\,(\text{concat}(1_{K_S}(i_S), 1_{K_T}(i_T)))$;
            // Reduce $z$ axis to 2D indexed mask $m = argmax_z 1_m(i, j, z)$;
            // Binarize mask $\forall i, j \in m,\ M = \begin{cases} 1 & if\ m(i,j) \in K_S \\ 0 & if\ m(i,j) \in K_T \end{cases}$;
            // Compute mixed image and labels as:
              $x_M = M \odot x_S + (1 - M) \odot x_T$;
            $y_M = M \odot y_S + (1 - M) \odot \hat{y}_T$;
            // Compute $w_M$ as in Eq. 5.3
        end
        Train on mixed $\mathcal{X}_M$ pairs
            // Compute augmented mixed batch
            $B_M = (\text{concat}(x_M, \tilde{x}_M), \text{concat}(y_M, \tilde{y}_M))$;
            // Train $f_\theta$ on mixed samples with
            // $L_{seg}(B_M)$, weighted by $w_M$
        end
    end
end
```

Bibliography

[1] Arbab Waseem Abbas et al. “K-Means and ISODATA clustering algorithms for landcover classification using remote sensing”. In: *Sindh University Research Journal-SURJ (Science Series)* 48.2 (2016).

[2] Nabila Abraham and Naimul Mefraz Khan. “A novel focal tversky loss function with improved attention u-net for lesion segmentation”. In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 683–687.

[3] Adekanmi Adeyinka Adegun, Serestina Viriri, and Jules-Raymond Tapamo. “Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis”. en. In: *Journal of Big Data* 10.1 (June 2023), p. 93. ISSN: 2196-1115. DOI: 10.1186/s40537-023-00772-x.

[4] Jiwoon Ahn and Suha Kwak. “Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4981–4990.

[5] Andrea Ajmar et al. “Response to Flood Events: The Role of Satellite-based Emergency Mapping and the Experience of the Copernicus Emergency Management Service”. en. In: *Geophysical Monograph Series*. Ed. by Daniela Molinari, Scira Menoni, and Francesco Ballio. 1st ed. Wiley, Aug. 2017, pp. 211–228. ISBN: 978-1-119-21792-3 978-1-119-21793-0. DOI: 10.1002/9781119217930.ch14.

[6] Abdulaziz Amer Aleissaee et al. “Transformers in Remote Sensing: A Survey”. en. In: *Remote Sensing* 15.7 (Jan. 2023). Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, p. 1860. ISSN: 2072-4292. DOI: 10.3390/rs15071860.

[7] Iftikhar Ali et al. “Sentinel-1 Based Near-Real Time Flood Mapping Service.” In: *ISCRAM*. 2018.

[8] Edoardo Arnaudo et al. “A Contrastive Distillation Approach for Incremental Semantic Segmentation in Aerial Images”. en. In: *Image Analysis and Processing – ICIAP 2022*. Ed. by Stan Sclaroff et al. Vol. 13232. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 742–754. DOI: 10.1007/978-3-031-06430-2_62.

[9] Edoardo Arnaudo et al. “Hierarchical instance mixing across domains in aerial segmentation”. In: *IEEE Access* 11 (2023). Publisher: IEEE, pp. 13324–13333.

[10] Edoardo Arnaudo et al. *Robust Burned Area Delineation through Multitask Learning*. arXiv:2309.08368 [cs]. Sept. 2023. URL: http://arxiv.org/abs/2309.08368.

[11] Barry C. Arnold. “Pareto Distribution”. en. In: *Wiley StatsRef: Statistics Reference Online*. Ed. by Ron S. Kenett et al. 1st ed. Wiley, Sept. 2015, pp. 1–10. ISBN: 978-1-118-44511-2. DOI: 10.1002/9781118445112.stat01100.pub2.

[12] Josef Aschbacher. “ESA’s Earth Observation Strategy and Copernicus”. In: *Satellite Earth Observations and Their Impact on Society and Policy*. Journal Abbreviation: Satellite Earth Observations and Their Impact on Society and Policy. June 2017, pp. 81–86. ISBN: 978-981-10-3712-2. DOI: 10.1007/978-981-10-3713-9_5.

[13] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. “Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks”. In: *ISPRS Journal of Photogrammetry and Remote Sensing*. Geospatial Computer Vision 140 (June 2018), pp. 20–32. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2017.11.011.

[14] Cesar Aybar et al. “CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2”. In: *Scientific data* 9.1 (2022). Publisher: Nature Publishing Group UK London, p. 782.

[15] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017). Publisher: IEEE, pp. 2481–2495.

[16] Favyen Bastani et al. “Satlaspretrain: A large-scale dataset for remote sensing image understanding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 16772–16782.

[17] C. Bayik et al. “Exploiting multi-temporal Sentinel-1 SAR data for flood extend mapping”. In: *The international archives of the photogrammetry, remote sensing and spatial information sciences* 42 (2018). Publisher: Copernicus Publications Göttingen, Germany, pp. 109–113.

[18] François Becker and Zhao-Liang Li. “Surface temperature and emissivity at various scales: Definition, measurement and related problems”. en. In: *Remote Sensing Reviews* 12.3-4 (Jan. 1995), pp. 225–253. ISSN: 0275-7257. DOI: 10.1080/02757259509532286.

[19] Nadir Bengana and Janne Heikkilä. “Improving land cover segmentation across satellites using domain adaptation”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2020). Publisher: IEEE, pp. 1399–1410.

[20] Abdelhakim Benoudjit and Raffaella Guida. “A novel fully automated mapping of the flood extent on SAR images using a supervised classifier”. In: *Remote Sensing* 11.7 (2019). Publisher: MDPI, p. 779.

[21] Adrian Boguszewski et al. “LandCover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1102–1110.

[22] 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 Workshops*. 2020, pp. 210–211.

[23] Kyle Bradbury et al. “Distributed solar photovoltaic array location and extent dataset for remote sensing object identification”. In: *Scientific data* 3.1 (2016). Publisher: Nature Publishing Group, pp. 1–9.

[24] William M. Brown. “Synthetic aperture radar”. In: *IEEE Transactions on Aerospace and Electronic Systems* 2 (1967). Publisher: IEEE, pp. 217–229. (Visited on 06/08/2024).

