Multimodal Artificial Intelligence Strategies for Remote Sensing Earth Observation

PH.D. SCHOOL
UNIVERSITY OF MILANO-BICOCCA

Department of Informatics, Systems and Communication
PhD Program in Computer Science - XXXVI Cycle

Ph.D. Dissertation of: Mirko Paolo Barbato

Supervisor: Prof. Paolo Napoletano
Co-Supervisor: Prof. Flavio Piccoli
Tutor: Prof. Paola Bonizzoni

This dissertation is submitted for the degree of Doctor of Philosophy

Abstract

Multimodal Artificial Intelligence Strategies for Remote Sensing Earth Observation

Ph.D. Dissertation of: Mirko Paolo Barbato

The study of the land is one of the most relevant tasks for the influence that Earth and the management of its resources have on our lives as individuals and as a society. From the location in which we live to the distribution of the population, the food we consume, the culture, and the socio-relationship between the different societies of the world are partially defined by the characteristics of the surrounding lands. These are the reasons that create our necessity of observing and studying the Earth. These studies have the scope to describe the features of the terrains and can be linked to many tasks, varying from classification, segmentation, estimation of soil characteristics, etc., with the final goal to obtain information that is fundamental in many applications from agriculture of precision to study of land cover and land use. To this end, the use of remote sensing technologies has exponentially increased, consequently enhancing the availability and collection of data. This increment and the use of new technologies open the remote sensing field to two crucial advantages: 1) the possibilities of using AI techniques for Earth Observation and 2) data that not only increase in cardinality but also in the kinds of information that they convey. The former of these opportunities allows for the use of incredibly efficient techniques derived in particular from computer vision that can greatly improve our ability to study Earth. The latter enables us to multimodal strategies. These strategies aim to combine different kinds of data (modalities), such as RGB images, hyperspectral data, LiDAR, etc., to exploit the information that comes from each of them. In many computer vision tasks, multimodal approaches have posed themselves as a new step for a better understanding of reality, thus improving our ability to handle data resources. However, in remote sensing applications, it is still difficult to consider these approaches together with AI techniques due to the lack of datasets that involve both high cardinality and modalities. This thesis wants to analyze and deepen the usefulness of multimodality in remote sensing. With this goal, different tasks that can characterize a remote sensing multimodal application will be investigated, starting from the acquisition of new data

to the study of specific tasks and their integration in a real scenario. In particular, the considered tasks will consist of 1) Hyperspectral Pansharpening for the enhancement of this kind of data; 2) Unsupervised Segmentation with Hyperspectral data; 3) Multimodal Supervised Semantic Segmentation; 4) Digital Soil Mapping for the estimation of soil parameters (such as chemical and texture features). For each of these tasks, the goal will be to demonstrate the usefulness of information that differs from the typical RGB images and the advantages that derive from combining these data using AI techniques. Finally, the knowledge derived from these studies will result in the creation of a real case pipeline for the estimation of the parameters in agricultural areas to help manage the resources. The analysis presented in this work demonstrates that each of these tasks benefits from the use of multimodality, also providing new data and techniques that can support future studies in Earth Observation.

Acknowledgements

Research developed in the context of the project PIGNOLETTO – Call HUB Ricerca e Innovazione CUP (Unique Project Code) n. E41B20000050007, co-funded by POR FESR 2014-2020 (Programma Operativo Regionale, Fondo Europeo di Sviluppo Regionale – Regional Operational Programme, European Regional Development Fund).

The author is also grateful to Prof. Begüm Demir for the valuable support, comments and stimulating discussions.

Contents

List of Figures

List of Tables

Acronyms

ARI	Adjusted Rand Index
ANN	Artificial Neural Network
BDEC	Balanced Deep Embedded Clustering
CA	Cross-Attention
CNN	Convolutional Neural Network
CPE	Channel Patch Embedding
DEC	Deep Embedded Clustering
DEM	Digital Elevation Mapping
DSM	Digital Soil Mapping
DTM	Digital Terrain Model
EC	Early Concatenation
EO	Earth Observation
ERGAS	Error Relative Global Dimensionless Synthesis
FR	Full Resolution
FQNR	Filtered-based Quality with No Reference
HS	Hyperspectral
IDW	inverse distance weighting
INFN	Istituto Nazionale di Fisica Nucleare (National Institute for Nuclear Physics)
LC	Late Concatenation
MS	Multispectral
MTF	Modulation Transfer Function
NMI	Normalized Mutual Information
OLCI	Ocean and Land Colour Instrument
PAN	Panchromatic
PCA	Principle Component Analysis
QNR	Quality with No Reference
RR	Reduced Resolution
RS	Remote Sensing
SAM	Spectral Angle Mapper
SAU	Soil Agricutaral Use
SCC	Spatial Correlation Coefficient
SLIC	Simple Linear Iterative Clustering
STB	Swin Transformer Block
SWIR	Short Wavelength Infrared
TFA	Token Fusion at Attention Level
TPE	Token Patch Embedding
UE	Undersegmentation Error
UQI	Universal Quality Index
VNIR	Visual Near Infrared

Chapter 1

Introduction

The land is the foundational element on which all human activities rely. It affects and differentiates the way human populations live, their distribution on Earth, and all the socio-economic dynamics that characterize our society. The land determines the food available in certain areas, water supplies, forest distribution, etc. In general, any kind of resources available in specific areas depend on the land, also influencing the economy and socio-relationship between countries [1]. Thus analyzing and understanding land characteristics is crucial in many fields, spanning environmental, economic, social, and scientific domains [2]. Earth Observation (EO) has been fundamental in environmental conservation, management of natural resources such as agriculture and food security, and infrastructure development. Most importantly, due to recent years climate changes, it is crucial to have instruments that help to mitigate the phenomenon. In such matters, land studies become one of the most important resources we dispose of for reducing the risk of natural disasters and intervening to reduce the damages [3]. The natural conclusion is that a comprehensive understanding of the characteristics of the land allows us for a better understanding of natural and human-made factors that influence different aspects of our life on Earth. This understanding is essential to balance human needs with environmental sustainability and emergencies, leading to a more resilient and efficient society, and less damage to the environment. These are some of the reasons that inspired the Pignoletto project, a collaboration between the Lombardia region, the Istituto Nazionale di Fisica Nucleare (INFN - National Institute for Nuclear Physics) and the University of Milano-Bicocca. This project aimed to combine knowledge from different departments of the scientific community to better observe, describe, and understand Earth. Pignoletto included the use of different technologies and sensors to extract the characteristics of the soil. From the use of satellites, drones, ground means of transportation, and direct collection on the field, the aim was to gather all the different information and knowledge to improve our efficiency in handling Earth's resources.

In this context, Remote Sensing (RS) is one of the most significant sources of information

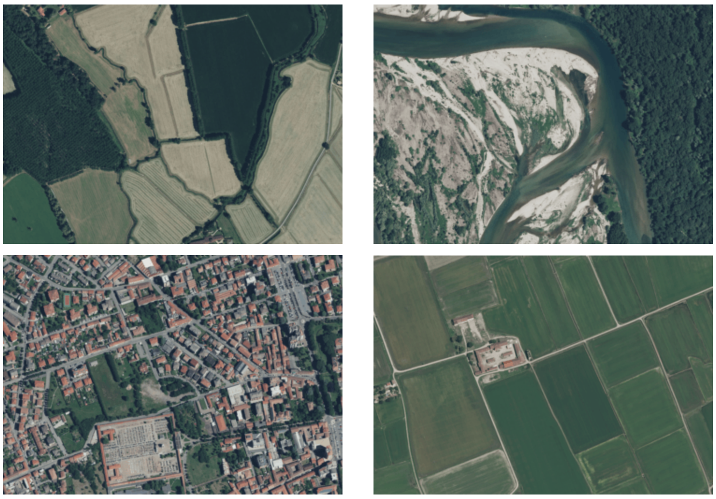


Figure 1.1: Examples of RGB remote sensing images.

that can lead to a full grasp of the land and its properties. RS allows us to use sensors mounted on drones, aircraft, or satellites to acquire images about the Earth's surface, atmosphere, and oceans from a distance, enabling the monitoring and study of our planet's environment and changes over time. Figure 1.1 shows examples of what are the information that RS can help us to collect. These images are firstly acquired by the sensors, then received from a Processing station on the surface, analyzed with the help of software tools and environments, such as QGIS [4], and finally used in different tasks dedicated to the description of Earth. This technology includes many satellites, such as Landsat-8, Sentinel-2, Sentinel-3 [5], ASI PRISMA [6] and many others, that in recent years have revolutionized our ability to observe and analyze our planet from a vantage point above the Earth's surface [7]. In fact, the analysis of data gathered by sensors onboard satellites or aircraft allows to infer useful information about the land, water, and atmospheric systems of the Earth. Thanks to the ease which characterizes the acquisition of new data, this technology has continuously gained more and more importance in several fields, such as agriculture [8], resource exploration [9], environmental monitoring [10], urban planning [11], and disaster management [12]. This new availability of RS data made it possible, even when it comes to EO studies, to exploit the advantages of AI, machine learning and, in particular, deep learning techniques. As already happened in other fields of computer vision, the combination of this technology and its capacity to collect data, and AI methodologies allowed for accurate and high-resolution soil mapping [13], speeding up and making the process of soil characterization [14], segmentation and mapping more scalable.

Following other computer vision tasks, and thanks to the development of RS sensors, different kinds of data (or modalities) that offer complementary information with respect to standard RGB information were introduced, allowing for multimodal approaches to

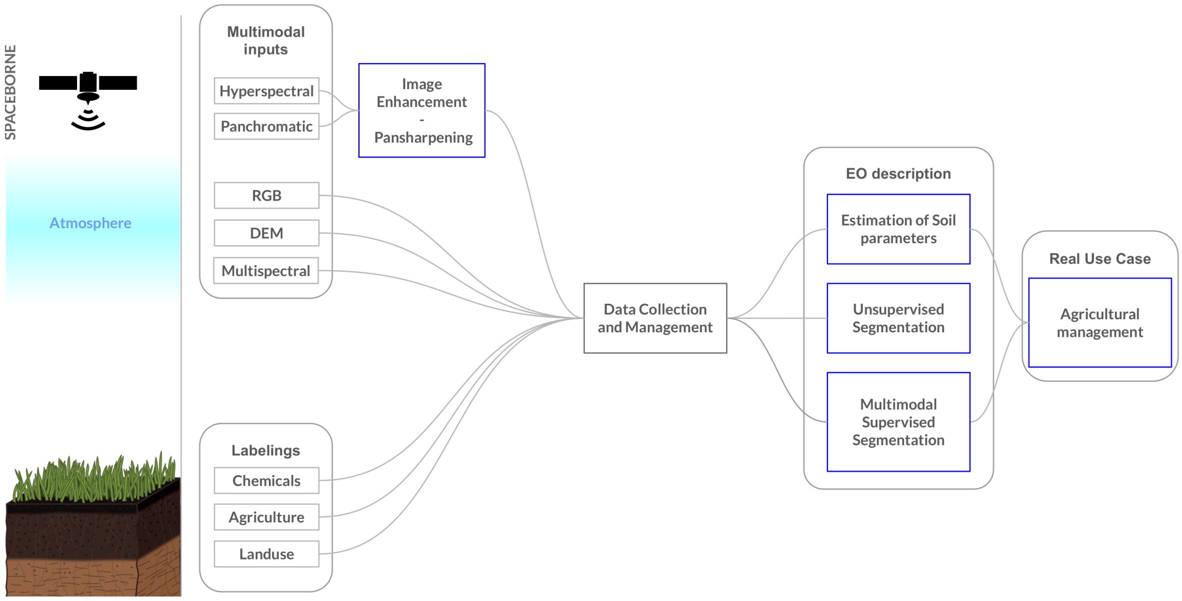


Figure 1.2: Representation of the multimodal remote sensing pipeline investigated in this work. In blue are highlighted the main work on which this thesis focuses.

become an important topic in this field of study [15]. A single sensor may not provide a complete understanding of a scene or object. A multimodal approach, instead, combines data from multiple sensors to overcome these limitations and enhance the analysis, potentially yielding a better description of the data, and better performance in many AI tasks. For instance, these approaches enhance material identification on the Earth's surface [16], therefore enabling more detailed and precise scene understanding, particularly in challenging scenarios. In fact, combining complementary modalities allows for exploiting the best advantages of each information kind, potentially achieving improved performance based on the task and the information. In RS, each modality is often associated with a specific sensor, serving as a distinct information source characterized by its own unique statistical attributes [17]. Apart from the typical RGB modalities, RS can count on complementary information such as LiDAR, multispectral (MS), hyperspectral (HS), panchromatic, etc. [18].

In this thesis, the usefulness of remote sensing and its novel multimodal power will be studied with a focus on land cover, use, and description of the properties of the terrain. The entire analysis, that will be elaborated in this work, will include the complete process that characterizes the use of RS and multimodality, starting from the sources (or modalities) used to the application of these data in a real scenario. Figure 1.2 shows how starting from the sensors and resources of the Earth, it is possible to collect different data, enhance them, and finally use them for different scopes in the EO to better understand the characteristics of the soil and, at the same time, can be combined together to improve our use of the terrain. With this goal, and the possibility to exploit different sources and thus combined information, multimodality and AI can become one of the most powerful tools at our disposal. The main scope is to fulfill EO gaps when it comes to multimodal and AI techniques in the RS field. The main issue that prevents the use of AI and multimodal approaches in RS is mainly the lack of multimodal datasets that do not allow for a proper introduction and study of these techniques and their advantages. This problem will be deepened in this thesis and the usefulness of multimodality will be discussed in many tasks that characterize EO, where the advantages of different kinds of data were analyzed and demonstrated.

Starting from the acquisition and thus the data used, where it was possible, existing datasets were used, in any other case, datasets that could reasonably fit the needs of high cardinality and multimodality were built from scratch. The investigation brought in this work focused on different aspects that characterize RS images and satisfy the possibly divergent necessities of different tasks. In particular, an all-around analysis of spatial, color, spectral and morphological information were considered in this work.

The spatial and color information is the most common modality when it comes to computer vision tasks and this includes even RS. A high-detail representation of these characteristics are the typical RGB images that present three high-resolution bands, covering the visible part of the spectrum (from 400 nm to 700 nm).

The spectral information, in particular, has been proven to be especially useful in many tasks. Nowadays, new sensors and satellites such as ASI PRISMA [6] and Sentinel-3 allow for collecting spectral information in multi- or hyperspectral format. These images consist of 3D image cubes where the first and second dimensions represent the spatial information, while the third one represents the spectral information. As shown in Figure 1.3, the difference between MS and HS images is that MS usually has a big grain spectral resolution that ends up in images with a limited number of bands usually less than 13), while HS has a really high spectral resolution which results in having images with a high number of bands (sometimes even more than one-hundred). They allow for a more precise differentiation of materials, that typically have spectral signatures with different characteristics [19]. For instance, in RS, this can help to better distinguish between buildings, cultivations, rivers etc. [10]. For these reasons, the spectral information helps improve accuracy and efficiency in many tasks spanning classification, segmentation, estimation of soil parameters, etc.

Nevertheless, these types of images are characterized by a significant trade-off between spatial resolution and the number of bands. The trade-off is mostly driven by the costs associated with launching a satellite into space. Consequently, it is necessary to carefully balance the selection of sensor design and components. This limitation, elaborated upon in chapter 3, plus the requirement to minimize energy usage, result in a disparity between spatial and spectral resolution. As a consequence, spectral images, especially HS, exhibit a significantly reduced spatial resolution. For these reasons, combining the information of high-spatial-resolution images, such as RGB, with spectral information can be crucial in

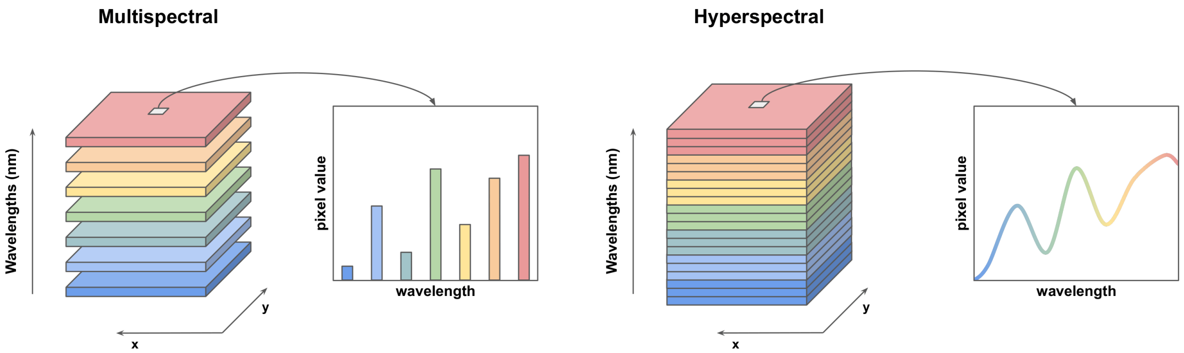


Figure 1.3: Difference between multispectral and hyperspectral signals. Differently from multispectral, hyperspectral presents a continuous signal for every pixel.

many tasks, like for instance semantic segmentation.

In this thesis, the morphological structure of the terrain is mainly described as Digital Elevation Model (DEM) and Digital Terrain Model (DTM). Usually, DEM is used to generally describe the elevation of soil and generated from Synthetic Aperture Radar interferometry [1], while DTM can be considered as a specialized version of DEM that usually describes the natural terrain elevation, thus being both extremely important in characterizing the relationship between different portions of the terrain and, together with other information, immediately identifying features of specific kinds of grounds.

Considering and focusing mainly on these kinds of data, in this thesis, different remote sensing tasks were addressed to demonstrate that, in each of them, specific kinds of data and combinations achieve better performance. In this work an analysis of these modalities is presented, including the creation of novel approaches to handle these data on different EO tasks. As mentioned, an important part consists of acquiring data from different complementary sources and harmonizing them with proper labeling based on the pursued task. Both inputs and labelings require refinement to make them compatible and increase the quality of the raw data. Once the data is acquired and refined it is possible to work on EO tasks that allow for a better understanding of the Earth and, finally, this comprehension can be transposed to real use cases that help in the management of the resources we dispose of. In this thesis, four main fields of EO are addressed covering the use of multimodality in the entire pipeline that goes from acquisition to application. Finally, the EO investigated are combined to demonstrate their usefulness in real use cases. These fields of study will be described in each of the following chapters as follows:

- Hyperspectral Pansharpening [20];
- Unsupervised Segmentation of hyperspectral images [10];
- Multimodal Supervised Semantic Segmentation [21, 22];

[1] DEM generation: https://step.esa.int/docs/tutorials/S1TBX DEM generation with Sentinel-1 IW Tutorial.pdf

- Digital Soil Mapping [23];
- Multimodality in a Real Scenario - Estimation of Soil Parameters for Agricultural Areas Management [24].