[25] F. Caltagirone et al. “COSMO-SkyMed program: Status and perspectives”. In: *Proceedings of the 3rd International Workshop on Satellite Constellations and Formation Flying*. 2003, pp. 11–16.

[26] Daniele Rege Cambrin, Luca Colomba, and Paolo Garza. “CaBuAr: California Burned Areas dataset for delineation”. In: *IEEE Geoscience and Remote Sensing Magazine* 11.3 (Sept. 2023). arXiv:2401.11519 [cs, eess], pp. 106–113. ISSN: 2168-6831, 2473-2397, 2373-7468. DOI: 10.1109/MGRS.2023.3292467.

[27] Daniele Rege Cambrin, Luca Colomba, and Paolo Garza. “Vision Transformers for Burned Area Delineation.” In: *MACLEAN@ PKDD/ECML*. 2022.

[28] Joseph Camilo et al. *Application of a semantic segmentation convolutional neural network for accurate automatic detection and mapping of solar photovoltaic arrays in aerial imagery*. arXiv:1801.04018 [cs]. Jan. 2018. URL: http://arxiv.org/abs/1801.04018.

[29] Gustavo Camps-Valls et al. *Remote Sensing Image Processing*. en. Synthesis Lectures on Image, Video, and Multimedia Processing. Cham: Springer International Publishing, 2012. ISBN: 978-3-031-01119-1 978-3-031-02247-0. DOI: 10.1007/978-3-031-02247-0.

[30] Simon J. Cantrell et al. *System characterization report on the WorldView-3 Imager*. en. Tech. rep. 2021-1030-I. ISSN: 2331-1258 Publication Title: Open-File Report. U.S. Geological Survey, 2021. DOI: 10.3133/ofr20211030I.

[31] Changyong Cao et al. "Early on-orbit performance of the visible infrared imaging radiometer suite onboard the Suomi National Polar-Orbiting Partnership (S-NPP) satellite". In: *IEEE Transactions on Geoscience and Remote Sensing* 52.2 (2013). Publisher: IEEE, pp. 1142–1156.

[32] Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *Advances in neural information processing systems* 33 (2020), pp. 9912–9924.

[33] Roberto Castello et al. "Deep learning in the built environment: Automatic detection of rooftop solar panels using Convolutional Neural Networks". In: *Journal of Physics: Conference Series*. Vol. 1343. Issue: 1. IOP Publishing, 2019, p. 012034.

[34] Eduardo Castro, Jaime S. Cardoso, and Jose Costa Pereira. "Elastic deformations for data augmentation in breast cancer mass detection". In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. Mar. 2018, pp. 230–234. (Visited on 09/29/2024).

[35] Fabio Cermelli et al. "Modeling the background for incremental learning in semantic segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9233–9242.

[36] Krishna Chaitanya et al. "Contrastive learning of global and local features for medical image segmentation with limited annotations". In: *Advances in neural information processing systems* 33 (2020), pp. 12546–12558.

[37] Yang-Lang Chang et al. "Ship detection based on YOLOv2 for SAR imagery". In: *Remote Sensing* 11.7 (2019). Publisher: MDPI, p. 786.

[38] 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 (2017). Publisher: IEEE, pp. 834–848.

[39] Liang-Chieh Chen et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[40] Liang-Chieh Chen et al. *Rethinking Atrous Convolution for Semantic Image Segmentation*. arXiv:1706.05587 [cs]. Dec. 2017. URL: http://arxiv.org/abs/1706.05587.

[41] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[42] Wuyang Chen et al. “Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 8924–8933.

[43] Zhao Chen et al. “Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks”. In: *International conference on machine learning*. PMLR, 2018, pp. 794–803.

[44] Bowen Cheng, Alex Schwing, and Alexander Kirillov. “Per-pixel classification is not all you need for semantic segmentation”. In: *Advances in neural information processing systems* 34 (2021), pp. 17864–17875.

[45] Bowen Cheng et al. “Masked-attention mask transformer for universal image segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 1290–1299.

[46] Dominic Cheng et al. “Darnet: Deep active ray network for building segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 7431–7439.

[47] Michele Chevrel, MICHEL Courtois, and G. Weill. “The SPOT satellite remote sensing mission”. In: *Photogrammetric Engineering and Remote Sensing* 47 (1981), pp. 1163–1171.

[48] Mingmin Chi et al. “Big data for remote sensing: Challenges and opportunities”. In: *Proceedings of the IEEE* 104.11 (2016). Publisher: IEEE, pp. 2207–2219.

[49] Mang Tik Chiu et al. “Agriculture-vision: A large aerial image database for agricultural pattern analysis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 2828–2838.

[50] Allison E. Cocke, Peter Z. Fulé, and Joseph E. Crouse. “Comparison of burn severity assessments using Differenced Normalized Burn Ratio and ground data”. In: *International Journal of Wildland Fire* 14.2 (2005). Publisher: CSIRO Publishing, pp. 189–198.

[51] Luca Colomba et al. “A Dataset for Burned Area Delineation and Severity Estimation from Satellite Imagery”. en. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. Atlanta GA USA: ACM, Oct. 2022, pp. 3893–3897. ISBN: 978-1-4503-9236-5. DOI: 10.1145/3511808.3557528.

[52] Marius Cordts et al. “The cityscapes dataset for semantic urban scene understanding”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016, pp. 3213–3223.

[53] Corine Land Cover. “Corine land cover”. In: *European Environment Agency, Copenhagen* (2000).

[54] Annarita D’Addabbo et al. “A Bayesian network for flood detection combining SAR imagery and ancillary data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 54.6 (2016). Publisher: IEEE, pp. 3612–3625.

[55] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05).* Vol. 1. Ieee, 2005, pp. 886–893.

[56] Ruth DeFries. “Remote Sensing and Image Processing”. en. In: Book Title: Encyclopedia of Biodiversity. Elsevier, 2013, pp. 389–399. ISBN: 978-0-12-384720-1. DOI: 10.1016/B978-0-12-384719-5.00383-X.

[57] Ilke Demir et al. “Deepglobe 2018: A challenge to parse the earth through satellite images”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.* 2018, pp. 172–181.

[58] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv:1810.04805 [cs]. May 2019. URL: http://arxiv.org/abs/1810.04805 (visited on 06/20/2024).

[59] Foivos I. Diakogiannis et al. “ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 162 (Apr. 2020), pp. 94–114. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2020.01.013.

[60] Craig Donlon et al. “The sentinel-3 mission: Overview and status”. In: *2012 IEEE International Geoscience and Remote Sensing Symposium.* IEEE, 2012, pp. 1711–1714.

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

[62] Gestore Servizi Energetici. *Statistical report - solar and photovoltaic.* 2022. URL: https://www.gse.it/dati-e-scenari/statistiche.

[63] Andreas Ess et al. “Segmentation-Based Urban Traffic Scene Understanding.” In: *BMVC.* Vol. 1. Citeseer, 2009, p. 2.

[64] European Commission. *Copernicus emergency management service.* 2020. URL: https://emergency.copernicus.eu/ (visited on 01/06/2024).

[65] European Space Agency. *Sentinel-2 MSI data.* 2015. URL: https://scihub.copernicus.eu/.

[66] (ESA) European Space Agency. *Mission ends for Copernicus Sentinel-1B satellite*. en. URL: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1/Mission_ends_for_Copernicus_Sentinel-1B_satellite (visited on 06/15/2024).

[67] (ESA) European Space Agency. *Sentinel-2 MSI Data*. 2016. URL: https://scihub.copernicus.eu/ (visited on 06/01/2024).

[68] (ESA) European Space Agency. *The Sentinel missions*. en. URL: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/The_Sentinel_missions (visited on 06/14/2024).

[69] Mark Everingham et al. "The pascal visual object classes challenge: A retrospective". In: *International journal of computer vision* 111 (2015). Publisher: Springer, pp. 98–136.

[70] Clement Farabet et al. "Learning hierarchical features for scene labeling". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2012). Publisher: IEEE, pp. 1915–1929.

[71] Abolfazl Farahani et al. "A Brief Review of Domain Adaptation". en. In: *Advances in Data Science and Information Engineering*. Ed. by Robert Stahlbock et al. Series Title: Transactions on Computational Science and Computational Intelligence. Cham: Springer International Publishing, 2021, pp. 877–894. ISBN: 978-3-030-71703-2 978-3-030-71704-9. DOI: 10.1007/978-3-030-71704-9_65.

[72] Alessandro Farasin, Luca Colomba, and Paolo Garza. "Double-step u-net: A deep learning-based approach for the estimation of wildfire damage severity through sentinel-2 satellite data". In: *Applied Sciences* 10.12 (2020). Publisher: MDPI, p. 4332.

[73] Alessandro Farasin et al. "Unsupervised Burned Area Estimation through Satellite Tiles: A Multimodal Approach by Means of Image Segmentation over Remote Sensing Imagery." In: *MACLEAN@ PKDD/ECML*. 2019.

[74] Mulham Fawakherji et al. "Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation". In: *2019 third IEEE international conference on robotic computing (IRC)*. IEEE, 2019, pp. 146–152.

[75] Yingchao Feng et al. "Continual learning with structured inheritance for semantic segmentation in aerial imagery". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021). Publisher: IEEE, pp. 1–17.

[76] Federico Filipponi. "BAIS2: Burned area index for Sentinel-2". In: *Proceedings*. Vol. 2. Issue: 7. MDPI, 2018, p. 364.

[77] Jun Fu et al. “Dual attention network for scene segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3146–3154.

[78] Marco Galatola et al. “Land Cover Segmentation with Sparse Annotations from Sentinel-2 Imagery”. In: *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 6952–6955.

[79] Alessandro Galdelli et al. “A synergic integration of AIS data and SAR imagery to monitor fisheries and detect suspicious activities”. In: *Sensors* 21.8 (2021). Publisher: MDPI, p. 2756.

[80] Alberto Garcia-Garcia et al. *A Review on Deep Learning Techniques Applied to Semantic Segmentation*. arXiv:1704.06857 [cs]. Apr. 2017. URL: http://arxiv.org/abs/1704.06857.

[81] Fan Ge et al. “A hierarchical information extraction method for large-scale centralized photovoltaic power plants based on multi-source remote sensing images”. In: *Remote Sensing* 14.17 (2022). Publisher: MDPI, p. 4211.

[82] Bo Geng, Dacheng Tao, and Chao Xu. “DAML: Domain adaptation metric learning”. In: *IEEE Transactions on Image Processing* 20.10 (2011). Publisher: IEEE, pp. 2980–2989.

[83] Golnaz Ghiasi et al. “Simple copy-paste is a strong data augmentation method for instance segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2918–2928.

[84] M. Alain Gleyzes, Lionel Perret, and Philippe Kubik. “Pleiades system architecture and main performances”. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 39 (2012). Publisher: Copernicus GmbH, pp. 537–542.

[85] Rafael C. Gonzalez. *Digital image processing*. Pearson education india, 2009.

[86] Haonan Guo et al. “Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 59.5 (2020). Publisher: IEEE, pp. 4287–4306.

[87] Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. “Learning to branch for multi-task learning”. In: *International conference on machine learning*. PMLR, 2020, pp. 3854–3863.

[88] Yanming Guo et al. “A review of semantic segmentation using deep neural networks”. en. In: *International Journal of Multimedia Information Retrieval* 7.2 (June 2018), pp. 87–93. ISSN: 2192-662X. DOI: 10.1007/s13735-017-0141-z.

[89] Ritwik Gupta et al. “Creating xBD: A dataset for assessing building damage from satellite imagery”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops.* 2019, pp. 10–17.

[90] Ephrem Habyarimana et al. “Towards predictive modeling of sorghum biomass yields using fraction of absorbed photosynthetically active radiation derived from sentinel-2 satellite imagery and supervised machine learning techniques”. In: *Agronomy* 9.4 (2019). Publisher: MDPI, p. 203.

[91] Jiaming Han et al. “Redet: A rotation-equivariant detector for aerial object detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2021, pp. 2786–2795.

[92] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016, pp. 770–778.