Each of the topics provides a different insight into the modalities, demonstrating the advantages of combining them and which modalities are better to be fused. Starting with describing the characteristics of the soil and ending with a semantic comprehension of the different areas in an RS image, each topic analyzes the usefulness of specific data, their advantages and disadvantages, how to reduce the influence of these disadvantages, and finally, the potential of combining them together exploiting their complementarity.

The Hyperspectral Pansharpening consists of fusing a PAN and an HS images to enhance the spatial resolution of the spectral cube, bringing the HS image to reach the same spatial resolution as the PAN image. A panchromatic image is a high-spatial-resolution gray-level image. If this single-band image shares the same geographical area, the same time of acquisition as the spectral data (HS or MS) and consequently the same content, then a new image can be obtained by fusing the spatial information of PAN with the spectral information of HS. This new pansharpened image will have the same number of bands as HS and the same m/px as PAN. This well-known strategy is really important in RS because it allows for compensating the low spatial resolution of spectral images, mainly caused by the high cost of the sensors. However, even if well known, the researchers in this field focused on MS data that as mentioned above has less potential in describing the spectral characteristics of an area. The main problem is relative to the lack of numerous HS datasets, whereas, on the other side, the availability of MS datasets is higher. This means that the state-of-the-art pansharpening techniques usually concern MS and when it considers HS data, it usually revolves around machine learning techniques or adapted deep learning techniques which are not fully capable of expressing the real advantages of these methodologies and cannot define a statically valid evaluation. In this thesis, a comprehensive analysis and adaptation of hyperspectral pansharpening techniques in the state of the art is reported. This analysis is based on a novel hyperspectral dataset, built and presented in this work, which allows for proper use of deep learning techniques and a statistically valuable evaluation of the methods. Chapter 3 describes in detail how the dataset was built and how all the methodologies were adapted to different data than MS.

The second main task described in this thesis is Unsupervised Segmentation. The Segmentation is fundamental for the description of the land use and cover because it allows to identify each pixel of an image with a specific class. The Unsupervised Segmentation, in particular, consists of making a prior analysis of the terrain and segmenting the area into different classes where the number of these classes and the real labels are not known. In segmentation and generally in EO, this is a really important task because creating maps and labels for supervised methods is a really complex, effortful and time-demanding manual work. Being able to provide a fast and efficient unsupervised analysis can make the work of creating new ground truths easy, thus improving the quality and speeding up the entire procedure of producing semantic segmentation labels. In this chapter 4, the HS information is used to provide a solid unsupervised segmentation technique, demonstrating at the same time the usefulness of this kind of data when it comes to semantic segmentation and providing important hints on how to work and adapt the algorithms to challenging data such as HS.

In chapter 5, the Multimodal Supervised Semantic Segmentation topic combines all the main modalities considered in this thesis and different fusion methodologies in a comprehensive work where a multimodal dataset called Ticino is presented to become a first step for any other research in this field. To the best of the author's knowledge, the proposed Ticino dataset is the biggest and most diverse dataset in terms of covered area for RS semantic segmentation, also including a high cardinality of images for all the modalities and presenting an HS source of information. This dataset includes spatial (RGB), spectral (HS) and morphological (DTM) information on a wide area in the South of Milan and wants to fill the lack of multimodal datasets with hyperspectral information in the state of the art. It also presents two pixel-wise semantic labelings, one dedicated to Land Cover and one to Soil Agricultural Use (SAU). The proposal of Ticino also comes with a benchmark on the dataset based on CNN, where different combinations of multimodal and single-modal are compared, demonstrating the advantages of multimodal even in the RS segmentation task. To further observe the impact of multimodal and fusion techniques on RS semantic segmentation, a comprehensive analysis of multimodal fusion methods with the use of a Transformer is also presented in this thesis. The main purpose of this work on Multimodal Supervised Semantic Segmentation is to show the real influence that a fusion method has on the resulting segmentation. Moreover, the utility of multimodal approaches is investigated in comparison with RGB, considering not only the performance but also the complexity of the different methods in terms of memory and computation. Once again, multimodal approaches are proven to be preferable.

Digital Soil Mapping involves the estimation of the parameters of the soil and the creation of different maps relative to each parameter. The estimation of these parameters is a regression task that consists of approximately esteeming texture and chemical characteristics of the terrain. These parameters can vary in nature. For instance, when it comes to texture characteristics, clay, silt, sand and coarse become important to estimation because they play a determining role when it comes to behaviors such as water-holding capacity, drainage characteristics, nutrient retention, and susceptibility to erosion, influencing plant growth and agricultural productivity [24]. Different plants have specific pH requirements for optimal growth. Chemical characteristics, such as the presence of specific substances like K, N, CaCO3 and pH can effectively help in soil and resource management, being particularly interesting in the agricultural field. For example, pH affects nutrient availability and microbial activity in the soil, giving important information on soil health. Not only this task is really important in EO and can benefit from RS technology, but it is also an interesting benchmark to understand the possibility and the potential of multimodal approaches. The investigation of this task, in this thesis, focused on the use of MS and DEM information, comparing the use of MS-only with the combination of the two modalities. The analysis, detailly described in Chapter 6, demonstrates that RS and AI together achieve good performance, and, in particular, multimodal approaches utterly improve the quality of the estimation. More importantly, it also demonstrates how including DEM information can guarantee a better description and understanding of the soil.

Finally, in chapter 7, these investigations will be combined to reach the goal of a real use case. In this chapter, a possible real strategy will be presented and discussed with the specific aim of estimating the necessities of agricultural areas and thus improving the management of resources. Even in this case, the advantages of multimodality over single modality will be reported to empower the importance of this field of study.

The main contributions of this dissertation are as follows:

- a study of multimodality for Digital Soil Mapping on the widest and variegated area in the state of the art;
- an original method of unsupervised segmentation using hyperspectral images that helps the creation of ground truth without the necessity of a-priori knowledge of the scene;
- an analysis of deep learning pansharpening techniques for hyperspectral data with a new pansharpening dataset that provides the highest cardinality and variety of scenes from all the World thus being statistically relevant for AI strategies;
- a multimodal dataset for semantic segmentation that includes a wide area of interest, a cardinality suitable for AI strategies and a total of five modalities including HS and DTM for the study of multimodal strategies;
- a study of different CNN and Transformers fusion methods for semantic segmentation that involve the introduction of original strategies was introduced;

Chapter 2

State of art of datasets, algorithms, and methods

As mentioned above, many factors affect RS, multimodality and AI techniques. In particular, AI, with machine learning and even more with deep learning, has many advantages but also flaws that prevent the use of such techniques. First of all, one of the most problematic issues is the necessity of large and diversified amounts of data to train a model and make it able to generalize a task in different contexts. When it comes to multimodal approaches and RS, this translates into many issues that depend on the specific tasks. Other difficulties when it comes to AI are represented by the engineering and design of the techniques used. Especially in multimodal approaches, methods can vary in how they extract and use data from different kinds of sources, combining them with strategies that can also vary based on the data and influencing the performance and results.

In this chapter, the state of the art related to RS will be deepened. It will be divided into sections, one for each topic of the pipeline described in Chapter 1. For each of them, the main datasets and techniques will be illustrated focusing on the results already achieved in the state of the art and challenges still open.

2.1 Hyperspectral Pansharpening

The field of pansharpening has made significant advancements thanks to data-driven approaches and novel methods. This section focuses on the current state of the art in pansharpening, highlighting two important aspects: the development of benchmarks for a comprehensive evaluation and the design of novel methods that effectively improve spatial resolution while maintaining spectral fidelity. By reviewing these benchmarks and methods, a comprehensive overview of the most recent advancements is provided.

Table 2.1: List of existing datasets used for RS image pansharpening. Here are reported, for each dataset, the number of images of the dataset, the number of bands, and the coverage in terms of wavelength. The image resolutions reported in this table are taken from the original dataset descriptions.

Dataset	Cardinality	Images resolution	Type	# of bands	Wavelength coverage
Pavia University [25]	1	610 × 610	airborne	103	430 - 838 nm
Pavia Center [25]	1	1096 × 1096	airborne	102	430 - 860 nm
Houston [26, 27]	1	349 × 1905	airborne	144	364 - 1046 nm
Chikusei [28]	1	2517 × 2335	airborne	128	363 - 1018 nm
AVIRIS Moffett Field [29]	1	37 × 79	airborne	224	400 - 2500 nm
Garons [29]	1	80 × 80	airborne	125	400 - 2500 nm
Camargue [29]	1	100 × 100	airborne	125	400 - 2500 nm
Indian Pines [25, 30]	1	145 × 145	airborne	224	400 - 2500 nm
Cuprite Mine	1	400 × 350	airborne	185	400 - 2450 nm
Salinas [25]	1	512 × 217	airborne	202 (224)	400 - 2500 nm
Washington Mall [31]	1	1200 × 300	airborne	191 (210)	400 - 2400 nm
Merced [26]	1	180 × 180	satellite	134 (242)	400 - 2500 nm
Halls Creek [32]	1	3483 × 567	satellite	171 (230)	400 - 2500 nm

2.1.1 Datasets

The choice and characteristics of the benchmark datasets play a crucial role in evaluating and comparing different algorithms for RS pansharpening. Each dataset has unique properties that differentiate it from others, including cardinality, image resolution, acquisition setup, number of spectral bands, and wavelength coverage. For an immediate comparison of the existing datasets in the state of the art, please refer to Table 2.1 which summarizes all of them and their properties.

It is possible to divide the existing datasets into different groups by mainly considering three properties: the wavelength coverage, image resolution, and acquisition setup of each dataset. Regarding the wavelength coverage, four datasets range from visible to near-infrared (VNIR), while the remaining ones cover the entire spectrum, from visible to short-wave infrared (SWIR). The use of data with limited spectral coverage for the design of pansharpening algorithms could limit their applicability to real-case scenarios, which may require the use of bands and data not covered by those datasets.

Another important aspect is the image resolution, associated to the dataset cardinality. While datasets like Halls Creek [32] can potentially be tiled in smaller samples for training or validation purposes, the other ones are limited due to the low cardinality and low resolution of the data. Furthermore, even if an image is tiled, the variety of the content of the scenes considered is limited to the area covered by the single image, making it hard to evaluate algorithms in different scenarios. Finally, most of the datasets in the state of the art are tagged as “airborne” type, which means that are collected by using airplanes or low-altitude flying devices, while only two are made of satellite-collected images.

2.1.2 Pansharpening Methods

Pansharpening methods can be grouped into five categories: component substitution (CS), multiresolution analysis (MRA), Bayesian, matrix factorization (as defined by Loncan et al. [29]) and deep learning.

Component substitution methods consist of substituting the spatial component of the spectral images with the high-resolution panchromatic images. The results are obtained by projecting a high-resolution version of the spectral image into its spatial component and then reverting the transformation using the panchromatic information instead of the spatial projection extracted. CS includes methods such as principal component analysis (PCA) [33], intensity-hue-saturation (IHS) [34], Gram-Schmidt (GS) [35], and GS Adaptive (GSA) [36]. These methods are usually easy to implement, achieve high spatial fidelity, and are robust to misregistration, but can create significant spectral distortions [29].

In multiresolution analysis, Loncan et al. [29] includes methods such as Decimated Wavelet Transform (DWT) [37], Undecimated Wavelet Transform (UDWT) [38], "à-trous" wavelet transform (ATWT) [39], and Laplacian pyramid [40], which consist of using a filtered version of the PAN signal, to extract high-resolution details and inject them into the spectral image. Compared to the CS methods, the MRA techniques are more difficult to implement and computationally more complex but also allow for achieving a better spectral consistency with the original spectral information.

Some other approaches, for instance, Guided Filter PCA [41], combine the two techniques to gather the advantages from both, but the results on hyperspectral images were not promising, being the technique with the worst results on hyperspectral data in Loncan et al.'s investigation [29].

Bayesian approaches are based on the estimation of the posterior probability of the full-resolution image that would be obtained considering the original panchromatic and spectral information. These methods typically consider the sensor characteristics to enhance the resolution, thus achieving state-of-the-art performance but also being less generalizable and more complex to use [29].

Matrix factorization techniques are described by Loncan et al. [29] as the only ones purposely used for hyperspectral pansharpening, and instead of using the panchromatic information, use a high-resolution multispectral image to convey the spatial information of the hyperspectral data into a higher resolution space. Even in this case, to exploit the best factorization to reconstruct the new image, the characteristics of the sensors are taken into consideration, making them less viable compared to other methods.

Deep Learning approaches, due to the recent success of the Neural Networks, have been proposed in recent years and using different models for multispectral pansharpening. [42]. Masi et al. [43] in 2016 propose the first Convolutional Neural Network designed for multispectral pansharpening. They presented a simple neural network made of few convolutional layers, capable to outperform the machine-learning-free approaches on three standard datasets. In 2018, Scarpa et al. [44] tried to compensate for the lack of data, proposing the Adaptive-PNN strategy by fine-tuning the model extracting samples from the reference image, and refining the pansharpened reconstruction by closing the gap between training and testing in pansharpening. Yang et al. [45] proposed a new model with the focus on spectral information preservation, which operates mainly on the high-frequency components of the multispectral images, while trying to keep low-frequency information as much as possible unaltered. In 2018, Yuan et al. [46] proposed a multi-branch network, being the first approach to fuse PAN and MS images in feature domain and reconstruct the pan-sharpened image from the fused features, instead of approaching the problem as a super-resolution task. In 2020, Liu et al. [47] presented a multi-resolution based approach for image fusion, based on wavelets decomposition, which represents the first approach to explicitly perform a deep fusion operation between the panchromatic information and the multispectral bands. In the same year, Cai et al. [48] proposed a method for progressive pansharpening, while recently, in 2021, Xie et al. [49] proposed a progressive PAN-injected fusion method based on super-resolution. This last approach extracts information from the panchromatic image with dedicated encoder branches, from both low and high frequencies, in order to better exploit features from the panchromatic image, achieving state-of-the-art results.

Some attempts at deep learning approaches for hyperspectral data have also been investigated. In 2019, He at al. [31] proposed HyperPNN, a phases CNN that firstly extracts spatial features from the panchromatic image and spectral features from the hyperspectral one, secondly fuses the spatial and spectral features with dedicated convolutional layers, and thirdly predicts the spectral information of the pansharpened image with convolutional layers that focus only on the spectral signatures. This model has been followed in 2020 by the improved version called HySpecNet [32]. In the same year of HyperPNN, Zheng et al. [50] investigated the use of the residual network for pansharpening, firstly guiding the upscaling and enhancing the edge details of the hyperspectral data with Contrast Limited Adaptive Histogram Equalization (CLAHE) and a guided filter to fuse the image with the panchromatic information, and then using a Deep Residual Convolutional Neural Network (DRCNN) to boost the reconstruction. Xie et al. [51] developed the HS Pansharpening method with Deep Priors (HPDP), exploiting the power of different deep learning modules to improve all parts of the pansharpening pipeline. In particular, they used a Super Resolution Deep Learning (SRDL) module to upscale the HS image and fuse it with the panchromatic information by also considering high-frequency information extracted by the proposed High-Frequency Net (HFNet). They finally obtained the final high-resolution HS by injecting the high-frequency structure in the upscaled HS, using a Sylvester equation. It is worth noticing that they used multispectral images for training to compensate for the limited number of training samples. Recently, in 2023, He et al. [26] proposed dynamic hyperspectral pansharpening that uses a learn-to-learn strategy to adapt the pansharpening to the spatial variations of an image.

Despite the increased adoption of CNNs and deep learning in the field of pansharpening, and the increased interest in the use of hyperspectral images for satellite image analysis, the limited number of hyperspectral samples is still an issue. In order to study the impact and the possible advantages of the application of deep neural models on hyperspectral pansharpening, these models are adapted and retrained to hyperspectral data, collected from the PRISMA dataset, and compared with a subset of machine-learning-free approaches.

2.2 Unsupervised Segmentation of hyperspectral images

Methods belonging to the state of the art for unsupervised and semi-supervised hyperspectral image segmentation should take into consideration both the spectral and the spatial information to avoid noisy results [52, 53]. Depending on the order by which these two information are addressed, it is possible to define a taxonomization of the methods composing the state of the art. Table 2.2 shows pros and cons of each method of the state of the art.

Spatial regulation methods Audebert et al. [54] perform in first place a per-pixel segmentation followed by a spatial regularization done with a context-dependent criteria. In this context, Wu et al. [55] propose a Laplacian Support Vector Machine (LapSVM) to classify each pixel and then use a Conditional Random Field (CRF) to regularize the results according to the surrounding of the pixel under consideration.

Pre-segmentation methods perform a first step of spatial regularization as an unsupervised pre-processing and then aggregate the spectral features for each segmented region to enforce spatial consistency. The spatial regularization is enforced through clustering or superpixels [56]. Gillis et al. [57] designed a fast hierarchical clustering algorithm for hyperspectral images (HNMF). The algorithm uses a rank-two nonnegative matrix factorization to split the data into clusters. This approach showed effectiveness on synthetic and real-world HS images, outperforming standard clustering techniques such as k-means, spherical k-means, and standard NMF. Zhang et al. [58] extend the SLIC algorithm for HS image segmentation. The authors state that the reduction of the spectrum, in their case, degrades the performance. Visual attention mechanisms can also be used to better highlight the salient parts [59, 60].

In this context, the compression of the spectral signal with dimensionality reduction techniques such as the Principal Component Analysis (PCA), helps to decrease the size of the problem and to improve the overall performance. In 2017, Zhang et al. [53] proposed a multiscale superpixel representation starting from the first principal component. Using this representation, a multiscale classification is achieved and then fused using a majority voting to exploit the final labels. Similarly, in 2019, Zhang et al. [61] used the Entropy Rate Superpixel segmentation and a kernel-based extreme learning machine. Starting from the first principle they successfully combined the spectral and spatial information with performance improvement over other spectral approaches. Zu et al. [62] took advantage of the spatial information combining a band reduction technique with SLIC segmentation and a feature extraction based on principal components. This approach demonstrated to achieve results comparable to other techniques with few labeled samples.

Self-supervised methods use autoencoders to learn a compact representation of the input data [63]. To learn a meaningful representation, those methods usually use a normalization term [64] or add noise during training [65]. Chen et al. [66] and Abdi et al. [67] use stacked autoencoders to create a latent space of lower dimensionality through the use of a normalization term. Similarly, Xing et al. [68] use stacked autoencoders but during training time they add noise to the embeddings and treat the problem as a denoising task. Nalepa et al.[69] introduce dependency among samples through the use of 3D convolutional autoencoders. Zhang et al.[70] use an information fusion network that combines hyperspectral images with light detection and ranging data (LiDAR). An autoencoder is trained to reconstruct both signals in a self-supervised way. The intermediate representation is then used by a two-branch CNN for final classification. Paul et al. [71] use a U-net architecture along with spectral partitioning. The proposed architecture is called HyperUnet. Tulczyjew et al. [72] propose the use of an asymmetric autoencoder based on recurrent neural networks to address the low cardinality and imbalance that is typical of HS image datasets. Chen et al. [73] use adversarial training to fill the lack of samples. As stated by Wambugu et al. [74], the generation of synthetic samples through data augmentation can improve robustness.

Graph theory is also broadly used in this context to leverage spatial relationships. In this context, Aletti et al. [75] propose a semi-supervised method that uses a random walker method to perform segmentation. Ding et al. [76] use a graph neural network (GNN) with autoregressive moving average filters for leveraging structures present in the HS images. Similarly, Luo et al. [77] use GNN in combination with a multi-structure unified discriminative embedding to enforce spatial consistency.

Joint learning methods attempt to learn simultaneously spatial and spectral features. Zhang et al. [78] developed a classification framework using gradient-fusion of bands combined with watershed superpixel segmentation to convey contextual information and spatial dependencies. By doing so, the classification will be less sensitive to noise and segmentation scales. Murphy et al. [19] proposed an unsupervised spectral-spatial diffusion learning technique (DLSS) that combines spectral and spatial information considering the modes of classes. This active learning strategy can be helpful in those contexts where the hyperspectral information varies along time and the system must adapt to new data. Xie et al. [79] proposed a Deep Embedded Clustering (DEC) to simultaneously learn feature representations and cluster assignments using deep neural networks. DEC maps the data space to a lower-dimensional feature space optimizing a clustering objective. The experimental evaluations on state of art images showed significant improvement. Starting from this approach, Obeid et al. [80] proposed a balanced version of DEC (BDEC). In particular, they developed an additional search and extraction step to balance the data before the training of DEC, making use of a-priori knowledge of the context of data and labeling, to further improve the overall quality of the segmentation on a variety of state-of-the-art datasets. They compared their method with other clustering techniques such as k-means [80], Gaussian mixture model (GMM [80]), and sparse manifold clustering and embedding (SMCE [81]).

2.3 Multimodal Supervised Semantic Segmentation

This section, as for pansharpening, is divided into two parts, one dedicated to the datasets and one to the semantic segmentation methods in computer vision and RS. The first part is particularly important to show the lack of RS multimodal datasets in the field of RS semantic segmentation, especially when HS information is considered. It describes the state-of-the-art datasets in detail for a better understanding of their lacks and weaknesses. The second part focuses on showing how even with a scarcity of multimodal datasets properly built for segmentation, this field is still one of the most studied and challenging.

2.3.1 RS Datasets for semantic segmentation

Remote sensing data have greatly increased in numerosity, modality, and variety. Nonetheless, due to labeling difficulties, when it comes to semantic segmentation, there is not the same quality of datasets available in other tasks like classification [82]. However, several datasets such as the TorontoCity dataset, the ISPRS 2D semantic labeling dataset, the Mnih dataset, the SpaceNet dataset, and the ISPRS Benchmark for Multi-Platform Photogrammetry have been proposed in the literature for semantic segmentation [83].

Three bands datasets Focusing on RGB images, one of the most important datasets is Deepglobe [83]. This dataset covers an area of 1716.9 km^2 that includes Thailand, Indonesia, and India. It is composed of 1156 images with a resolution of 0.5 meters per

Table 2.2: Pros and cons of the methods representing the state of the art.

Method	Pros	Cons
[67]	The architecture is very simple and easily replicable	The use of NNs instead of CNNs makes the system more data-hungry
[75]	Consensus-based methods show more robustness to noise	Dataset-tailored similarity index can lead to misjudgement
[66]	Self-supervised approach helps to deal with small amount of data	The size of the neighbor region has a huge impact on the performance of the system and it is dataset-dependent
[73]	The local manifold learning helps to discover relationships among samples	Adversarial prediction can face a partial or a total mode collapse
[76]	Graph neural networks can be effective in describing structures	Requires a great amount of labeled data that often is not available
[81]	Through the use of sparse coding the method is robust to data nuisances such as noise and outliers	the refactorization of each sample as a linear combination of remaining samples can be misleading in presence of many outliers
[57]	Nonnegative matrix factorization can be a powerful splitting technique	Requires to know in advance the number of classes
[52]	Only five parameters required and a moderate number of training samples	High computational time both in training and testing due to the construction of the similarity graph
[77]	Intraclass and interclass neighborhood structure graphs can help to improve the description in the feature space of the HS images	It's unclear how much is the contribution of the tangential structure information on the final performance
[19]	The proposed algorithm can be complemented with few real samples, boosting the performance in an active learning fashion	Based on the assumption that different classes have different densities
[69]	Representation learned in a self-supervised fashion	3D convolutions introduce in-batch samples dependency
[80]	Extremely fast (up to 2600X w.r.t. [79]). Address the problem of data imbalance	The extraction of data subsets that are equally representative can lead to misrepresentation in presence of low cardinality classes
[56]	Fast and simple as the computation is limited to neighbors	Generalization limited by bag of words which must be small to avoid a performance drop
[55]	Pixelwise classification is simple and fast	low resolution of hyperspectral images leads to pixel misclassification
[79]	Feature representation is learned during the process. Less sensitive to hyperparameters	Kullback-Leibler divergence minimization can lead to errors when the auxiliary distribution has low cardinality
[68]	Very simple architecture	Poor results
[78]	The proposed fusion method requires less samples with respect to other methods	The use of Local Binary Patterns (LBPs) heavily affects the performance as the scale changes. This is due to the limited field of view of LBPs.
[53]	The multiscale approach avoids the choice of the optimal superpixel size	Major voting strategy considers all scales equally
[58]	Exploits the consolidated SLIC algorithm to define superpixels, extending it to hyperspectral information	It is semi-supervised and requires some labeled samples to propagate their label to the pixels in the same superpixel
[61]	Captures local as well as global spatial characteristics of the HS images	Extreme learning machine (EML) uses single hidden layer feedforward neural networks (SLFNs), whose representativity is low
[70]	Improves performance by integrating LiDAR data	requires LiDAR data
[62]	Lower dimensionality promotes meaningfulness of feature vectors	Superpixel-independent dimensionality reduction through robust PCA can lead to non-comparable feature vectors

pixel and 2448x2448 pixels per sample. The labeling consists of 7 classes for land cover and land use segmentation: urban, agriculture, rangeland, forest, water, barren, and unknown.

The TorontoCity dataset [84] includes RGB and LIDAR information, focusing on building footprints and road segmentation and covering an area of 712.5 $km2$ in Toronto with a resolution of 0.10 meters per pixel. Some RGB datasets as a variant of the SpaceNet dataset [85] and the INRIA aerial dataset [86] focus on the binary segmentation of buildings and non-buildings. The former considers only the RGB components of the standard SpaceNet dataset [87]. It wants to be an alternative and more approachable dataset than the original SpaceNet. It consists of 300x300 pixels per image, divided into a training set of 280741 images, a validation set of 60317 images, and a test set of 60697 images, focusing on the segmentation of building and non-building classes. The latter dataset presents 360 images of 1500x1550 pixels that cover an area of 810 km^2, with a resolution of 0.3 meters per pixel.

Another RGB dataset is the Urban dataset from the Campinas region in Brazil [88]. This dataset was created in 2003 including three different urban classes that consist of residential, commercial, and industrial areas. The other regions were all included in a non-urban area which included highways, roads, native, vegetation, crops, and rural buildings. The dataset is composed of 9 images of 1000×1000 pixels and a spatial resolution of 0.62 meters per pixel. The same article also presents the Coffee dataset [88]. It considers images of 3 bands using the NIR-R-G part of the spectrum instead of the classical RGB. Taken by the SPOT sensor in 2005 over Monte Santo de Minas in Brazil, the dataset focuses on the detection of manually segmented coffee crops and, even in this case, is composed of 9 images. Each image has 1000x1000 pixels with a spatial resolution of 2.5 meters per pixel.

Multispectral datasets To better extract information from the spectrum, other datasets consider multispectral modality alone or together with other source types. The Zurich Summer dataset [89] consists of 20 images by the QuickBird satellite in 2002. The data include four bands in the NIR-RGB part of the spectrum, an average size of 1000x1150 pixels, and a spatial resolution of 0.61 meters per pixel achieved after pan-sharpening. The labeling represents eight urban classes: roads, trees, bare soil, rails, buildings, grass, water, and pools.

Multimodal datasets with multispectral information SpaceNet [90, 87] consist of images from different sensors: WorldView-1, WorldView-2, WorldView-3, WorldView-4, and GeoEye-1 [91]. Each sensor presents a variety of data and the most complementary between them is the WorldView-3 which includes a panchromatic image and 8 multispectral data, respectively in the VNIR and SWIR portions of the spectrum. Spacenet presents tiles of 666x666 with a resolution of 0.3/0.5 meters per pixel for the panchromatic data, depending on the sensor. In this case, the original multispectral images have different resolutions in a range from 1.24 to 1.85 meters per pixel but are upscaled using the pansharpening and in their final versions share the same resolution of the panchromatic component. It covers different cities and presents various kinds of segmentation depending on the type of task aimed.

Other multimodal datasets are 2D Semantic Labeling Potsdam and Vaihingen datasets [92] that present both multispectral/RGB and Digital Surface Model (DSM) information (another specialization of the Digital Elevation Model which characterizes the heights of the surface, including artificial and natural elements). They respectively have 38 and 33 images of 6000x6000 and 2000x2500 pixels [90]. Potsdam also involves more versions of the same ground tiles. It includes two three-band images in the RGB or the IR-RG part of the spectrum and a third multispectral image of 4 bands with IR-RGB information that comprehends all the spectral information of the dataset. Vaihingen instead presents only RGB information when it comes to the spectrum. The spatial resolution is different for the two datasets, with Potsdam with a high resolution of 5cm and Vaihingen with a resolution of 9cm. The two datasets present a labeling that includes six classes. The classes involve impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background.

Another multi-source dataset that includes multispectral information is the DSTL dataset [93]. The dataset consists of 1km x 1km satellite samples from the WorldView-3 sensor. For each sample, it includes an RGB image from Deepglobe [83], a one-band panchromatic image, an eight-band multispectral image with NIR and visible information from 400 to 1400nm (red, red edge, coastal, blue, green, yellow, near-IR1, and near-IR2), and an eight-band multispectral image in the short wavelengths part of the spectrum (from 1195 to 2365nm). Each source has different spatial resolutions. Panchromatic image is 0.31 meters per pixel, multispectral 1.24m, and Swir 7.5m. The dataset is built to identify 10 classes: buildings (large building, residential, non-residential, fuel storage facility, fortified building), misc (Manmade structures), road, track (poor/dirt/cart track, footpath/trail), trees (woodland, hedgerows, groups of trees, standalone trees), crops (contour ploughing/cropland, grain crops, row with potatoes and turnips), waterway, standing water, large vehicle, small vehicle.

Hyperspectral datasets Another group of datasets represents hyperspectral data that possess the highest spectral resolution compared to any other sources. Typically, these datasets include single images, so they are not suitable for standard modern deep learning techniques due to a low cardinality and variety of data. Some of the most popular datasets are the Indian Pines [30], Salinas, SalinasA, Pavia Center, and Pavia University datasets [25].

The Indian Pines by AVIRIS sensor consists of 145x145 pixels and 224 spectral bands in the 400–2500 nm wavelength range and a spatial resolution of 3.7 meters per pixel.

The final number of bands is reduced to 200 by removing the region of water absorption bands. The ground truth available includes sixteen classes: alfalfa, corn-notill, corn-mintill, corn, grass-pasture, grass-trees, grass-pasture-mowed, hay-windrowed, oats, soybean-notill, soybean-mintill, soybean-clean, wheat, woods, buildings-grass-trees-drives, stone-steel-towers.

Salinas and SalinasA [25] have also been acquired by the AVIRIS sensor and present 224 bands in the 400–2500nm portion of the spectrum, like Indian Pines. Even in this case, the final datasets have 204 bands because the 20 noisy channels in the region of water absorption have been discarded. The spatial resolution is 3.7 meters per pixel, and the two images are respectively of 512x217 and 86x83 pixels. In particular, the SalinasA dataset represents a subset of the Salinas dataset. Consequently, the labeling is different between the two datasets. Salinas is annotated with 16 classes representing the region of cultures such as broccoli, fallow, grapes, etc., while SalinasA, being a subset of the Salinas labeling, includes only six classes.

Pavia Center and Pavia University datasets [25] by the ROSIS sensor are respectively 1096x1096 and 610x610 images with 102 and 103 channels. In both cases, part of the samples has been discarded because of missing information, resulting in two images respectively of 1096x715 and 610x340 pixels with a spatial resolution of 1.3 meters per pixel. The labeling includes nine labels for both datasets, representing typical classes of land cover such as asphalt, meadows, trees, bare soil, etc.

The use of different modalities can achieve better performance [94] when it comes to remote sensing semantic segmentation, but due to the problem of providing semantic labelings [10], the existing remote sensing datasets for semantic segmentation usually consist of single modality or sources that possess less discriminative power such as multispectral instead of hyperspectral. These are the main reasons to push the creation of a dataset capable of exploiting the advantages and complementarities of multimodal approaches using more discriminative sources. The Ticino dataset proposed in this work aims to address these challenges.

2.3.2 Deep Learning for Semantic Segmentation

In this section, the most popular deep learning models used for semantic segmentation will be described. In computer vision, the Convolutional Neural Network and recently Transformers represent the most used architectures to build models and analyze images.

Convolutional Neural Networks In semantic segmentation, different CNN models have been created to purposely extract features from each pixel and classify them to achieve segmentation. Usually, the general idea of these architectures consists of convolution, deconvolution, and fusion that is typically in the form of skip connection to recover spatial information during the deconvolution [90]. Some models [95] extended classic CNN architectures such as AlexNet [96], VGGNet [97], GoogleNet [98], and ResNet [99] to adapt them to semantic segmentation. Unet [100], SegNet [101], DeepLab [102], and DenseNet [103] are created directly with the aim of semantic segmentation.

Unet [100] is a CNN divided into two main phases: convolution and deconvolution. The former encodes the images using convolutional 3x3 filters, ReLU, and pooling to extract feature maps. The latter starts from the feature maps to reconstruct the dimension of the original image and classify each pixel. Between the encoder and the decoder part, Unet also implements skip connections betwixt the correspondent layer to recover the spatial structure of the original images.

SegNet [101] is divided into encoder and decoder as Unet. The encoder is the same as Unet and other CNNs and it is built to extract features of greater semantic meanings. At the same time, the spatial information in the deepest layers becomes ambiguous [90]. SegNet address this issue by storing the element index and using it for the upsampling of the decoder. Basically, SegNet differs from Unet because instead of concatenating the outputs of the encoder with the inputs of the decoders through skip connection, it reconstructs the spatial information by guiding the upsampling process using indexes that memorize the spatial position before operations of convolution or pooling.

DeepLab [102] extends CNN changing the classic convolution filters with dilated filters. By enlarging the neighborhood considered, these filters incorporate more context at each convolution operation without increasing the number of parameters of the network. To improve the spatial localization and edge accuracy of segmentation, degraded from downsampling and pooling operations [90], DeepLab applies Conditional Random Field (CRF). DeepLab V3 [104] and DeepLab V3+ [105] are variants of DeepLab that abandon the use of CRF and implement standard convolution and concatenation of feature maps to recover spatial information. In addition to DeepLab V3, DeepLab V3+ introduces a decoder to refine the boundary details.

DenseNet [103] is an extension of ResNet [99]. It introduces the concept of dense blocks helping to reduce the problem of vanishing gradient, enabling the reuse of features, and consequently needing fewer parameters to learn [90]. A dense block consists of a layer that receives in input every output of the previous layers incrementing the existing connections in the standard ResNet.

When it comes to remote sensing semantic segmentation, SegNet [106, 107, 108, 109], Unet [110, 111, 112, 113, 114, 115], DeepLab [111, 116, 106], and DenseNet [103] have been tested, achieving state of the art performance.

Transformers Since 2017, Transformers have become a reference point in the deep learning community, gradually increasing their importance in the state of the art [117]. The first Transformer architecture [118] had the aim to overcome long short-term memory methods for natural language processing. Transformers introduce the concept of self-attention that, simply using fully connected layers and multiplication between matrices, is capable of computing an association between every element of the input with each other (e.g.: every word of a sentence with every word of the same sentence). Transformer models have recently demonstrated exceptional performance on a broad range of language tasks, machine translation, and question-answering [119]. In computer vision, some models have been developed to take advantage of these architectures and have been proven to be comparable to or better than CNNs [120]. Contrary to the CNNs that have a local point of view of the images thanks to convolutional operations, the main advantage of Transformers is the global context knowledge achieved thanks to self-attention. Moreover, compared to their convolutional counterparts, Transformers assume minimal prior knowledge about the structure of the problem. Consequently, they can be pre-trained on pretext tasks on large-scale (unlabelled) datasets reducing the necessity of annotations [119].

Vision Transformer (ViT) [121] is the first Transformer applied in computer vision. The main problem addressed by ViT is how to create the input token of the network. Generally, in natural language processing, the main idea is that every word of a sentence becomes an input token of the Transformer, and the model computes the self-attention between the tokens. In computer vision, this ideally translates into using every pixel as an independent token and elaborating the attention between them. Due to token cardinality, this approach would be computationally unfeasible. ViT addresses this problem using patches in a grid shape to generate tokens instead of single pixels. The original model works perfectly for classification, but it is easily adaptable to segmentation problems by attaching a decoder for segmentation to the Transformer and using the latter as an encoder of features of the patches.

Pyramid Vision Transformer (PVT) [122] is the first hierarchical Transformer built for pixel-dense prediction. The encoder architecture consists of 4 stages where the dimension of the patches continues to increment. Instead of dividing the images in a regular grid of big patches as ViT [121], PVT starts with 4x4 patches. Each stage computes the relationship between the tokens using a Transformer block, and at the end of each stage, the features are reshaped as an image, and the process is repeated, doubling the dimensions of the patches.

Shifted Window Transformer (SWIN Transformer) [120] is another hierarchical Transformer, and as such, can extract hierarchical features considering different scales and also has linear complexity, being efficient both in terms of performance and computation. As PVT [122], SWIN starts with 4x4 patches but uses a sliding window of 4x4 patches to compute the attention only on the ones inside. The model also uses a Shifted Window approach where the window alternately changes shape to capture the relationship between the patches that overlap the borders of the previous window shape. After the computation with the two different forms, the patches merge, composing new patches, and the process is iterated till the window covers the entire image. So, dissimilarly from ViT, SWIN can capture relationships almost at pixel level without renouncing efficiency.