[93] Kaiming He et al. “Momentum contrast for unsupervised visual representation learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2020, pp. 9729–9738.

[94] Lifeng He et al. “The connected-component labeling problem: A review of state-of-the-art algorithms”. In: *Pattern Recognition* 70 (2017). Publisher: Elsevier, pp. 25–43.

[95] Judy Hoffman et al. “Cycada: Cycle-consistent adversarial domain adaptation”. In: *International conference on machine learning.* Pmlr, 2018, pp. 1989–1998.

[96] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. “Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2022, pp. 9924–9935.

[97] Lukas Hoyer et al. “MIC: Masked image consistency for context-enhanced domain adaptation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2023, pp. 11721–11732.

[98] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018, pp. 7132–7141.

[99] Yuansheng Hua et al. “Semantic segmentation of remote sensing images with sparse annotations”. In: *IEEE Geoscience and Remote Sensing Letters* 19 (2021). Publisher: IEEE, pp. 1–5.

[100] Zilong Huang et al. “Ccnet: Criss-cross attention for semantic segmentation”. In: *Proceedings of the IEEE/CVF international conference on computer vision.* 2019, pp. 603–612.

[101] David Hutchison et al. “What, Where and How Many? Combining Object Detectors and CRFs”. en. In: *Computer Vision – ECCV 2010*. Ed. by Kostas Daniilidis, Petros Maragos, and Nikos Paragios. Vol. 6314. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 424–437. ISBN: 978-3-642-15560-4 978-3-642-15561-1. DOI: 10.1007/978-3-642-15561-1_31. (Visited on 06/15/2024).

[102] Interagency Implementation, NASA Advanced ConceptsTyler, and IEEE GRSS Earth Science Informatics Technical Committee. *ETCI 2021 competition on flood detection*. 2021. URL: https://nasa-impact.github.io/etci2021/ (visited on 05/01/2022).

[103] Gareth Ireland, Michele Volpi, and George P. Petropoulos. “Examining the capability of supervised machine learning classifiers in extracting flooded areas from Landsat TM imagery: a case study from a Mediterranean flood”. In: *Remote sensing* 7.3 (2015). Publisher: MDPI, pp. 3372–3399.

[104] Sen Jia, Zhichao Min, and Xiyou Fu. “Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion”. In: *Information Fusion* 96 (2023). Publisher: Elsevier, pp. 117–129. (Visited on 06/19/2024).

[105] S. Jutz and M. P. Milagro-Perez. “Copernicus: the European Earth Observation programme”. In: *Revista de Teledetección* 56 (2020), pp. V–XI.

[106] Wenchao Kang et al. “Flood detection in Gaofen-3 SAR images via fully convolutional networks”. In: *Sensors* 18.9 (2018). Publisher: MDPI, p. 2915.

[107] Gabriel Kasmi et al. *Towards unsupervised assessment with open-source data of the accuracy of deep learning-based distributed PV mapping*. arXiv:2207.07466 [cs, eess]. Feb. 2023. URL: http://arxiv.org/abs/2207.07466.

[108] Bala Bhavya Kausika et al. “GeoAI for detection of solar photovoltaic installations in the Netherlands”. In: *Energy and AI* 6 (Dec. 2021), p. 100111. ISSN: 2666-5468. DOI: 10.1016/j.egyai.2021.100111.

[109] Alex Kendall, Yarin Gal, and Roberto Cipolla. “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7482–7491.

[110] Hoel Kervadec et al. “Boundary loss for highly unbalanced segmentation”. In: *International conference on medical imaging with deep learning*. PMLR, 2019, pp. 285–296.

[111] Prannay Khosla et al. “Supervised contrastive learning”. In: *Advances in neural information processing systems* 33 (2020), pp. 18661–18673.

[112] Alexander Kirillov et al. “Segment anything”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[113] Victor Klemas. "Remote sensing of floods and flood-prone areas: An overview". In: *Journal of Coastal Research* 31.4 (2015). Publisher: The Coastal Education and Research Foundation, pp. 1005–1013.

[114] Lisa Knopp et al. "A deep learning approach for burned area segmentation with Sentinel-2 data". In: *Remote Sensing* 12.15 (2020). Publisher: MDPI, p. 2422.

[115] Philipp Krähenbühl and Vladlen Koltun. "Efficient inference in fully connected crfs with gaussian edge potentials". In: *Advances in neural information processing systems* 24 (2011).

[116] Herbert J. Kramer. *Observation of the Earth and Its Environment: Survey of Missions and Sensors*. en. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. ISBN: 978-3-642-62688-3 978-3-642-56294-5. DOI: 10.1007/978-3-642-56294-5.

[117] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[118] Haitao Lang, Siwen Wu, and Yongjie Xu. "Ship classification in SAR images improved by AIS knowledge transfer". In: *IEEE Geoscience and Remote Sensing Letters* 15.3 (2018). Publisher: IEEE, pp. 439–443.

[119] Francesco Lattari et al. "Deep learning for SAR image despeckling". In: *Remote Sensing* 11.13 (2019). Publisher: MDPI, p. 1532.

[120] Jose M. Leiva-Murillo, Luis Gómez-Chova, and Gustavo Camps-Valls. "Multitask remote sensing data classification". In: *IEEE transactions on geoscience and remote sensing* 51.1 (2012). Publisher: IEEE, pp. 151–161.

[121] Biao Li et al. *A Survey on Semantic Segmentation*. Pages: 1240. Nov. 2018. DOI: 10.1109/ICDMW.2018.00176.

[122] Lusi Li, Haibo He, and Jie Li. "Entropy-based sampling approaches for multi-class imbalanced problems". In: *IEEE Transactions on Knowledge and Data Engineering* 32.11 (2019). Publisher: IEEE, pp. 2159–2170.

[123] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. "Bidirectional learning for domain adaptation of semantic segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 6936–6945.

[124] Zhizhong Li and Derek Hoiem. "Learning without forgetting". In: *IEEE transactions on pattern analysis and machine intelligence* 40.12 (2017). Publisher: IEEE, pp. 2935–2947.

[125] Qing Lian et al. “Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6758–6767.