SegFormer [123] is a Transformer based architecture built to aim semantic segmentation. As PVT [122] and SWIN [120], it is a hierarchical Transformer, but differently from the other Transformer models, it doesn't need positional encodings that are usually considered together with the tokens to imply positional information. The model is divided into a hierarchical Transformer encoder to generate multi-scale features and a lightweight all-MLP decoder to fuse them and produce the final semantic segmentation mask.

Recently, another Transformer based model called MaskFormer [124] has been proposed. The idea behind MaskFormer is to change the base paradigm used to segment an image. Instead of classifying every pixel of an image, this model creates and classifies different sets of masks that combined achieve the final semantic segmentation.

In remote sensing, semantic segmentation some attempts to use Transformers alone or in combination with CNNs have been explored, for instance, Efficient-T [125], CCT-Net [126], Stranfuse [127], Trans-CNN [128], SwinTF [129], UnetFormer [130] have achieved competitive results combining the advantages of CNNs and Transformers.

Multimodal and fusion approaches Multimodal approaches have been really useful in many tasks such as image fusion, change detection, image localization, target recognition and tracking, image matching, etc. [131]. These strategies are used in various fields, from medical analysis, language translation, and image annotation, to remote sensing monitoring [132, 131]. The combination of different sources to achieve relevant improvements by exploiting the advantages of each type of data has been and is still being investigated. In semantic segmentation, various methods to fuse modalities have been studied and depend on the datasets, the model chosen, and the kind of fusion applied.

In RS, these kinds of fusion can be divided into two groups: heterogenous and homogenous [94]. Heterogeneous fusions represent the fusion techniques to combine modalities with different meanings, such as hyperspectral, LIDAR, and DTM. The homogeneous group indicates the fusion of modalities of the same type. For instance, these kinds of fusions include spatio-temporal fusion and pansharpenig [29, 94, 133]. The former is the fusion of the same images collected in diverse timesteps. The latter is typically used to upscale multispectral images and recently also involved the spatial improvement of hyperspectral data to the resolution of the panchromatic component, characterized by high spatial information.

Typically, with CNNs, the methods can be divided into 3 groups: early fusion (or data-level), middle fusion, and late fusion [134]. The fusion strategy usually involves concatenating the features of the different modalities to merge their characteristics. The main difference between each strategy resides in the position of the CNN model where the concatenation is applied. For instance, in early fusion, concatenation is the first step of the pipeline, directly combining the sources in a shallow way. The late fusion methods imply the merging after extracting high-level features from each modality independently and using the decisions on each source to define the final segmentation. Middle fusion, as the name suggests, is a middle-ground strategy between early and late techniques. The combination is in the middle of the pipeline, extracting features from each modality before concatenating them.

With Transformers's popularity and performance, new architectures purposely adapted for fusing have been investigated. Nonetheless, Transformer-based architectures have emerged as a prominent choice in multimodal learning research, but their utilization for semantic segmentation of RS images remains relatively underexplored [135]. These architectures give rise to different possible strategies [136]. For instance, the fusion of the tokens, by summing or concatenating them, can characterize the multimodal approaches with Transformers [137, 138]. Following the concatenation approach, hierarchical attention represents another example of multimodal techniques with transformers. It consists of concatenating and splitting the tokens before or after the attention mechanism. The hierarchical attention can be divided into two versions depending on the application order of the concatenation and splitting operations (from multi-stream to one-stream or vice versa) [139]. Another strategy modify the structure of the self-attention mechanism. A typical method among these strategies is represented by the Cross Attention [140] that swaps the query of a modality with another one during the computation of the classical attention. Finally, combinations of these techniques used concatenation and Cross Attention together [141, 142].

2.4 Digital Soil Mapping

Remote sensing, as described before, involves the use of sensors mounted on drones, aircraft, or satellites for observing and monitoring the Earth from a distance. When coupled with Machine Learning, this technology can also be applied to soil characterization, expediting and scaling up the soil characterization process [14]. The integration of these technologies facilitates accurate and high-resolution soil mapping, empowering land managers, farmers and policymakers with vital information for sustainable land use planning, precision agriculture and environmental conservation initiatives [13].

Several research make use of machine learning techniques to improve the quality of the estimation of soil parameters. For instance, Ladoni et al. [143] use partial least square regression to estimate the soil organic carbon (SOC) from hyperspectral images. Forkuor et al. [144] tackled the estimation of sand, silt, clay, CEC, SOC, and nitrogen (N) across a West African landscape by leveraging RapidEye and Landsat images, employing random forests (RF), support vector machines (SVM), and stochastic gradient boosting (SGB) techniques. Similarly, Safanelli et al. [145] estimate the clay, sand, SOC, calcium carbonates (CaCO3), CEC and pH present in water using the gradient boosting regression algorithm. Zhou et al. [146] harnessed RF, SGB, and SVM to foresee SOC and the C:N ratio from multispectral data in Switzerland [147].

Hu et al. [148] utilized RF with hyperspectral and multispectral data to estimate soil salinity. Guo et al. [149] perform the SOC prediction using vis-NIR (visible near-infrared) technology. Two approaches are compared: a direct method, which estimates directly the SOC from spectral information, and an indirect method, which in the first place estimates the soil organic matter (SOM) and the soil bulk density (SBD), and then computes the SOC value on the basis of the estimated variables.

Meng et al. [150] estimated SOC from hyperspectral images obtained from the Gaofen-5 (GF-5) satellite, employing a multiscale approach that featured the three bands with the highest SOC correlation as input features, and prediction models like RF [151], SVM [152], and backpropagation neural networks. Chambers et al. [153] created two datasets for soil nutrient prediction (P, K and M) and then investigated the use of several machine learning techniques for the prediction. The two datasets are, respectively, called global and local datasets, as the former covers several locations in Slovenia while the latter corresponds to a local farm. Both datasets are augmented using subsampling, reaching a total of 350 and 56 samples, respectively. Data acquisition of the spectra was performed within the UV-VIS (ultra-violet visible) range, specifically between 200 nm and 11,000 nm.

Li et al. [154] performed a prediction of soil properties (OC, N, and Clay) starting from vis-NIR signals. The investigation was conducted on two datasets: a small dataset collected by the authors and the LUCAS (Land Use and Coverage Area frame Survey) dataset [155]. The prediction is achieved through a multi-branch neural network that evaluates both the signal as-is and the corresponding 2D spectrum map, obtained through the use of the Fourier transform. The latter treats the vis-NIR as a temporal signal. Three different preprocessing methods are investigated: the Savitzky–Golay (S-G) smoothing algorithm [156, 157], multivariate scattering correction (MSC) [158, 159] and centralization methods.

Chapter 3

Hyperspectral Pansharpening

The hyperspectral pansharpening analysis described in this chapter can be found on *arXiv* and refers to the article entitled *Deep Learning Hyperspectral Pansharpening on large scale PRISMA dataset* [20].

Remote Sensing (RS) has revolutionized our ability to observe and analyze our planet from a vantage point beyond the Earth's surface [7]. The analysis of data gathered by sensors on board of satellites or aircraft, in fact, allows the inference of useful information about the land, water and atmospheric systems of the Earth. This technology became fundamental in several fields, such as environmental monitoring [10], agriculture [8], urban planning [11], disaster management [12], and resource exploration [9].

However, the costs to put a satellite in Earth orbit are very high. They range from 2.6k$/kg with SpaceX to 22k$/kg with NASA, with an intermediate value of 17.6k$/kg with Soyuz, the Russian rockets [160, 161]. Minimizing the payload is therefore the major goal that drives the choice and the design of every component on a satellite [162]. This constraint, in combination with the need to use as little energy as possible, results in a huge trade-off between the spatial resolution and the number of acquired bands when designing optical remote sensing devices. On the one hand, in fact, several orbital expeditions such as Landsat 6/7 [163], SPOT 6/7 [164], Sentinel-2 [165] included a panchromatic imaging device acquiring at high resolution [166]. On the other hand, missions carrying hyperspectral imaging devices such as ASI PRISMA [6], had to decrease the spatial resolution in favor of a higher number of acquired bands [167].

The loss of spatial resolution can be partially solved through the use of pansharpening [168]. In this context, the panchromatic image could be used as a source of information to extend the spatial resolution of the multispectral (MS) and hyperspectral (HS) images.

The first attempts of image pansharpening are machine-learning-free approaches [169], designed to handle data in the range of visible radiations (400 - 700 nm). These approaches assume the possibility of exploiting the existing relation between the panchromatic image and the spectral bands in the input data. This assumption may not be valid when handling

Table 3.1: List of existing datasets used for RS image pansharpening. For each dataset, the number of images, the number of bands, and the coverage in terms of wavelength are reported and compared with the dataset presented in this investigation. The image resolutions reported in this table are taken from the original dataset descriptions. The last row describes the dataset used in this work.

Dataset	Cardinality	Images resolution	Type	# of bands	Wavelength coverage
Pavia University [25]	1	610 × 610	airborne	103	430 - 838 nm
Pavia Center [25]	1	1096 × 1096	airborne	102	430 - 860 nm
Houston [26, 27]	1	349 × 1905	airborne	144	364 - 1046 nm
Chikusei [28]	1	2517 × 2335	airborne	128	363 - 1018 nm
AVIRIS Moffett Field [29]	1	37 × 79	airborne	224	400 - 2500 nm
Garons [29]	1	80 × 80	airborne	125	400 - 2500 nm
Camargue [29]	1	100 × 100	airborne	125	400 - 2500 nm
Indian Pines [25, 30]	1	145 × 145	airborne	224	400 - 2500 nm
Cuprite Mine	1	400 × 350	airborne	185	400 - 2450 nm
Salinas [25]	1	512 × 217	airborne	202 (224)	400 - 2500 nm
Washington Mall [31]	1	1200 × 300	airborne	191 (210)	400 - 2400 nm
Merced [26]	1	180 × 180	satellite	134 (242)	400 - 2500 nm
Halls Creek [32]	1	3483 × 567	satellite	171 (230)	400 - 2500 nm
OURS based on PRISMA	190	1259 × 1225	satellite	203 (230)	400 - 2505 nm

data outside the range of visible wavelengths, i.e. when there is partial or missing spectral overlap between the panchromatic image and the spectral bands to be processed.

Alongside these methods, neural network-based approaches have been recently developed, showing promising results. However, data-driven approaches are limited by the lack of high-cardinality datasets in the state of the art. Table 3.1 reports the most relevant datasets in the literature used for multispectral and hyperspectral RS pansharpening. The majority of them are composed of only one single satellite or aerial image covering a small portion of land (at most few km^2), with limited variability in the content of the scene.

In order to overcome the limitations relative to the lack of data, this work presents:

- a new large-scale dataset covering 262200 km^2 for qualitative assessment of deep neural models for hyperspectral image pansharpening. Such dataset is collected from the PRISMA satellite, preprocessed and adopted for the retraining of current state-of-the-art approaches for image pansharpening;

- an in-depth comparison, both in quantitative and qualitative terms, of the current deep learning approaches, re-trained and tested on the newly proposed large-scale dataset, and traditional machine-learning-free approaches.

The presented study is the first one based on a large-scale dataset, covering a wide variety of ground areas. The proposed investigation wants to be a starting point for the design of new deep-learning approaches for RS image pansharpening.

Table 3.2: Ranges of wavelengths covered by the panchromatic image and by the hyperspectral cubes VNIR and SWIR, and the corresponding number of bands. The PAN image covers most of the range of the VNIR cube, while the SWIR cube is completely outside of that range.

Cube	Wavelengths covered nm	# of bands	Resolution m/px	 pixels
panchromatic	400 – 700	1	5	7554 × 7350
VNIR	400 – 1010	66	30	1259 × 1225
SWIR	920 – 2505	174	30	1259 × 1225

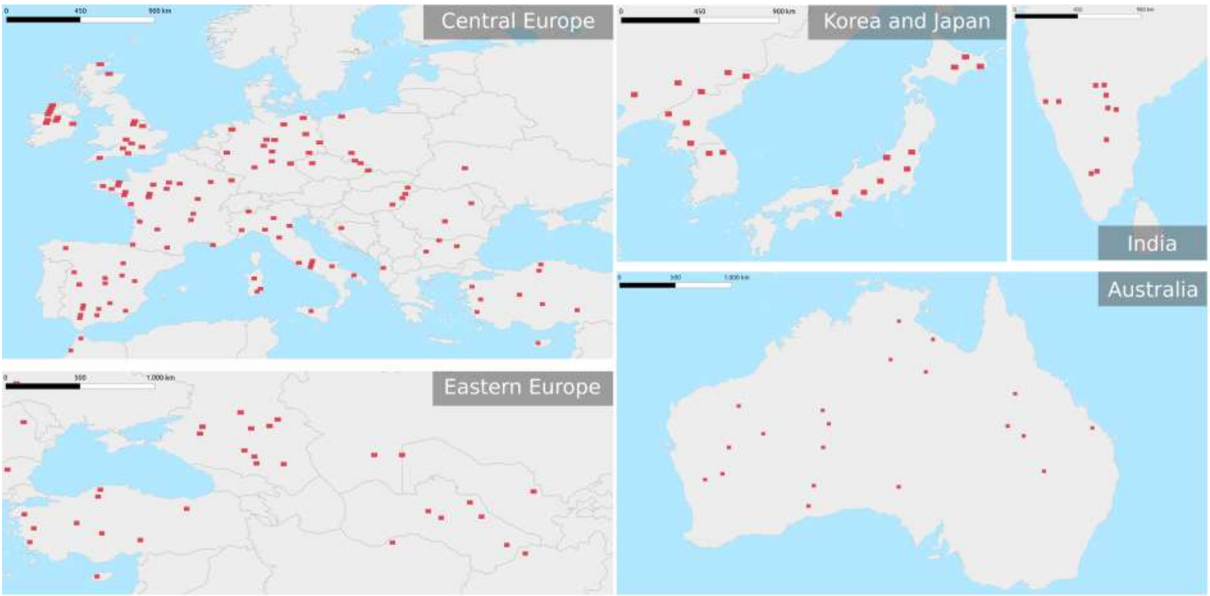


Figure 3.1: Map of the patches acquired using the PRISMA satellite. On average, every patch covers about 1380 km^2 of soil.

3.1 Data

In order to assess the performance of the approaches for hyperspectral image pansharpening, a new dataset of HS images has been collected using the PRISMA satellite. A collection of 190 images covering different areas, from Europe, Japan, Korea, India and Australia for a total of about 262200 km^2 has been gathered. The actual locations of the images are shown in Figure 3.1, while in the last row of Table 3.1, the proposed dataset is reported with a summary of its characteristics, compared with other existing datasets.

The data used for the construction of the proposed datasets has been collected by using the level-2D image data product downloaded from the ASI PRISMA portal for data distribution [6]. Visible and Near-InfraRed (VNIR), Short-Wave InfraRed (SWIR) cubes and the panchromatic (PAN) band have been extracted from these downloaded products, according to the Hierarchical Data Format (HDF5) standard. The hyperspectral cubes from the level-2D refer to the geocoded at-surface (Bottom-of-Atmosphere) reflectance data [170]. PAN images are at a spatial resolution of 5 meters per pixel, while VNIR and SWIR cubes (respectively 66 and 174 spectral bands) are at a spatial resolution of

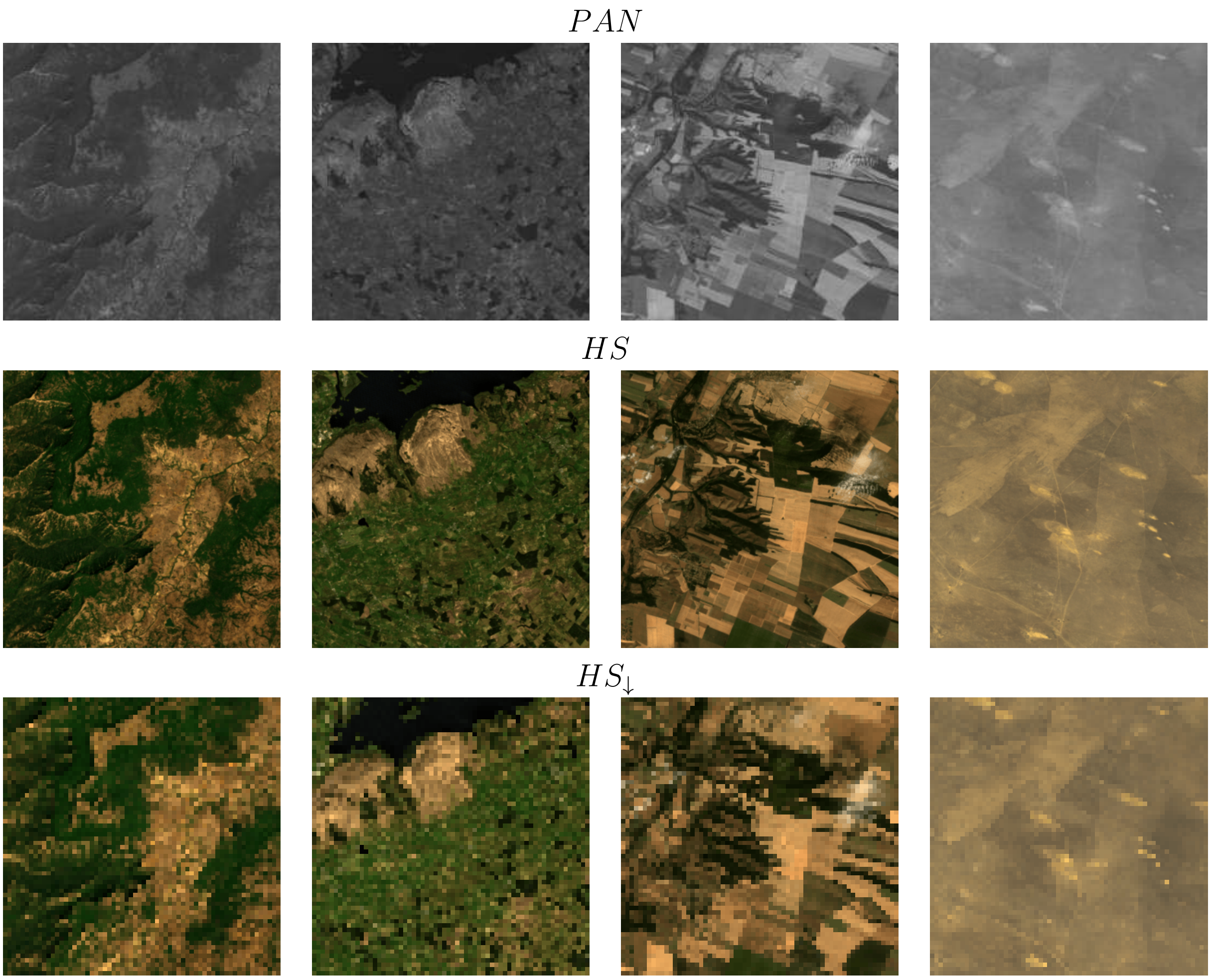


Figure 3.2: Examples of PRISMA dataset entries, visualized in true color RGB (641 nm, 563 nm, 478 nm). PAN image is at a resolution of 5 meters per pixel, HS images at 30 meters per pixel and $HS_{\downarrow}$ at 180 meters per pixel

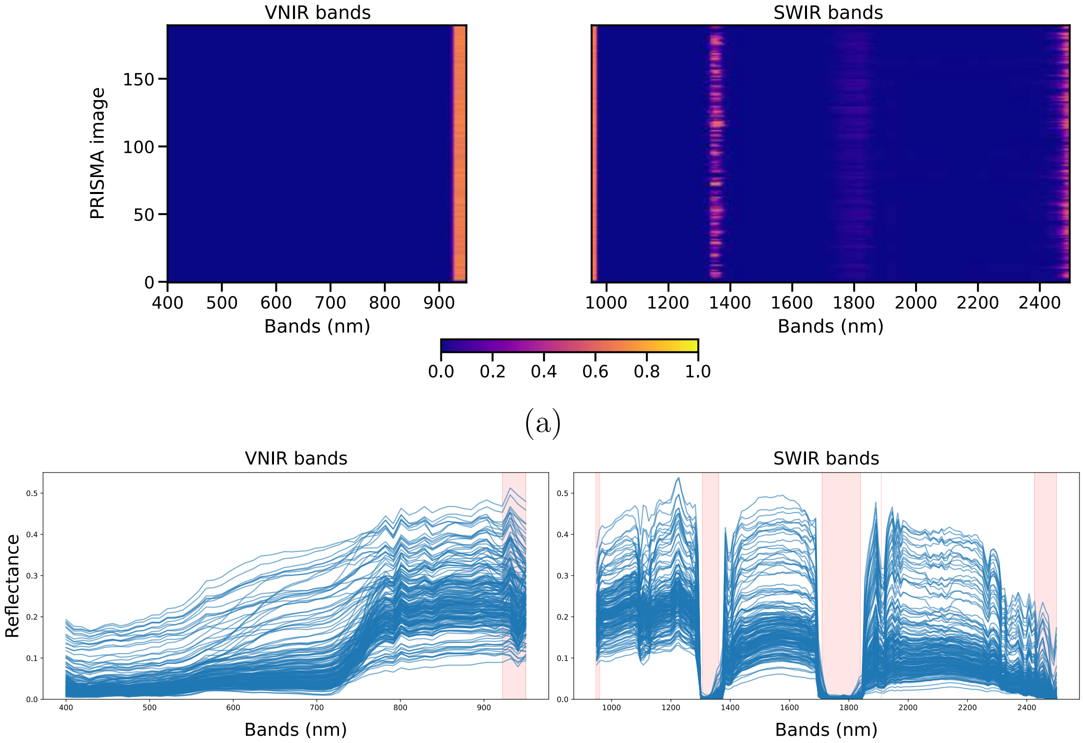
30 meters per pixel. Table 3.2 reports the details of each cube, while Figure 3.2 shows some examples of PRISMA data visualized in true-color RGB. Each PAN image is at a resolution of 7554 × 7350 pixels, while HS bands are at a resolution of 1259 × 1225 pixels.

Each collected image has been pre-processed by performing an image co-registration step and a cleaning step, with the last one used to remove bands that contain noisy or invalid data. Each image is then divided into tiles at different resolutions, to produce two sets of images for two different training and evaluation protocols: the Full Resolution (FR) protocol and Reduced Resolution (RR) protocol.

3.1.1 Data cleaning procedure

The VNIR and SWIR PRISMA level-2D product cubes cannot be directly used because of two problems:

- slight misalignment between panchromatic image and VNIR and SWIR cubes (VNIR



(a)

(b)

Figure 3.3: Distribution of the invalid band for each PRISMA image collected. In (a), the scale indicates the percentage of invalid pixels for each band of each image collected from the PRISMA satellite. Bands that are considered invalid for at least one image (with more than 5% of the entries invalid), have been selected for removal. In (b) are shown the average spectral signatures per image and the bands excluded in the final version of the dataset.

and SWIR are assumed to be aligned already);

- presence of pixels marked as invalid from the Level-2D Prisma pre-processing.

To tackle the first problem, we adopted the AROSIC framework [171] to align the VNIR and SWIR cubes to the corresponding panchromatic images. A reference band has been manually selected for the VNIR and SWIR cube to be used for the calculation of the transformation. The same alignment has been used for all the 240 bands of VNIR and SWIR cubes.

Invalid bands are removed through a cleaning procedure. From the PRISMA HDF5 data, the VNIR_PIXEL_L2_ERROR and SWIR_PIXEL_L2_ERROR matrices have also been extracted. These matrices contain pixel-specific annotations regarding the status of the information collected by the satellite. The bands having at least 5% of the pixels labeled as INVALID have been removed. More details on the labeling system are available at [170]. The selected bands have been removed from all the scenes collected from PRISMA. Figure 3.3(a) shows the distribution of the invalid bands (x-axis) over all the 190 selected PRISMA images (y-axis). Figure 3.3(b) shows the average spectral signature for each image (blue lines) and the spectral bands that have been removed (pink stripes). After

Table 3.3: Size and resolution of the input PAN images, HS bands, and pansharpened outputs in both RR and FR protocols.

	Size (px)	Resolution (m/px)	Usage FR	 RR
PAN	2304 × 2304	5	input	-
$PAN_{\downarrow}$	384 × 384	30	-	input
HS	384 × 384	30	input	reference
$HS_{\downarrow}$	64 × 64	180	-	input
$\hat{HS}_{FR}$	2304 × 2304	5	output	-
$\hat{HS}_{RR}$	348 × 348	30	-	output

this cleaning procedure, VNIR and SWIR bands are concatenated, thus obtaining a final HS cube of 203 spectral bands.

3.1.2 Full Resolution and Reduced Resolution datasets

The experimentation is made adopting two different protocols which require two different versions of the dataset:

1. *Full Resolution (FR)*: this dataset is used to evaluate the goodness of pansharpening algorithms without a reference image. Due to missing reference images, this dataset cannot be used for model training but only for evaluation purposes. This dataset is made of couples of the type $< PAN, HS >$.

2. *Reduced Resolution (RR)*: this dataset is created in order to perform full-reference evaluation since it presents reference bands alongside the input HS and the PAN, and for training the deep learning model. This dataset is made of triplets of the type $< PAN_{\downarrow}, HS_{\downarrow}, HS >$.

To create the two versions of the dataset, the original collected images were tiled and resized with different parameters. Table 3.3 provides a summary of the characteristics of the two versions.

The FR dataset is made of tiles of size 2304 × 2304 at the original spatial resolution of $5m/px$, for the PAN image, and 384 × 384 pixels at $30m/px$ resolution for the HS bands. In this study's experimental setup, the pansharpening algorithms are used to scale up the HS bands from $30m/px$ by a factor of 6×, thus obtaining a no-reference reconstruction $\hat{HS}_{FR}$ of HS bands at a size of 2304 × 2304 at a spatial resolution of $5m/px$.

The RR dataset is obtained by subsampling the FR version and generating triples of the type $< PAN_{\downarrow}, HS_{\downarrow}, HS >$. Firstly, the VNIR-SWIR bands are tiled at a dimension of 384 × 384 pixels, which corresponds to a resolution of 30 m/px (HS). These images are used as references for the evaluation of the algorithms performance. Then, the same cubes are further reduced at 1/6 of their original resolution, obtaining new tiles at size 64×64 at a spatial resolution of 180 m/px ($HS_{\downarrow}$) which are the input of the pansharpening algorithm. The panchromatic images are also reduced to 1/6 of the original resolution and tiled at size 384×384 pixels at a spatial resolution of 30 m/px ($PAN_{\downarrow}$), in order to be used as input for the pansharpening operation. The pansharpening algorithm is defined as a function that takes as input the pair $< PAN_{\downarrow}, HS_{\downarrow} >$, and it outputs an approximation $\hat{HS}_{RR}$ of the original HS, which is a 6× version of the $HS_{\downarrow}$.

3.2 Reduced Resolution Metrics

The following evaluation metrics have been used to compare the pansharpened $\hat{HS}_{RR}$ image and the reference HS:

- ERGAS [172] (Error Relative Global Dimensionless Synthesis) is an error index that tries to propose a global evaluation of the quality of the fused images. This metric is based on the $RMSE$ distance between the bands that constitute the fused and the reference images and is computed as:

$$RMSE(x, y) = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(x_j - y_j)^2} \tag{3.1}$$

$$ERGAS(x, y) = 100\frac{h}{l}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{RMSE(x_i, y_i)}{\mu(y_i)}\right)^2} \tag{3.2}$$

  where x and y are the output pansharpened image and the reference, respectively, m is the number of the pixels in each band, h and l are the spatial resolution of the PAN image and HS image, respectively, $\mu(y_i)$ is the mean of the $i-th$ band of the reference and N is the number of total bands.

- The Spectral Angle Mapper (SAM) [173] denotes the absolute value of the angle between two vectors v and $\hat{v}$.

$$SAM(v, \hat{v}) = cos^{-1}\frac{< v, \hat{v} >}{||v||_2 \cdot ||\hat{v}||_2} \tag{3.3}$$

  where v and $\hat{v}$ are respectively the flattened versions of $\hat{HS}_{RR}$ and HS. A SAM value of zero denotes complete absence of spectral distortion but possible radiometric distortion (the two vectors are parallel but have different lengths).

- The Spatial Correlation Coefficient (SCC)[174] is a spatial evaluation index that analyses the difference in high-frequency details between two images. SCC is computed as follows:

$$SCC(x,y) = \frac{\sum_{i=1}^{w}\sum_{j=1}^{h}(F(x)_{i,j} - \mu_{F(x)})(F(y)_{i,j} - \mu_{F(y)})}{\sqrt{\sum_{i=1}^{w}\sum_{j=1}^{h}(F(x)_{i,j} - \mu_{F(x)})^2 \sum_{i=1}^{w}\sum_{j=1}^{h}(F(y)_{i,j} - \mu_{F(y)})^2}} \tag{3.4}$$

  where $\mu_{F(x)}$ and $\mu_{F(y)}$ are the means of $F(x)$ and $F(y)$ respectively and and w and h are the weight and height of an image. F is a filter for the extraction of high-frequency details, defined as follows:

$$F = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{3.5}$$

- The $Q2^n$ index is a generalization of the Universal Quality Index (UQI) defined by Wang et al. [175] for an image x and a reference image y.

$$Q2^n(x,y) = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \cdot \frac{2\bar{x}\bar{y}}{(\bar{x})^2 + (\bar{y})^2} \cdot \frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \tag{3.6}$$

  Here $\sigma_{x,y}$ is the covariance between x and y, and σ_x and $\bar{x}$ are the standard deviation and mean of x, respectively. The $Q2^n$ metric represents a good candidate to give an overall evaluation of both radiometric and spectral distortions in the pansharpened images.

3.3 Full Resolution Metrics

For the FR assessment, it has been decided to use the *Quality with No Reference* index (QNR), as done by Vivone at al. [133]. This index is obtained as the product of the spectral distortion index D_λ and the spatial distortion index D_s.

The spectral distortion index D_λ is computed as proposed in the Filtered-based QNR (FQNR) quality index [176]. In this definition, each fused HS band is spatially degraded using its specific Modulation Transfer Function (MTF) matched filter[1], then the $Q2^n$ index between the set of spatially degraded HS images and the set of original HS data is computed, and eventually the unit complementary value is taken in order to obtain a

[1] The filter is defined for ensuring the consistency property of the Wald's protocol [177]. As done by Vivone et al. [133], it has been assumed that the HS sensor's MTFs follow a Gaussian shape with a standard deviation set all equal to 0.3.

distortion measure:

$$D_\lambda = 1 - Q2^n(\hat{H}_{L\downarrow}, H) \tag{3.7}$$

where, $\hat{H}_{L\downarrow}$ is the pansharpened image which has been spatially degraded using the MTF filter and decimated to input spatial dimension and H are the input hyperspectral bands. As done by Vivone et al. [133], the Q (UQI) index has been adopted instead of the $Q2^n$ index for computational reasons due to the high number of HS bands. As stated by Vivone et al. [133], comparable performance can be obtained with this modification, while drastically improving the computation time.

Spatial consistency D_s is computed as described by Alparone et al. [178]. Adopting a linear regression framework, the PAN image is modeled as a linear combination of the fused HS bands. To measure the extent of the spatial matching between the fused HS bands and the PAN image, the coefficient of determination is exploited [178].

$$D_s = 1 - R^2 \tag{3.8}$$

Finally, the QNR index is calculated as:

$$QNR = (1 - D_\lambda)^\alpha \cdot (1 - D_s)^\beta \tag{3.9}$$

Here the two exponents α and β determine the non-linearity of response in the interval $[0, 1]$. The value of these two parameters has been set to 1, based on previous work choices [133].

3.4 Hyperspectral Pansharpening Methods

Six deep learning and three traditional machine-learning-free approaches have been compared. The selection of the methods has been done taking into consideration two factors: how recent is the approach and the availability of the source code. For what concerns the machine-learning-free approaches, the methods considered were Principal Component Analysis (PCA) [33], Gram-Schmidt Adaptive (GSA) [36] and HySure [179]. For all these methods, the implementation available in the Mini Toolbox PRISMA [2] has been used. Regarding the deep learning methods, PNN [43], PanNet [45], MSDCNN [46], TFNet [47], SRPPNN [48] and DIPNet [49] have been selected.

Since the main interest is the evaluation of the 6× upscaling pansharpening task, the methods originally designed for scale factors power of 2 (e.g. 2×, 4×, 8×, etc...) have been modified. These methods are:

[2] `https://openremotesensing.net/wp-content/uploads/2022/11/Mini-Toolbox-PRISMA.zip`

- **SRPPNN [48]:** the architecture proposed by Cai et al. [48] is characterized by multiple progressive upsampling steps, that correspond to a first 2× and a secondary latter 4× upscaling operations. Those two upsampling operations have been changed by modifying the scale factors to 3× and 6× respectively. The rest of the original architecture has not been changed.

- **DIPNet [49]:** this model is made of 3 main components. The first two are feature extraction branches, respectively for the low-frequency and high-frequency details of the panchromatic image; here, the stride value of the second convolutional layer used to reduce the features' spatial resolution has been changed, from 2 to 3, in order to bring the extracted features at the same dimension of the input bands to perform feature concatenation. The third component is the main branch, which uses the features extracted from the previous components along with the input images to perform the actual pansharpening operation. The main branch can be also divided into two other components: a first upsampling part and an encoder-decoder structure for signal post-processing. The scaling factor of the upscaling module has been changed from 2 to 3, and in the encoder-decoder part, the stride values of the central convolutional and deconvolutional layers have been changed from 2 to 3.

Each method has been retrained on the proposed PRISMA dataset (RR version) by using a workstation equipped with a Titan V GPU and Ubuntu 22.04 Operating System. The environment for the training has been written in PyTorch `v1.10.0`. For all the methods the training process lasted 1000 epochs, with a learning rate of $1e^{-4}$ and the Adam optimizer. The loss functions used are the ones adopted in the original papers of each method.

3.5 Results

3.5.1 Quantitative Comparison

Table 3.4 and Table 3.5 report the numerical results of the different selected approaches with the RR and FR protocols respectively.

The two best methods for RR pansharpening protocol are DIPNet and TFNet. In-depth analysis reveals that DIPNet outperforms all other approaches across various metrics, except for the SCC index, where it ranks fourth. As the second-best algorithm, TFNet demonstrates commendable results that are comparable to those achieved by SRPPNN. Notably, machine-learning-free approaches generally exhibit lower performance compared to the majority of neural-network-based methods.

Figure 3.4 reports a graphical comparison between network-based approaches (in the RR protocol). The comparison evaluates the performance in terms of ERGAS versus SAM

Table 3.4: Results of the methods for the Reduced Resolution (RR) protocol. The dimensions (millions of parameters) of each model are reported alongside the results.

Method	# of parameters (M)	ERGAS ↓	SAM ↓	SCC ↑	$Q2^n$ ↑
PCA [33]	-	8.9545	4.8613	0.6414	0.6071
GSA [36]	-	7.9682	4.3499	0.6642	0.6686
HySure [179]	-	8.3699	4.8709	0.5832	0.5610
PNN [43]	0.08	12.8840	3.8465	0.8237	0.6702
PanNet [45]	0.19	6.7062	2.7951	0.8705	0.7659
MSDCNN [46]	0.19	9.9105	3.0733	0.8727	0.7537
TFNet [47]	2.36	<u>6.4090</u>	2.4644	<u>0.8875</u>	<u>0.7897</u>
SRPPNN [48]	1.83	6.4702	<u>2.3823</u>	**0.8890**	0.7708
DIPNet [49]	2.95	**5.1830**	**2.3715**	0.8721	**0.7929**

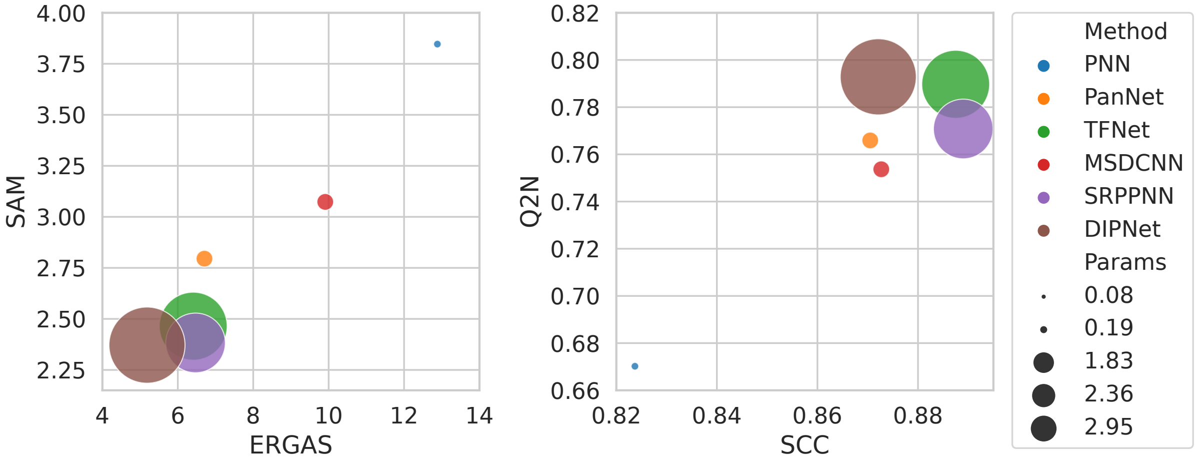


Figure 3.4: Graph comparison of the results of the analyzed methods with the RR protocol. The larger is the size of the circle the higher is the number of parameters (measured in millions).