[126] Zhiyuan Liang et al. “Tree energy loss: Towards sparsely annotated semantic segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16907–16916.

[127] Di Lin et al. “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3159–3167.

[128] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[129] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. en. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Vol. 8693. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 740–755. ISBN: 978-3-319-10601-4 978-3-319-10602-1. DOI: 10.1007/978-3-319-10602-1_48.

[130] Shilong Liu et al. *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. arXiv:2303.05499 [cs]. Mar. 2023. URL: http://arxiv.org/abs/2303.05499 (visited on 06/19/2024).

[131] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022. (Visited on 06/09/2024).

[132] Zhenjie Liu et al. “Moving Ship Optimal Association for Maritime Surveillance: Fusing AIS and Sentinel-2 Data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022). Publisher: IEEE, pp. 1–18.

[133] 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.

[134] Mohammad Reza Loghmani et al. “Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition”. In: *IEEE Robotics and Automation Letters* 5.4 (2020). Publisher: IEEE, pp. 6631–6638.

[135] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[136] Mingsheng Long et al. “Learning transferable features with deep adaptation networks”. In: *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[137] David G. Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International journal of computer vision* 60 (2004). Publisher: Springer, pp. 91–110.

[138] Jun Lu et al. “Automated flood detection with improved robustness and efficiency using multi-temporal SAR data”. en. In: *Remote Sensing Letters* 5.3 (Mar. 2014), pp. 240–248. ISSN: 2150-704X, 2150-7058. DOI: 10.1080/2150704X.2014.898190.

[139] Yawei Luo et al. “Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2507–2516.

[140] Ning Lv et al. “Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in SAR images”. In: *IEEE transactions on industrial informatics* 14.12 (2018). Publisher: IEEE, pp. 5530–5538.

[141] Zhenhua Lv et al. “Parallel K-Means Clustering of Remote Sensing Images Based on MapReduce”. In: *Web Information Systems and Mining*. Ed. by Fu Lee Wang et al. Vol. 6318. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 162–170. ISBN: 978-3-642-16514-6 978-3-642-16515-3. DOI: 10.1007/978-3-642-16515-3_21.

[142] Radek Malinowski et al. “Automated production of a land cover/use map of Europe based on Sentinel-2 imagery”. In: *Remote Sensing* 12.21 (2020). Publisher: MDPI, p. 3523.

[143] Arun Mallya and Svetlana Lazebnik. “Packnet: Adding multiple tasks to a single network by iterative pruning”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pp. 7765–7773.

[144] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. “Attentive single-tasking of multiple tasks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1851–1860.

[145] Sandro Martinis, Jens Kersten, and André Twele. “A fully automated TerraSAR-X based flood service”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 104 (2015). Publisher: Elsevier, pp. 203–212.

[146] Sandro Martinis, André Twele, and Stefan Voigt. “Towards operational near real-time flood detection using a split-based automatic thresholding procedure on high resolution TerraSAR-X data”. In: *Natural Hazards and Earth System Sciences* 9.2 (2009). Publisher: Copernicus GmbH, pp. 303–314.

[147] Valérie Masson-Delmotte et al., eds. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2021. DOI: 10.1017/9781009157896.

[148] Maxar. *Maxar Intelligence & Maxar Space Systems*. URL: https://www.maxar.com/ (visited on 06/01/2024).

[149] Michael McCloskey and Neal J. Cohen. "Catastrophic interference in connectionist networks: The sequential learning problem". In: *Psychology of learning and motivation*. Vol. 24. Elsevier, 1989, pp. 109–165.

[150] Ke Mei et al. "Instance Adaptive Self-training for Unsupervised Domain Adaptation". en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Vol. 12371. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 415–430. ISBN: 978-3-030-58573-0 978-3-030-58574-7. DOI: 10.1007/978-3-030-58574-7_25.

[151] *Microsoft Global Building Footprints*. original-date: 2022-04-22T22:09:24Z. June 2024. URL: https://github.com/microsoft/GlobalMLBuildingFootprints (visited on 06/20/2024).

[152] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. "Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs". In: *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 2229–2235.

[153] Jay D. Miller and Andrea E. Thode. "Quantifying burn severity in a heterogeneous landscape with a relative version of the delta Normalized Burn Ratio (dNBR)". In: *Remote sensing of Environment* 109.1 (2007). Publisher: Elsevier, pp. 66–80.

[154] Mohsen Mirzaei and Mostafa Zamani Mohiabadi. "A comparative analysis of long-term field test of monocrystalline and polycrystalline PV power generation in semi-arid climate conditions". In: *Energy for Sustainable Development* 38 (2017). Publisher: Elsevier, pp. 93–101.

[155] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 6707–6717.

[156] Ishan Misra et al. "Cross-stitch networks for multi-task learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3994–4003.

[157] *MMSegmentation*. 2024. URL: https://github.com/open-mmlab/mmsegmentation (visited on 06/01/2024).

[158] MMSegmentation Contributors. *OpenMMLab semantic segmentation toolbox and benchmark*. 2020. URL: https://github.com/open-mmlab/mmsegmentation.

[159] Andrea Bordone Molini et al. "Speckle2Void: Deep self-supervised SAR despeckling with blind-spot convolutional neural networks". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021). Publisher: IEEE, pp. 1–17.

[160] Simone Monaco et al. "Attention to fires: Multi-channel deep learning models for wildfire severity prediction". In: *Applied Sciences* 11.22 (2021). Publisher: MDPI, p. 11060.

[161] Fabio Montello, Edoardo Arnaudo, and Claudio Rossi. "Mmflood: A multimodal dataset for flood delineation from satellite imagery". In: *IEEE Access* 10 (2022). Publisher: IEEE, pp. 96774–96787.

[162] L C Morena, K V James, and J. Beck. "An introduction to the RADARSAT-2 mission". en. In: *Canadian Journal of Remote Sensing* 30.3 (Jan. 2004), pp. 221–234. ISSN: 0703-8992, 1712-7971. DOI: 10.5589/m04-004.

[163] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12416–12425.

[164] Ruigang Niu et al. "Hybrid multiple attention network for semantic segmentation in aerial images". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021). Publisher: IEEE, pp. 1–18.