Table 3.5: Results of the methods for the Full Resolution (FR) protocol. The dimensions (millions of parameters) of each model are reported alongside the results.

Method	# of parameters (M)	D_λ ↓	D_s ↓	QNR ↑
PCA [33]	-	0.9411	1.5277	0.0558
GSA [36]	-	0.3820	0.0016	0.6170
HySure [179]	-	0.4151	0.0009	0.5843
PNN [43]	0.08	0.3801	0.0101	0.6136
PanNet [45]	0.19	**0.3507**	0.0203	<u>0.6360</u>
MSDCNN [46]	0.19	0.3915	<u>0.0068</u>	0.6044
TFNet [47]	2.36	<u>0.3552</u>	**0.0066**	**0.6405**
SRPPNN [48]	1.83	0.3948	0.0139	0.5965
DIPNet [49]	2.95	0.3681	0.0348	0.6098

(Figure 3.4a) and SCC versus $Q2^n$ (Figure 3.4b), along with the number of parameters associated with the neural models. The size of the circles in the figure corresponds to the number of parameters, measured in millions. Larger circles indicate a higher number of parameters. Ideally, the optimal approach would be represented by a small circle positioned in the bottom-left part of Figure 3.4a and the top-right part of Figure 3.4b. In practice, the best neural methods are DIPNet, SRPPNN, and TFNet which have a number of parameters that is about 30 times the number of parameters of less-performing approaches, such as PanNet and MSDCNN.

The results obtained in the FR protocol are presented in Table 3.5, revealing a significant shift in the behavior of the models. Notably, TFNet emerges as the top-performing model in terms of QNR index. Surprisingly, DIPNet, which was the winning method in the RR protocol, demonstrates considerably poorer results compared to other approaches. Even the simpler and smaller PanNet outperforms DIPNet, securing the second position in the comparison.

Analyzing the spatial distortion aspect (D_s), the top-performing models are TFNet and MSDCNN, while DIPNet exhibits the weakest performance among the deep learning models. It is worth mentioning that HySure is the best method in terms of D_s; however, additional insights regarding its performance can be found in Section 3.5.2, showing various issues in the spatial reconstruction of this technique.

On the other hand, from a spectral distortion perspective (D_λ^k), PanNet emerges as the best approach, followed by TFNet and DIPNet. A comprehensive qualitative comparison of these two aspects of the reconstruction is presented in the subsequent section. Notably, PanNet's achievement of the second-best position in the FR leaderboard is particularly noteworthy, given its comparatively smaller size compared to TFNet and other more recent approaches.

In conclusion, TFNet emerges as the most successful approach when evaluating both

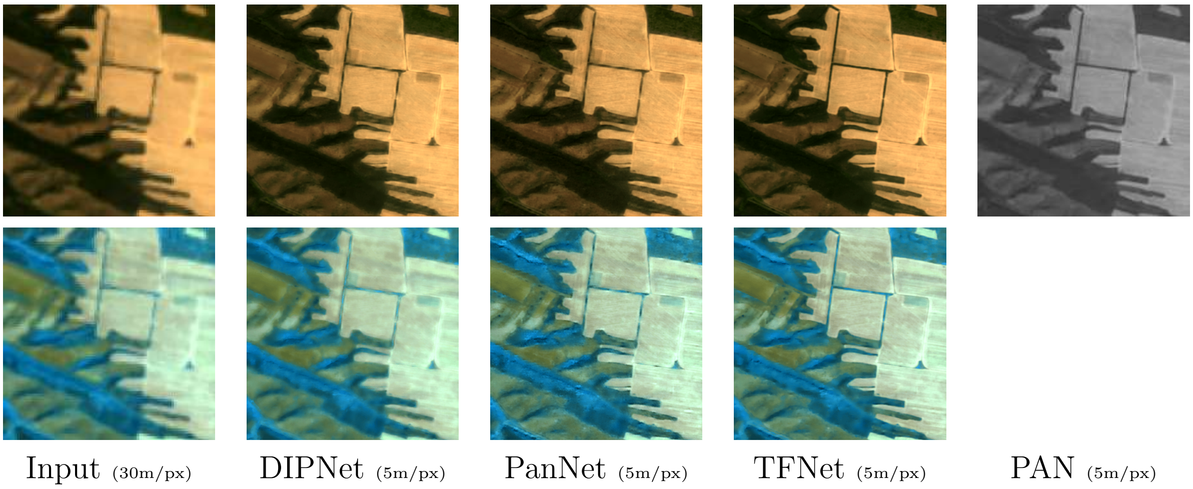


Figure 3.5: Pansharpening results on a 512 × 512 *tile of a test set image. For visualization purposes, images have been linearly stretched between the 1 and 99 percentile of the image histogram. In the first row, images are visualized in true color (*641 *nm*, 563 *nm*, 478 *nm), and in the second row, images are in false color (*1586 *nm*, 1229 *nm*, 770 *nm).*

the RR and FR protocols. Notably, TFNet exhibits a commendable ability to strike a balance between preserving spectral and spatial information throughout the pansharpening process, particularly evident in the FR test case. When compared to SRPPNN and DIPNet, TFNet demonstrates superior generalization capabilities when transitioning from the training resolution of 180 m/px to the native resolution of 30 m/px of the PRISMA satellite hyperspectral images.

3.5.2 Qualitative Comparison

In this section, a qualitative comparison of the results on a selection of test images is presented, analyzing the results in terms of the preservation of spatial and spectral distortions after the pansharpening process.

Figures 3.5, 3.6, and 3.7 show the results of the best models on three images of the FR protocol. Here are shown center crops of dimensions 512 × 512 of the pansharpened images (5 m per pixel) alongside the same crop of the original input image (30 m per pixel). For visualization purposes, images have been linearly stretched between the 1 and 99 percentile of the image histogram. In the first row, images are visualized in true color (641 nm, 563 nm, 478 nm), and in the second row, images are in false color (1586 nm, 1229 nm, 770 nm). For what concerns the spatial information, as can be seen here and as already highlighted by the quantitative comparison, TFNet presents the overall best-looking structures and details. Among the considered methods and especially in comparison with DIPNet, TFNet reconstructs much cleaner images, with good edges and a lot more details. PanNet still achieves good results compared to DIPNet but with still few artifacts and aberrations of different kinds. In Figures 3.6 and 3.7, it is easy to notice the presence of such artifacts,

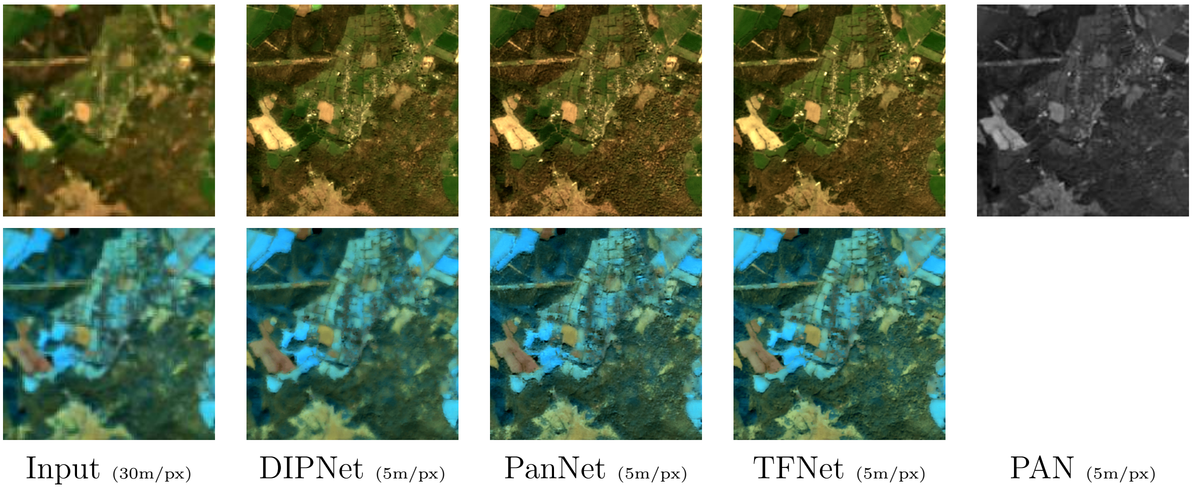


Figure 3.6: Pansharpening results on a 512 × 512 *tile of a test set image. For visualization purposes, images have been linearly stretched between the 1 and 99 percentile of the image histogram. In the first row, images are visualized in true color* (641 *nm*, 563 *nm*, 478 *nm*)*, and in the second row, images are in false color* (1586 *nm*, 1229 *nm*, 770 *nm*)*.*

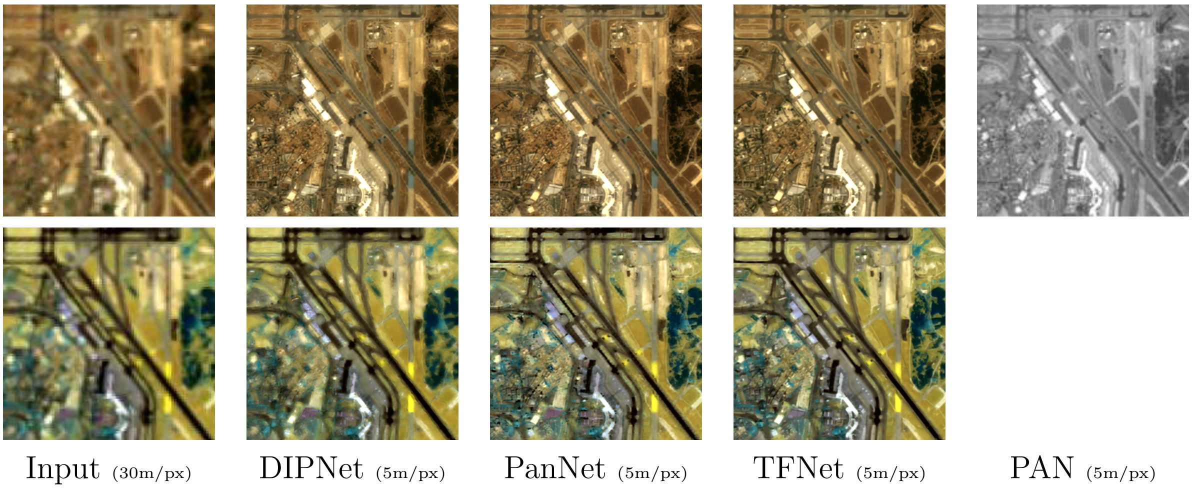


Figure 3.7: Pansharpening results on a 512 × 512 *tile of a test set image. For visualization purposes, images have been linearly stretched between the 1 and 99 percentile of the image histogram. In the first row, images are visualized in true color* (641 *nm*, 563 *nm*, 478 *nm*)*, and in the second row, images are in false color* (1586 *nm*, 1229 *nm*, 770 *nm*)*.*

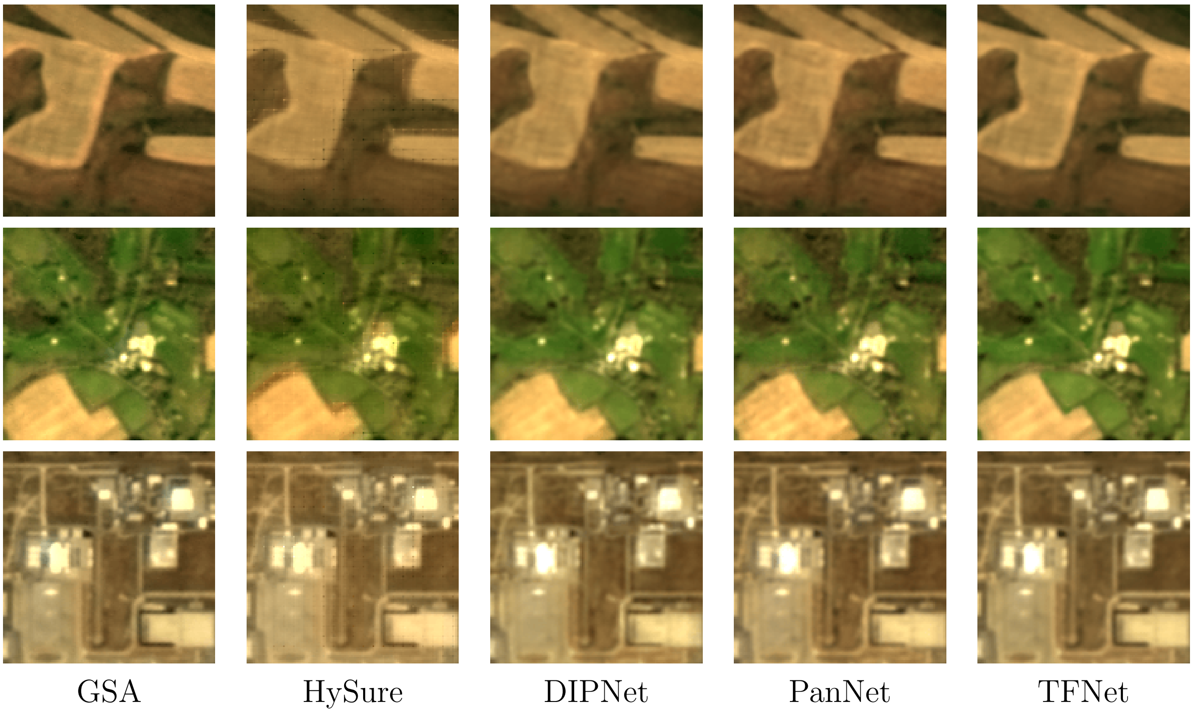


Figure 3.8: Zoom of areas from the test images. Crops of dimension 128 × 128*, at resolution* 5 m/px*, in true color (*641 nm, 563 nm, 478 nm*). As can be seen, repeated artifacts along the edges can be observed for the HySure method.*

due to the high amount of details in the scenes. Overall, DIPNet presents the most blurry results, with a very poor amount of details and the presence of artifacts, particularly noticeable in the false color composite version of the reported scenes.

Figure 3.8 shows zoomed crops at dimension 128 × 128 of areas of the same images processed by the best neural-based and the two best machine-learning-free methods. As can be seen, even if HySure numerically represents the best approach from the spatial point of view (see Table 3.5, D_s index), a pattern of artifacts occurs over all the images processed by the HySure algorithm. This last comparison shows a potential problem in the adoption of QNR index as a metric for the no-reference analysis when this type of artifacts occurs in the pansharpened images.

Figure 3.9 reports the average differences between the spectral signatures of each method and the reference one, alongside the normalized version of the same difference, computed on 5 different tiles. These tiles have been randomly extracted from the test set. From this comparison, it is possible to notice how DIPNet's average error is much smaller with respect to the other approaches. Compared with the results from the quantitative evaluation with the FR protocol, where DIPNet reaches third place, this is the only unexpected behavior. This unexpected result could be an insight into a possible flaw in adopting the spectral component of the QNR metric as it is usually done in the literature.

Figure 3.10 shows the spectral signatures of both input and pansharpened images for each method. Here only selected groups of pixels are considered, specifically labeled

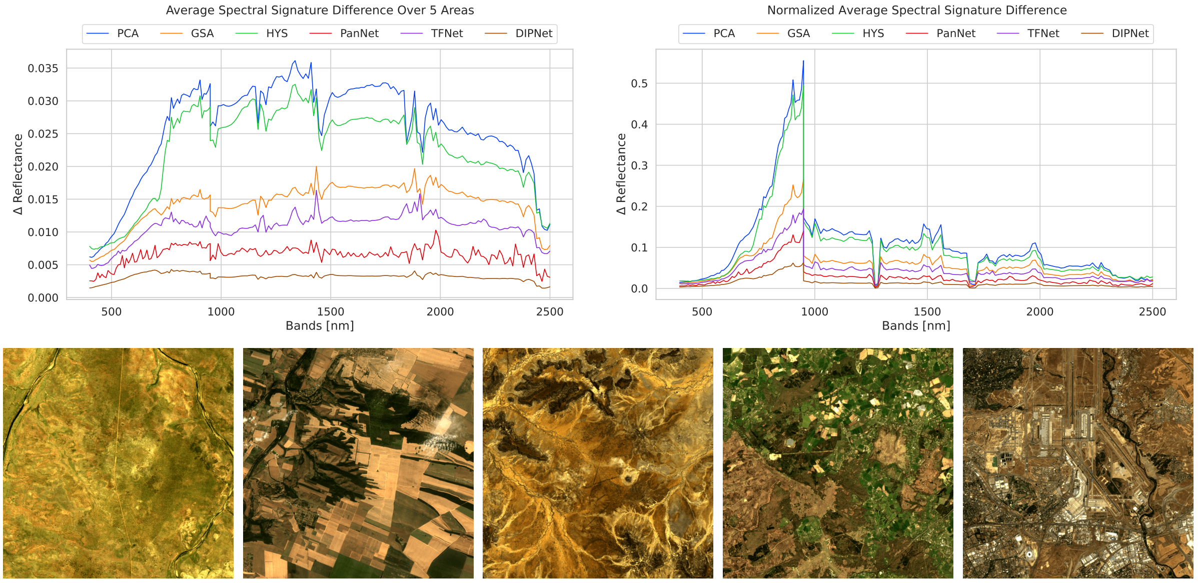


Figure 3.9: Difference between spectral signatures of the fused images with respect to the input image. The difference is evaluated as the average of the differences for each pixel of the five images reported in the row below the graph. The graph on the left shows the average spectral difference, while the graph on the right shows the difference normalized for each band.

as *Forest*, *Urban*, *Agriculture*, and *Water*, highlighted in red in the images aside the graphs. The best method is expected to show signatures closer to the input ones; both the machine-learning-free and the best deep-learning approaches are reported. From this qualitative comparison and according to the numbers in Table 3.5, PCA and HySure are the worst-performing methods in terms of spectral fidelity. For all the reported classes, these methods perform badly over the entire spectrum, and, in particular, PCA and Hysure performance is even worse than GSA. Regarding the deep learning approaches, DIPNet seems to show the lowest difference from the input, despite the score obtained in terms of D_λ (see Table 3.5). TFNet and PanNet instead have a behavior more coherent with the results obtained in the quantitative evaluation with the FR protocol. PanNet seems to perform better from a spectral fidelity point of view with respect to TFNet, which, however, performs better than all of the machine-learning-free approaches. Overall, the results in Table 3.5 can be confirmed, with DIPNet, PanNet, and TFNet as the best approaches, with performance higher in comparison to the machine-learning-free approaches.

Images at higher resolution are available at `https://thezino.github.io/HSbenchmarkPRISMA/`.

3.6 Final remarks on hyperspectral pansharpening techniques

The increasing availability of hyperspectral remote sensing data presents new opportunities for studying Earth's soil. However, this type of data is typically collected at low resolution,

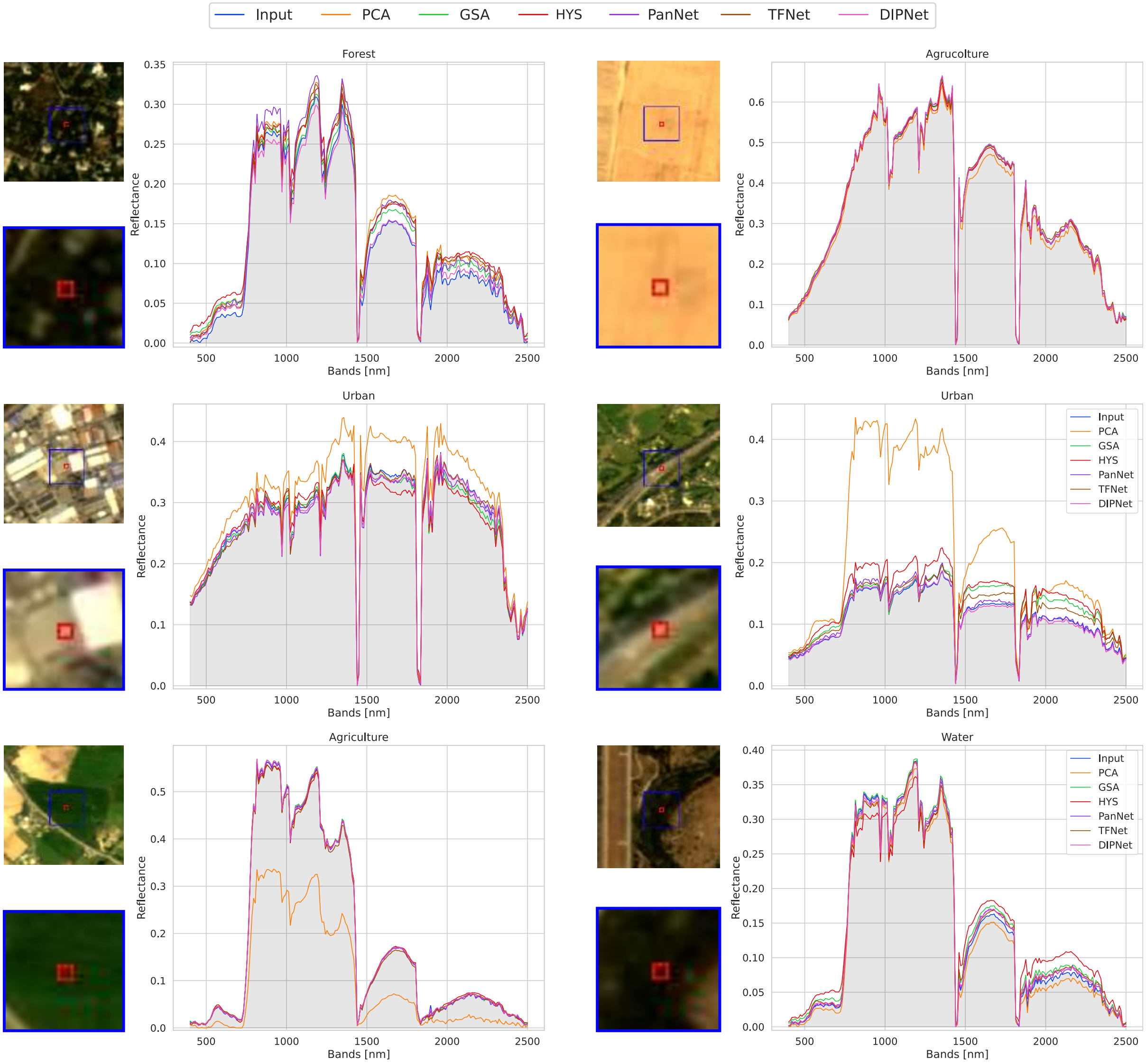


Figure 3.10: Spectral signatures obtained in six different areas, labeled as Forest, Urban, Agriculture, *and* Water *areas. For each area are reported the spectral signatures of the input bands and the ones obtained by each pansharpening method. The area used to extract the signatures is the one in the red box highlighted in each image. The images thumbnails are in true colors* (641 *nm*, 563 *nm*, 478 *nm*).

posing challenges in their effective usability in RS tasks. Therefore, the process of image pansharpening becomes crucial in the enhancement of hyperspectral remote sensing images.

In literature, deep learning approaches have shown promising results. These methods, however, are data hungry and state-of-the-art datasets strive to support sufficient data. To overcome this limitation, a newly collected large-scale dataset is proposed, using the PRISMA ASI satellite for training and assessment models.

This work is the first to present an analysis based on a big and various hyperspectral dataset for pansharpening in the state of the art (more than 1000 tiles, of which more than 200 are used for testing). The dataset tiles have been collected from 190 PRISMA images with 203 bands[3] covering both the VNIR and SWIR parts of the spectrum, making this investigation the optimal benchmark for hyperspectral pansharpening.

The comparison includes machine-learning-free and deep learning techniques tested using two experimental protocols for 6× upscaling factor: Reduced Resolution (RR) and Full Resolution (FR). The former is used for training and testing, and the latter to test the methods on the original resolution, evaluating their ability to generalize the upsampling operation at different starting resolutions with respect to the training phase.

The RR protocol consists of a comparison between the reconstructed data and the original hyperspectral images as target references. The results show that the neural networks generally work better than the machine-learning-free methods for spatial information improvement and spectral information coherency. In particular, DIPNet and TFNet architectures outperform any other techniques evaluated. In the FR protocol, the comparison with ground truths is not possible thus quantitative and also qualitative evaluations have been reported to have a complete understanding of the investigated methods. Based on both assessments, the architecture that achieves the best overall performance is TFNet which remains coherent with the RR results. DIPNet instead shows worse results when it comes to spatial reconstruction, not demonstrating good abilities of adaptation when the original resolution is involved and thus not being the best option for tests in real-world applications. It is also valuable to notice that machine-learning-free methods are generally worse at reconstructing the spectral information, degrading the signals.

The investigation conducted in this work clearly shows that data-driven neural architectures are generally better for hyperspectral pansharpening, both in spectral and spatial reconstruction, using a dataset that allows for meaningful analysis of the different approaches. On the contrary, The machine-learning-free methods are not adaptable to the new environment based on hyperspectral data and wavelengths outside the visible part of the spectrum.

To further improve hyperspectral pansharpening performance, future research should

[3] The original number of PRISMA spectral bands is 240. The number reported above is obtained after a proposed pre-processing procedure.

focus on creating new data-driven neural architectures which directly focus on hyperspectral data and the relationship between the different portions of the spectrum.

Chapter 4

Unsupervised Segmentation of hyperspectral images

The study reported in this chapter has been published in *Elsevier Remote Sensing Applications: Society and Environment* journal and refers to the paper *Unsupervised segmentation of hyperspectral remote sensing images with superpixels* [10].

The analysis of hyperspectral remote sensing images has become more and more important in a wide number of fields, such as environmental monitoring, conservation goals, spatial planning enforcement, or ecosystem-oriented natural resources management [2]. The use of HS images permits the analysis of specific electromagnetic ranges that allow a precise differentiation of observed materials, based on their spectral signatures [19]. Its ability to distinguish among several materials has shown to be a great boost in terms of performance for *HS image classification.* For example, buildings, cultivations, and rivers can be easily discerned in the images as their spectral profile is different.

To fruitfully exploit HS image classification, in particular for data-hungry methods such as neural networks [180], it is necessary to rely on large and properly annotated image datasets. Unfortunately, labeled datasets publicly available in the state of the art are few and extremely small, with most of them composed of a single image [90]. The main problem in the generation of remote sensing image ground truth is that the labeling creation is usually an interactive task that takes a lot of time and effort [82], and it is an operation susceptible to errors [181]. To overcome these problems in pixel annotations, a pre-processing of segmentation can be applied to divide the data into homogeneous regions and objects [182]. Most segmentation methods do not directly extract meaningful image objects, but clusters with generic labels, which can be used as the foundation of a succeeding procedure [183]. To provide a pixelwise image segmentation that can be later exploited in an interactive labeling process to speed up ground truth creation [184], different approaches of unsupervised clustering have been proposed [61, 80].

Previous studies report on the possibility of using *superpixels* to group pixel-sharing homogeneous characteristics for semi-supervised image segmentation [185, 58]. A recent and relevant segmentation method is the Balanced Deep Embedded Clustering (BDEC [80]) that works directly on the raw hyperspectral image. This method is a variant of the Deep Embedded Clustering (DEC - [79]) adapted to work properly with imbalanced data. This method is not completely unsupervised since it requires an *a-priori* knowledge about the land to be segmented (e.g. water, vegetation, building, etc.).

In this study, it is proposed an unsupervised method for hyperspectral remote sensing pixelwise image segmentation that exploits the mean-shift algorithm [186] that takes as input a preliminary superpixels segmentation together with the spectral pixel information. The preliminary superpixels segmentation is obtained using a modified version of the Simple Linear Iterative Clustering (SLIC) [187] algorithm that considers as input a concatenation between the hyperspectral image and a clustered-hyperspectral information achieved by using unsupervised clustering. The use of clustered information reduces the effect of noise, typical in hyperspectral remote sensing images.

The proposed method, differently from the state of the art, does not require the number of segmentation classes as input parameters [188], as well as it does not require a-priori knowledge about the type of land-cover or land-use to be segmented. The effectiveness of the proposed method is demonstrated with respect to the state of the art on four publicly available datasets of hyperspectral remote sensing images.

Despite its simplicity, the proposed method outperforms the state of the art in terms of average Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). In addition, it permits to overcome some limitations found in the literature that limit the applicability in real case scenarios, in particular:

- it does not require the training of a specific neural model since it is based on handcrafted features
- it does not require a-priori knowledge about the number of classes present in the image;
- it does not require external knowledge about the image content such as the vegetation index, water, etc;

Moreover, two variants of the proposed method are presented, one totally automatic and one that can be easily tweaked through a single parameter to improve the performance on a new dataset.

4.1 Proposed method

Figure 4.1 and 4.2 show the pipeline of the proposed method. It can be divided into two steps: 1) augmented hyperspectral superpixels are achieved by using a modified version of

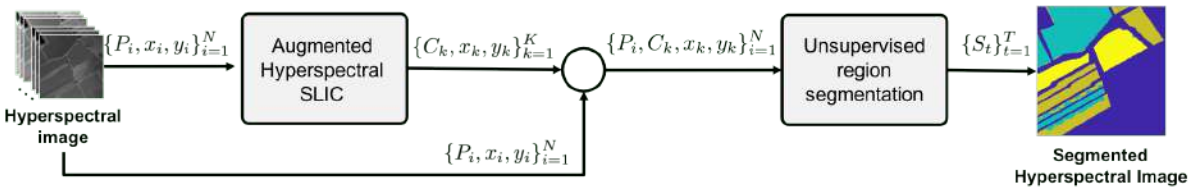

$\{P_i, x_i, y_i\}_{i=1}^N$
Hyperspectral image
Augmented Hyperspectral SLIC
$\{C_k, x_k, y_k\}_{k=1}^K$
$\{P_i, C_k, x_k, y_k\}_{i=1}^N$
Unsupervised region segmentation
$\{S_t\}_{t=1}^T$
$\{P_i, x_i, y_i\}_{i=1}^N$
Segmented Hyperspectral Image

Figure 4.1: Pipeline of the proposed method. In the first step, given the hyperspectral image, the augmented hyperspectral superpixels are calculated. In the second step, the centers of the superpixels along with the hyperspectral image are then used for unsupervised region segmentation using the mean-shift algorithm and major voting.

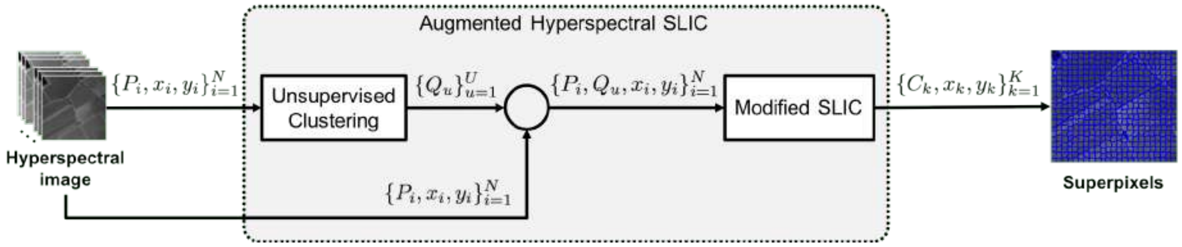

Augmented Hyperspectral SLIC
$\{P_i, x_i, y_i\}_{i=1}^N$
Hyperspectral image
Unsupervised Clustering
$\{Q_u\}_{u=1}^U$
$\{P_i, Q_u, x_i, y_i\}_{i=1}^N$
Modified SLIC
$\{C_k, x_k, y_k\}_{k=1}^K$
$\{P_i, x_i, y_i\}_{i=1}^N$
Superpixels

Figure 4.2: Augmented Hyperspectral Simple Linear Iterative Clustering (SLIC). In the first step, given the hyperspectral image, an unsupervised clustering is applied and the output is merged with the input hyperspectral image to finally achieve the augmented hyperspectral superpixels.

the Simple Linear Iterative Clustering (SLIC) that takes as input the hyperspectral image and a clustered-hyperspectral information achieved by using unsupervised clustering; 2) augmented superpixels along with the hyperspectral image are used by an unsupervised region segmentation module to achieve the final segmentation. In the following, each step of the pipeline is presented.

4.1.1 Augmented Hyperspectral SLIC superpixels

Hyperspectral remote sensing images are usually very large images that, due to the imaging conditions, may be afflicted by different artifacts. As a consequence, superpixels calculated on hyperspectral images may present imprecise boundaries. To improve the goodness of superpixels, a modification of the original SLIC algorithm is proposed. This new version of SLIC is composed of two steps: a preliminary unsupervised clustering using the mean-shift algorithm [189] whose output (that here is defined as clustered-hyperspectral information) is used together with the original image by the SLIC algorithm.

4.1.1.1 Unsupervised clustering

The pipeline of the augmented superpixels method is depicted in Figure 4.2. Given an input hyperspectral image I made of N pixels $P_i = (b_1^i, \cdots, b_L^i)$ and L bands, pixel similarity information is extracted using the mean-shift algorithm [190] which is an unsupervised clustering algorithm that iteratively finds the best number of clusters U that better fits the input data. Each cluster center is defined by the mean of pixels of the hyperspectral

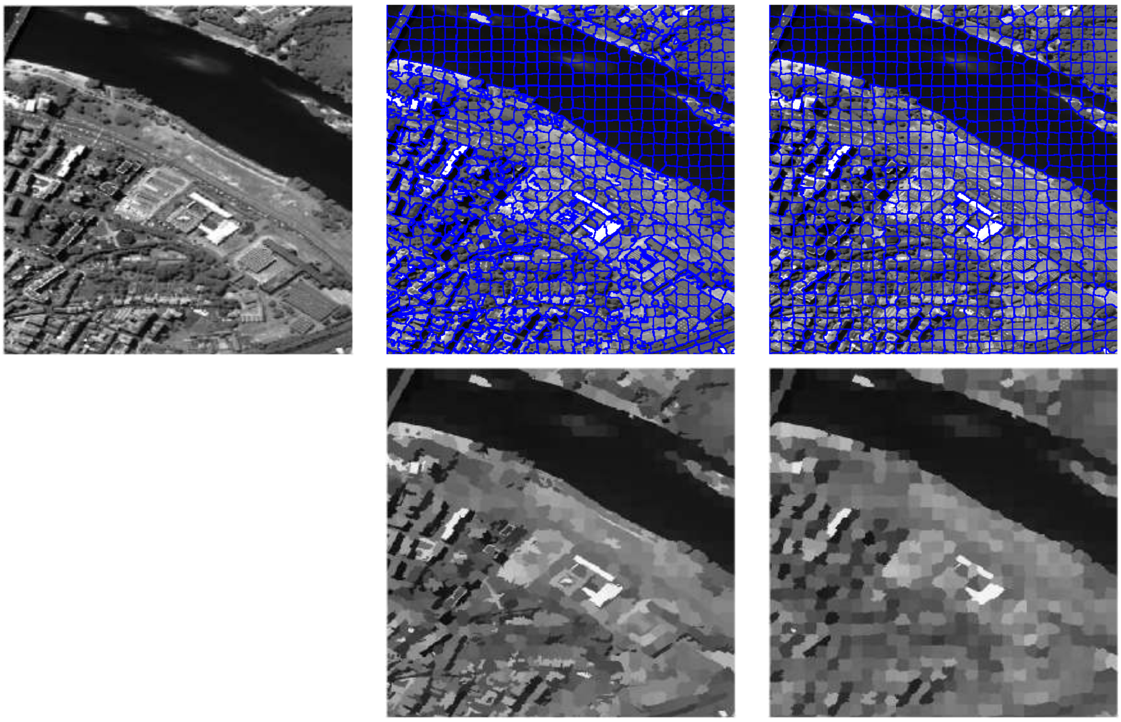


Figure 4.3: Results of the modified SLIC algorithm. On the left, it is possible to see a projection of a portion of the Pavia Center dataset on a single band, while in the center and on the right it is shown the results of the modified SLIC algorithm, respectively with $m = 0.2$ and $m = 1$. In these examples, m_{clust} is fixed to 0.

image that are assigned to the u-th cluster defined as follows: $Q_u = (q_1^u, \cdots, q_L^u)$.

To avoid outliers in the spectral signal, a normalization of the hyperspectral image is performed by considering its maximum value V as for the 95% of the spectral data. All the values of the image are clipped between 0 and V and then divided by the same value V so that the final image is normalized between 0 and 1.

Once the algorithm has found the clusters, the spectral information of each pixel of the image is concatenated with the center of the cluster to which it belongs to. The resulting vector describing a pixel at a position x_i and y_i is therefore $F_i = \langle P_i, Q_u, x_i, y_i \rangle$ which is of a size equal to $(2 \times L) + 2$. The augmented hyperspectral image $\{P_i, Q_u, x_i, y_i\}_{i=1}^N$ is the input of a modified version of the SLIC method.