[165] Adrien Nivaggioli and Hicham Randrianarivo. "Weakly supervised semantic segmentation of satellite images". In: *2019 Joint urban remote sensing event (JURSE)*. IEEE, 2019, pp. 1–4.

[166] Keiller Nogueira et al. "Learning to semantically segment high-resolution remote sensing images". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 3566–3571.

[167] Mehdi Noroozi and Paolo Favaro. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles". en. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Vol. 9910. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 69–84. ISBN: 978-3-319-46465-7 978-3-319-46466-4. DOI: 10.1007/978-3-319-46466-4_5.

[168] I. Ogashawara, C. Ferreira, and M. P. Curtarelli. "Hurricane coastal flood analysis using multispectral spectral images". In: *AGU Fall Meeting Abstracts*. Vol. 2013. 2013, NH51C–1634.

[169] Igor Ogashawara, Marcelo Curtarelli, and Celso Ferreira. "The use of optical remote sensing for mapping flooded areas". In: *International Journal of Engineering Research and Application* 3 (Oct. 2013), pp. 1956–1960.

[170] Hugo Oliveira et al. "Fully convolutional open set segmentation". en. In: *Machine Learning* 112.5 (May 2023), pp. 1733–1784. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-021-06027-1. (Visited on 06/20/2024).

[171] Viktor Olsson et al. "Classmix: Segmentation-based data augmentation for semi-supervised learning". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 1369–1378.

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

[173] Lucas Prado Osco et al. "The segment anything model (sam) for remote sensing applications: From zero to one shot". In: *International Journal of Applied Earth Observation and Geoinformation* 124 (2023). Publisher: Elsevier, p. 103540. URL: https://www.sciencedirect.com/science/article/pii/S1569843223003643 (visited on 06/19/2024).

[174] Batuhan Osmanoğlu et al. "Time series analysis of InSAR data: Methods and trends". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 115 (2016). Publisher: Elsevier, pp. 90–102.

[175] Thomas S. Pagano and Rodney M. Durham. "Moderate resolution imaging spectroradiometer (MODIS)". In: *Sensor Systems for the Early Earth Observing System Platforms*. Vol. 1939. SPIE, 1993, pp. 2–17.

[176] Giulio Palomba, Alessandro Farasin, and Claudio Rossi. "Sentinel-1 Flood Delineation with Supervised Machine Learning." In: *ISCRAM*. 2020, pp. 1072–1083.

[177] Bin Pan et al. "CoinNet: Copy initialization network for multispectral imagery semantic segmentation". In: *IEEE Geoscience and Remote Sensing Letters* 16.5 (2018). Publisher: IEEE, pp. 816–820.

[178] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. en. Dec. 2019. URL: https://arxiv.org/abs/1912.01703v1.

[179] Nathalie Pettorelli et al. *Conservation Technology Series Issue 4: Satellite Remote Sensing for Conservation*. Apr. 2018. DOI: 10.13140/RG.2.2.25962.41926.

[180] Nicolas Pielawski et al. "CoMIR: Contrastive multimodal image representation for registration". In: *Advances in neural information processing systems* 33 (2020), pp. 18433–18444.

[181] Sankaranarayanan Piramanayagam et al. "Supervised Classification of Multisensor Remotely Sensed Images Using a Deep Learning Framework". en. In: *Remote Sensing* 10.9 (Sept. 2018). Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 1429. ISSN: 2072-4292. DOI: 10.3390/rs10091429.

[182] Wolfgang Pitz and David Miller. "The terrasar-x satellite". In: *IEEE Transactions on Geoscience and Remote Sensing* 48.2 (2010). Publisher: IEEE, pp. 615–622.

[183] Veerle Plakman, Job Rosier, and Jasper Van Vliet. "Solar park detection from publicly available satellite imagery". en. In: *GIScience & Remote Sensing* 59.1 (Dec. 2022), pp. 462–481. ISSN: 1548-1603, 1943-7226. DOI: 10.1080/15481603.2022.2036056.

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

[185] Ratna Prastyani and Abdul Basith. "Utilisation of Sentinel-1 SAR imagery for oil spill mapping: a case study of Balikpapan Bay oil spill". In: *Journal of Geospatial Information Science and Engineering* 1.1 (2018), pp. 22–26.

[186] Luca Pulvirenti et al. "An algorithm for operational flood mapping from Synthetic Aperture Radar (SAR) data using fuzzy logic". In: *Natural Hazards and Earth System Sciences* 11.2 (2011). Publisher: Copernicus Publications Göttingen, Germany, pp. 529–540.

[187] Kunlun Qi et al. "Rotation invariance regularization for remote sensing image scene classification with convolutional neural networks". In: *Remote Sensing* 13.4 (2021). Publisher: MDPI, p. 569.

[188] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[189] Maryam Rahnemoonfar et al. "FloodNet: a high resolution aerial imagery dataset for post flood scene understanding". In: *IEEE access : practical innovations, open solutions* 9 (2021), pp. 89644–89654. DOI: 10.1109/ACCESS.2021.3090981.

[190] 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.

[191] Rubén Ramo and Emilio Chuvieco. "Developing a random forest algorithm for MODIS global burned area classification". In: *Remote Sensing* 9.11 (2017). Publisher: MDPI, p. 1193.

[192] R. Keith Raney et al. "Radarsat (sar imaging)". In: *Proceedings of the IEEE* 79.6 (1991). Publisher: IEEE, pp. 839–849.

[193] Sylvestre-Alvise Rebuffi et al. “icarl: Incremental classifier and representation learning”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017, pp. 2001–2010.

[194] Tianhe Ren et al. *Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks*. arXiv:2401.14159 [cs]. Jan. 2024. URL: http://arxiv.org/abs/2401.14159 (visited on 06/20/2024).

[195] John A. Richards. *Remote Sensing Digital Image Analysis*. en. Cham: Springer International Publishing, 2022. ISBN: 978-3-030-82326-9 978-3-030-82327-6. DOI: 10.1007/978-3-030-82327-6.

[196] Tal Ridnik et al. “Tresnet: High performance gpu-dedicated architecture”. In: *proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 1400–1409.

[197] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. en. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Vol. 9351. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24573-7 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

[198] German Ros et al. “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3234–3243.

[199] Franz Rottensteiner et al. *2D semantic labeling contest*. 2020. URL: https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx (visited on 01/06/2024).

[200] D. P. Roy et al. “Landsat-8: Science and product vision for terrestrial global change research”. In: *Remote Sensing of Environment* 145 (Apr. 2014), pp. 154–172. ISSN: 0034-4257. DOI: 10.1016/j.rse.2014.02.001.

[201] Sebastian Ruder. *An Overview of Multi-Task Learning in Deep Neural Networks*. arXiv:1706.05098 [cs, stat]. June 2017. URL: http://arxiv.org/abs/1706.05098.

[202] Alan Saalfeld. “Topologically Consistent Line Simplification with the Douglas-Peucker Algorithm”. en. In: *Cartography and Geographic Information Science* 26.1 (Jan. 1999), pp. 7–18. ISSN: 1523-0406, 1545-0465. DOI: 10.1559/152304099782424901.

[203] Sebastien Saunier et al. “SkySat data quality assessment within the EDAP framework”. In: *Remote Sensing* 14.7 (2022). Publisher: MDPI, p. 1646.

[204] Thomas J. Schmugge et al. “Advanced spaceborne thermal emission and reflection radiometer (ASTER)”. In: *Remote Sensing for Agriculture, Ecosystems, and Hydrology IV*. Vol. 4879. SPIE, 2003, pp. 1–12.

[205] Max Schwarz et al. “RGB-D object detection and semantic segmentation for autonomous manipulation in clutter”. en. In: *The International Journal of Robotics Research* 37.4-5 (Apr. 2018), pp. 437–451. ISSN: 0278-3649, 1741-3176. DOI: 10.1177/0278364917713117.

[206] Ozan Sener and Vladlen Koltun. “Multi-task learning as multi-objective optimization”. In: *Advances in neural information processing systems* 31 (2018).

[207] *Sentinel-1*. en. URL: https://sentiwiki.copernicus.eu/web/sentinel-1 (visited on 06/15/2024).

[208] Yu Shen et al. “Bdanet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021). Publisher: IEEE, pp. 1–14.

[209] Hao Sheng et al. “Effective data fusion with generalized vegetation index: Evidence from land cover segmentation in agriculture”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 60–61.

[210] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. “Semantic texton forests for image categorization and segmentation”. In: *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[211] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556 [cs]. Apr. 2015. URL: http://arxiv.org/abs/1409.1556.

[212] Sinergise Solutions d.o.o., a Planet Labs company. *Sentinel hub*. 2023. URL: https://www.sentinel-hub.com.

[213] Suriya Singh et al. “Self-supervised feature learning for semantic segmentation of overhead imagery”. In: (2018). Publisher: BMVA Press.

[214] Microsoft Open Source et al. *Microsoft Planetary Computer*. Oct. 2022. DOI: 10.5281/zenodo.7261897. URL: https://doi.org/10.5281/zenodo.7261897.

[215] Robin Strudel et al. “Segmenter: Transformer for semantic segmentation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 7262–7272.

[216] Gencer Sumbul et al. “BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]”. In: *IEEE Geoscience and Remote Sensing Magazine* 9.3 (2021). Publisher: IEEE, pp. 174–180.

[217] Onur Tasar, Yuliya Tarabalka, and Pierre Alliez. “Incremental learning for semantic segmentation of large-scale remote sensing data”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.9 (2019). Publisher: IEEE, pp. 3524–3537.

[218] Antonio Tavera et al. “Augmentation invariance and adaptive sampling in semantic segmentation of agricultural aerial images”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 1656–1665.

[219] Antonio Tavera et al. “Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 1626–1635.

[220] Alan A. Thompson*. “Overview of the RADARSAT Constellation Mission”. en. In: *Canadian Journal of Remote Sensing* 41.5 (Sept. 2015), pp. 401–407. ISSN: 0703-8992, 1712-7971. DOI: 10.1080/07038992.2015.1104633.

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

[222] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention”. In: *International conference on machine learning*. PMLR, 2021, pp. 10347–10357.

[223] Wilhelm Tranheden et al. “Dacs: Domain adaptation via cross-domain mixed sampling”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1379–1389.

[224] Yi-Hsuan Tsai et al. “Learning to adapt structured output space for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7472–7481.

[225] Gabriel Tseng et al. “Cropharvest: A global dataset for crop-type classification”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

[226] André Twele et al. “Sentinel-1-based flood mapping: a fully automated processing chain”. en. In: *International Journal of Remote Sensing* 37.13 (July 2016), pp. 2990–3004. ISSN: 0143-1161, 1366-5901. DOI: 10.1080/01431161.2016.1192304.

[227] Eric Tzeng et al. “Deep domain confusion: Maximizing for domain invariance”. In: *arXiv preprint arXiv:1412.3474* (2014). URL: https://arxiv.org/abs/1412.3474.

[228] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. “Self-supervised model adaptation for multimodal semantic segmentation”. In: *International Journal of Computer Vision* 128.5 (2020). Publisher: Springer, pp. 1239–1285.

[229] Jakob J. Van Zyl. “The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography”. In: *Acta astronautica* 48.5-12 (2001). Publisher: Elsevier, pp. 559–565.

[230] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. “MTI-Net: Multi-scale Task Interaction Networks for Multi-task Learning”. en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Vol. 12349. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 527–543. ISBN: 978-3-030-58547-1 978-3-030-58548-8. DOI: 10.1007/978-3-030-58548-8_31.

[231] Simon Vandenhende et al. “Multi-task learning for dense prediction tasks: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021). Publisher: IEEE, pp. 3614–3633.

[232] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[233] Kaupo Voormansik et al. “Flood mapping with TerraSAR-X in forested regions in Estonia”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7.2 (2013). Publisher: IEEE, pp. 562–577.

[234] Di Wang et al. “An empirical study of remote sensing pretraining”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2022). Publisher: IEEE.

[235] Di Wang et al. “Samrs: Scaling-up remote sensing segmentation dataset with segment anything model”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[236] Guoli Wang et al. “Feature extraction by rotation-invariant matrix representation for object detection in aerial image”. In: *IEEE Geoscience and Remote Sensing Letters* 14.6 (2017). Publisher: IEEE, pp. 851–855.

[237] Haoran Wang et al. “Classes Matter: A Fine-Grained Adversarial Approach to Cross-Domain Semantic Segmentation”. en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Vol. 12359. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 642–659. ISBN: 978-3-030-58567-9 978-3-030-58568-6. DOI: 10.1007/978-3-030-58568-6_38.

[238] Huiyu Wang et al. “Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation”. en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Vol. 12349. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 108–126. ISBN: 978-3-030-58547-1 978-3-030-58548-8. DOI: 10.1007/978-3-030-58548-8_7.

[239] Jiaqi Wang et al. “Review of large vision models and visual prompt engineering”. In: *Meta-Radiology* (2023). Publisher: Elsevier, p. 100047.

[240] Jie Wang et al. “Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 154 (2019). Publisher: Elsevier, pp. 189–201.

[241] Jingdong Wang et al. “Deep high-resolution representation learning for visual recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.10 (2020). Publisher: IEEE, pp. 3349–3364.

[242] Junjue Wang et al. “LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation”. In: *arXiv preprint arXiv:2110.08733* (2021).

[243] Xiang-Yang Wang, Ting Wang, and Juan Bu. “Color image segmentation using pixel wise support vector machine classification”. In: *Pattern Recognition* 44.4 (2011). Publisher: Elsevier, pp. 777–787.

[244] Ximei Wang et al. “Transferable normalization: Towards improving transferability of deep neural networks”. In: *Advances in neural information processing systems* 32 (2019).

[245] 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). Publisher: IEEE, pp. 98–106.

[246] Yufeng Wang et al. “Boundary-aware multitask learning for remote sensing imagery”. In: *IEEE Journal of selected topics in applied earth observations and remote sensing* 14 (2020). Publisher: IEEE, pp. 951–963.

[247] Shiqing Wei, Shunping Ji, and Meng Lu. “Toward automatic building footprint delineation from aerial images using CNN and regularization”. In: *IEEE Transactions on Geoscience and Remote Sensing* 58.3 (2019). Publisher: IEEE, pp. 2178–2189.

[248] Shunjun Wei et al. “HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation”. In: *Ieee Access* 8 (2020). Publisher: IEEE, pp. 120234–120254.

[249] John Weier and David Herring. “Measuring vegetation (ndvi & evi)”. In: *NASA Earth Observatory* 20.2 (2000).

[250] Darrel L. Williams, Samuel Goward, and Terry Arvidson. “Landsat”. In: *Photogrammetric Engineering & Remote Sensing* 72.10 (2006). Publisher: American Society for Photogrammetry and Remote Sensing, pp. 1171–1178. (Visited on 06/14/2024).

[251] Zuxuan Wu et al. "Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 518–534.

[252] Gui-Song Xia et al. "DOTA: A large-scale dataset for object detection in aerial images". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3974–3983.

[253] Junshi Xia et al. "Openearthmap: A benchmark dataset for global high-resolution land cover mapping". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 6254–6264.

[254] Tete Xiao et al. "Unified perceptual parsing for scene understanding". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 418–434.

[255] 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.

[256] Dan Xu et al. "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 675–684.

[257] Siwei Yang et al. "Reducing the feature divergence of RGB and near-infrared images using Switchable Normalization". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 46–47.

[258] Yanchao Yang and Stefano Soatto. "Fda: Fourier domain adaptation for semantic segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 4085–4095.

[259] Jiafan Yu et al. "DeepSolar: A machine learning framework to efficiently construct a solar deployment database in the United States". In: *Joule* 2.12 (2018). Publisher: Elsevier, pp. 2605–2617.

[260] Qinglie Yuan et al. "Multiscale semantic feature optimization and fusion network for building extraction using high-resolution aerial images and LiDAR data". In: *Remote Sensing* 13.13 (2021). Publisher: MDPI, p. 2473.

[261] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. "A review of deep learning methods for semantic segmentation of remote sensing imagery". In: *Expert Systems with Applications* 169 (2021). Publisher: Elsevier, p. 114417.

[262] Yuhui Yuan, Xilin Chen, and Jingdong Wang. “Object-Contextual Representations for Semantic Segmentation”. en. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Vol. 12351. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 173–190. ISBN: 978-3-030-58538-9 978-3-030-58539-6. DOI: 10.1007/978-3-030-58539-6_11.

[263] Mirko Zaffaroni and Claudio Rossi. “Water Segmentation with Deep Learning Models for Flood Detection and Monitoring”. en. In: (2020).

[264] Daniele Zanaga et al. “ESA WorldCover 10 m 2021 v200”. In: (2022). Publisher: Zenodo.

[265] Friedemann Zenke, Ben Poole, and Surya Ganguli. “Continual learning through synaptic intelligence”. In: *International conference on machine learning*. PMLR, 2017, pp. 3987–3995.

[266] Yang Zhan, Zhitong Xiong, and Yuan Yuan. “RSVG: Exploring data and models for visual grounding on remote sensing data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023). Publisher: IEEE, pp. 1–13.

[267] Jielu Zhang et al. *Text2Seg: Remote Sensing Image Semantic Segmentation via Text-Guided Visual Foundation Models*. arXiv:2304.10597 [cs]. Apr. 2023. URL: http://arxiv.org/abs/2304.10597 (visited on 06/19/2024).

[268] Yu Zhang and Qiang Yang. “A survey on multi-task learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 34.12 (2021). Publisher: IEEE, pp. 5586–5609.

[269] Zhenyu Zhang et al. “Pattern-affinitive propagation across depth, surface normal and semantic segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4106–4115.

[270] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.

[271] Sixiao Zheng et al. “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 6881–6890.

[272] Bolei Zhou et al. “Scene parsing through ade20k dataset”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 633–641.

[273] Manfred Zink et al. “TanDEM-X: The new global DEM takes shape”. In: *IEEE Geoscience and Remote Sensing Magazine* 2.2 (2014). Publisher: IEEE, pp. 8–23.

[274] Yang Zou et al. “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 289–305.

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.