4.1.1.2 Modified SLIC

The original SLIC algorithm takes as input RGB color and spatial information of the image, namely N pixels $P_i = (b_R^i, b_G^i, b_B^i)$ at position x_i and y_i, and it exploits the k-means algorithm to cluster them into superpixels [187]. The algorithm initially considers a number K of superpixel cluster centers C_k taken at regular grid intervals $S = \sqrt{N/K}$. The higher is the number of K and the smaller is the size of the initial superpixels. Ideally, S^2 represents the area of each superpixel. To assign the pixel to the cluster k a search region of $2S \times 2S$ around the cluster center is used. This strategy reduces the complexity of the SLIC algorithm so that it is linear to the number of pixels N and independent from the number of superpixels K.

After the initialization of the grid, the algorithm iteratively assigns each pixel P_i to the nearest superpixel, whose search area overlaps the pixel itself. For each iteration, the assignment is determined by a distance that, in the original formulation of the SLIC algorithm, is defined as follows:

$$D_s = d_{rgb} + m\left(\frac{d_{xy}}{S}\right) \tag{4.1}$$

where m is a parameter used to control the regularity of superpixels using spatial information, while d_{rgb} and d_{xy} are respectively, the color and spatial differences between a pixel and the center of the corresponding superpixel.

The algorithm repeats the assignment process between pixels and superpixels until the Euclidian distance between the old centers and the new centers is lower than a certain threshold.

In the modified version of the SLIC algorithm, instead of D_s, a new distance D_{ahs} is defined in order to handle the hyperspectral image and the clustered hyperspectral information at the same time. The new distance D_{ahs} between a pixel F_i and k-th superpixel center C_k is defined as follows:

$$D_{ahs} = \frac{d_{spec}}{\sqrt{L}} + m_{clust}\left(\frac{d_{clust}}{\sqrt{L}}\right) + m\left(\frac{d_{xy}}{S\sqrt{2}}\right) \tag{4.2}$$

where

$$d_{spec} = \sqrt{\sum_{j=1}^{L}(b_j^k - b_j^i)^2} \tag{4.3}$$

$$d_{clust} = \sqrt{\sum_{j=1}^{L}(q_j^k - q_j^u)^2} \tag{4.4}$$

and

$$d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \tag{4.5}$$

To make the distance D_{ahs} independent from the number of bands, the distances d_{spec} and d_{clust} are normalized by a factor $\sqrt{L}$ that is achieved considering the maximum L2 distance between two most diverse hyperspectral pixels. Since each hyperspectral band ranges between 0 and 1, the most diverse pixels are the following: $P_0 = (0, \cdots, 0)$ and $P_1 = (1, \cdots, 1)$. The d_{spec} (as well as d_{clust}) between these two pixels is therefore $\sqrt{L}$.

The same idea is applied for the normalization of the spatial information. Considering the standard search region of the SLIC algorithm, the maximum spatial distance between

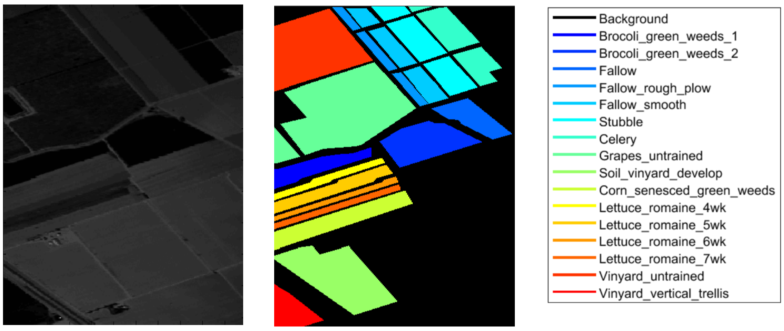


Figure 4.4: On the left it is possible to see a projection of the Salinas dataset on a single band, while on the right its ground truth and the corresponding color labeling. Note that the black color represents the background.

a superpixel center and a pixel in the search region is $2S/\sqrt{2} = S\sqrt{2}$.

The parameters m and m_{clust} are used to control the regularity of superpixels using spatial and clustered hyperspectral information respectively. Figure 4.3 shows an example of the proposed SLIC achieved on a hyperspectral image as m varies between 0 and 1. As you can see the higher is the value of m, the higher is the compactness of the superpixels that tend to resemble the square shape of standard pixels. In these examples, m_{clust} is fixed to 0, but analogous considerations can be done by varying the parameter m_{clust}.

4.1.2 Unsupervised region segmentation

This module takes as input the concatenation of the original hyperspectral image $\{P_i = (b_1^i, \cdots, b_L^i)\}_{i=1}^N$ with the corresponding centers of the superpixels $\{C_k = (b_1^k, \cdots, b_L^k, x_k, y_k)\}_{i=1}^K$ that are the average color (over the L bands of the hyperspectral image) of the pixels and their spatial positions. The input, in short $\langle P_i, C_k, x_k, y_k \rangle$, is processed by the mean-shift algorithm [189] along with a major voting strategy to generate the final segmentation map $\{S_t\}_{t=1}^T$.

The use of mean-shift is motivated by the fact that, differently from other clustering methods, the knowledge of the number of clusters to be predicted is not required, moreover, it is demonstrated to reduce the number of mislabeled samples with respect to other methods in the state of the art [181].

At the end of the clustering, the segmentation results are further improved by eliminating small regions that are likely due to noise. This is done by re-assigning a label to a given region that is more frequent in its neighborhood.

4.2 Hyperspectral Datasets

The first two datasets used are Salinas and SalinasA (respectively in figure 4.4 and 4.5) [25]. These images have been collected by the AVIRIS sensor and present 224 bands in the

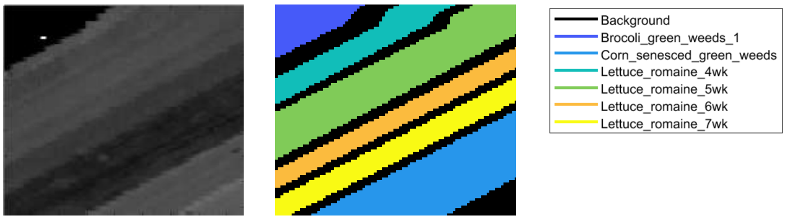


Figure 4.5: On the left it is possible to see a projection of the SalinasA dataset on a single band, while on the right its ground truth and the corresponding color labeling. Note that the black color represents the background.

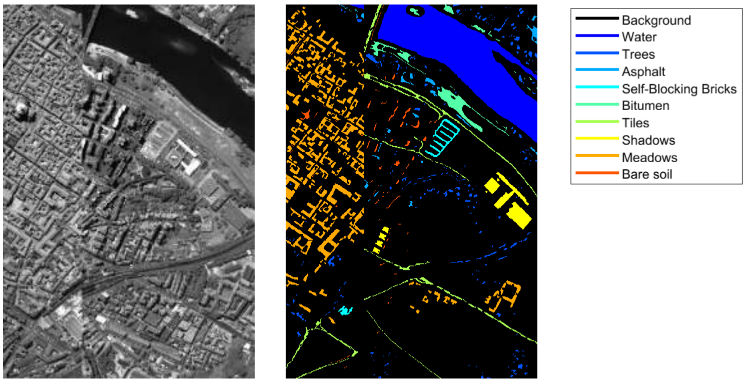


Figure 4.6: On the left it is possible to see a projection of the Pavia Center dataset on a single band, while on the right its ground truth and the corresponding color labeling. Note that the black color represents the background.

400–2500nm portion of the spectrum, where 20 noisy bands in the region of water absorption have been discarded, resulting in two images with 204 channels. The dimensions of the datasets are respectively 512x217 and 86x83 with a high spatial resolution of 3.7m per pixel. The labeling is different between the two datasets. Salinas presents 16 classes that represent the region of culture such as broccoli, fallow, grapes, etc., while SalinasA is segmented in a subset of the former dataset, representing only 6 classes from broccoli to corn and different variations of lettuce.

The Pavia Center and Pavia University datasets [25] have also been considered. They are respectively images of 1096x1096 with 102 channels and 610x610 with 103 channels (see figure 4.6 and 4.7). However, in both images part of the samples have been discarded because the information was missing, resulting in two images respectively of 1096x715 and 610x340. The two scenes have been acquired by the ROSIS sensor with a geometric resolution of 1.3 meters. These datasets are both annotated with 9 labels typical of a city such as asphalt, meadows, trees, bare soil, etc..

4.3 Evaluation metrics

The method is composed of two modules. The former segments the input image using superpixels while the latter clusters data using spectral and spatial information to perform

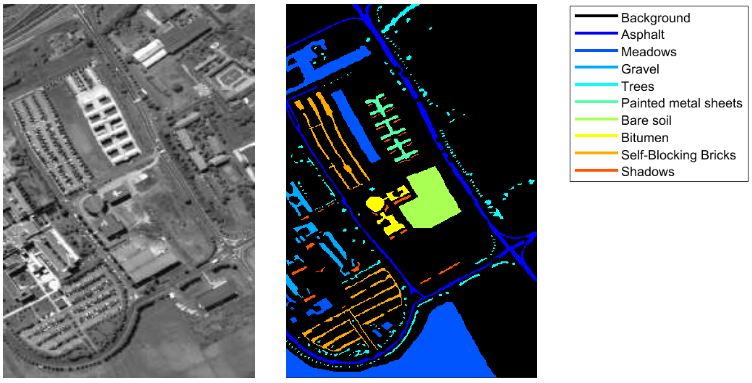


Figure 4.7: On the left it is possible to see a projection of the Pavia University dataset on a single band, while on the right its ground truth and the corresponding color labeling. Note that the black color represents the background.

the final segmentation. Following the guidelines of Achanta et al. [187], the superpixel segmentation step has been evaluated with the UE (Undersegmentation Error), while the second step has been evaluated by using ARI (Adjusted Rand Index), NMI (Normalized Mutual Information) and F-measure [80].

4.3.1 Superpixel Segmentation - Undersegmentation error

An undersegmentation error (UE) occurs when pixels belonging to different classes considered in the task are grouped together into a single region/class. Given a region of the ground truth g_i, the UE is defined as follows:

$$UE = \frac{1}{N}\left[\sum_{i=1}^{M}\Big(\sum_{S_j|S_j\cap g_i>B}|S_j|\Big) - N\right] \tag{4.6}$$

where M is the number of ground truth segments, B is a minimum number of pixels in S_j overlapping g_i and N is the number of pixels of the image. B is used to compensate for possible errors in the ground truth segmentation data. The lower is the UE and the better is the method.

4.3.2 Unsupervised region segmentation

To measure the performance of the whole method, three evaluation metrics have been used: normalized mutual information (NMI), adjusted rand index (ARI) [191, 192], and F1-score [193]. NMI and ARI are defined as:

$$NMI = \frac{\sum_i \sum_j n_{ij} \log(\frac{n \cdot n_{ij}}{n_i \cdot n_j})}{\sqrt{\sum_i n_i \log \frac{n_i}{n} \sum_j n_j \log \frac{n_j}{n}}} \tag{4.7}$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}] - [\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}]/\binom{n}{2}} \tag{4.8}$$

where n is the total number of samples, n_i is the number of samples in a cluster i, n_j is the number of samples in class j, and n_{ij} is the number of samples in both clusters i and class j. For both the NMI and ARI, the higher is the value the better is the method.

F1-score [193] for unsupervised clustering, where the number of classes is unknown is defined as:

$$Precision = \frac{\sum_k \max_s\{a_{ks}\}}{\sum_k \sum_s a_{ks}} \tag{4.9}$$

$$Recall = \frac{\sum_s \max_k\{a_{ks}\}}{\sum_k \sum_s a_{ks}} \tag{4.10}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.11}$$

where k is the number of clusters predicted, s is the number of classes, and a_{ks} denotes the number of samples clustered to cluster k and belonging to class s.

4.4 Results

The results section is divided into two subsections, the first one discusses the results achieved with just the augmented version of the superpixel segmentation while the second one presents the results with the entire pipeline applied for unsupervised region segmentation.

4.4.1 Superpixel segmentation

The metric adopted for the evaluation of the superpixel segmentation is the under-segmentation error (UE), which strongly depends on the precision of both the segmentation and the ground truth on the boundaries [187]. Since the most precise ground truth annotation is available for the Salinas and SalinasA datasets, the focus here is only on these datasets. Due to the imprecision of the ground truths on the boundaries derived from the remote sensing nature of the images, the parameter B of the equation (4.6) has been fixed to the 15% of the pixels in S_j.

In the experiments, the number of superpixels required is set to $K = 1000$ for Salinas, while $K = 500$ for SalinasA. The bandwidth for the clustering has been empirically fixed to 0.1 for both datasets.

The augmented SLIC segmentation has two weight parameters: spatial m and clustered information m_{clust}. By setting the value $m_{clust} = 0$ the augmented SLIC is turned into the original SLIC applied to a hyperspectral image.

Table 4.1: Evaluation of superpixel segmentation on (a) SalinasA and (b) Salinas datasets with under-segmentation error. The table shows the error considering difference values of m and m_{clust}. For each column, cells in orange show the maximum values (worst results). Values in bold-face represent the best value for each column while underlined values show the best value overall.

	$m = 0.2$	$m = 0.4$	$m = 0.6$	$m = 0.8$	$m = 1$
$m_{clust} = 0$	0.2148	0.2102	0.2102	0.2098	0.2098
$m_{clust} = 0.2$	0.2088	0.2073	0.2067	0.2067	0.2067
$m_{clust} = 0.4$	0.2051	0.2080	0.2071	0.2067	0.2067
$m_{clust} = 0.6$	0.2053	0.2088	0.2073	0.2067	0.2067
$m_{clust} = 0.8$	**0.2030**	0.2061	0.2047	0.2043	0.2040
$m_{clust} = 1$	0.2055	**0.2050**	**0.2032**	**0.2024**	**0.2018**

(a) SalinasA

	$m = 0.2$	$m = 0.4$	$m = 0.6$	$m = 0.8$	$m = 1$
$m_{clust} = 0$	0.2766	0.2856	0.2938	0.2998	0.3019
$m_{clust} = 0.2$	0.2714	0.2845	0.2986	0.3087	0.3066
$m_{clust} = 0.4$	0.2661	0.2867	0.2954	0.3011	0.3065
$m_{clust} = 0.6$	0.2649	0.2828	0.2874	0.2986	0.3071
$m_{clust} = 0.8$	**0.2575**	0.2800	0.2846	0.2950	0.3015
$m_{clust} = 1$	0.2583	**0.2773**	**0.2844**	**0.2911**	**0.2985**

(b) Salinas

In table 4.1(a), the UE achieved by the algorithm is reported with different values of m and m_{clust} on Salinas A. The lower is the value and the better is the result. The first row corresponds to the original SLIC while the other rows correspond to the proposed algorithm with different settings of the parameter m_{clust}. Whatever is the value of m, with m_{clust} higher than zero, better results than the original SLIC are always achieved. As visible from the table, the best value is when $m = 0.2$ and $m_{clust} = 0.8$.

The table 4.1(b) reports the UE on Salinas and it also demonstrates and confirms that considering also the clustered information improves the results of the superpixel segmentation. Even in this case, the best overall result is achieved when $m = 0.2$ and $m_{clust} = 0.8$.

4.4.2 Unsupervised region segmentation

The entire pipeline for unsupervised segmentation has been evaluated on all the four datasets described in section 4.2. For the sake of comparison, the unsupervised region segmentation has been evaluated by considering different types of input information (see the pipeline in Figure 4.1) using both the k-means (assuming known the number of classes) and the mean-shift:

- hyperspectral information: the clustering algorithm applied directly to the input hyperspectral image;
- superpixels centers: the clustering algorithm applied to the superpixels C_k;
- hyperspectral and superpixels centers: the clustering algorithm applied to the concatenation of superpixels C_k and the input hyperspectral image;
- reduced hyperspectral information: the clustering algorithm applied to the output of the feature reduction phase that is based on Principal Component Analysis (PCA) applied to the input hyperspectral image;

Table 4.2: The number of bands before and after the application of PCA is shown for both the original spectral information of the image and the information of the superpixels centers.

Dataset	Spectral original	Superpixels original	Spectral PCA	Superpixels PCA
Pavia Center	102	104	15	7
Pavia University	103	105	16	8
Salinas	204	206	13	9
SalinasA	204	206	17	10

Table 4.3: The table shows the performance on the Pavia Center dataset using different combinations of methods and inputs. In the mean-shift methods, the table reports the results on each metric with a fixed bandwidth, that has been selected considering the best results achieved on NMI.

Clustering method	Input Information	PCA	ARI	NMI	F1
K-means	Spectral	No	0.78	0.76	0.82
K-means	Spectral	Yes	0.78	0.76	0.82
K-means	Superpixels	No	0.78	0.70	0.80
K-means	Superpixels	Yes	0.77	0.69	0.81
K-means	Spectral + Superpixels	No	0.78	0.74	0.81
K-means	Spectral + Superpixels	Yes	0.79	0.77	0.84
Mean-shift	Spectral	No	0.80	0.77	0.85
Mean-shift	Spectral	Yes	0.80	0.78	0.84
Mean-shift	Superpixels	No	0.82	0.75	0.88
Mean-shift	Superpixels	Yes	0.82	0.75	0.88
Mean-shift	Spectral + Superpixels	No	**0.88**	**0.87**	**0.90**
Mean-shift	Spectral + Superpixels	Yes	**0.88**	0.86	0.89

- reduced superpixels centers: the clustering algorithm applied to the superpixels centers achieved on the reduced hyperspectral information;
- reduced hyperspectral and superpixels centers: the clustering algorithm applied to the concatenation of superpixels C_k and the input hyperspectral image after a feature reduction using PCA;

When using PCA, 99.9% of the total variance represented by each principal component is considered. The table 4.2 shows the reduction of the number of bands after the application of the PCA for each dataset.

In the tables 4.3, 4.4, 4.5, and 4.6, mean-shift methods with k-means applied with the correct and previously known number of clusters are compared. All the experiments have the same values fixed for the parameters $m = 0.4$ and $m_{clust} = 0.8$ but a different number of superpixels dependent on the dataset (2000 for Pavia center and Pavia University, 800 for Salinas, 300 for SalinasA). Empirically, the bandwidth parameter of mean-shift has been selected based on the better NMI performance. The figures 4.8, 4.9, 4.10, 4.11 show the results of the proposed algorithm, considering that the color map of the labels between ground truth and segmentation is not the same.

The results show that for each of the datasets, better results are achieved when the combination of mean-shift, spectral information, and superpixel centers is involved. PCA

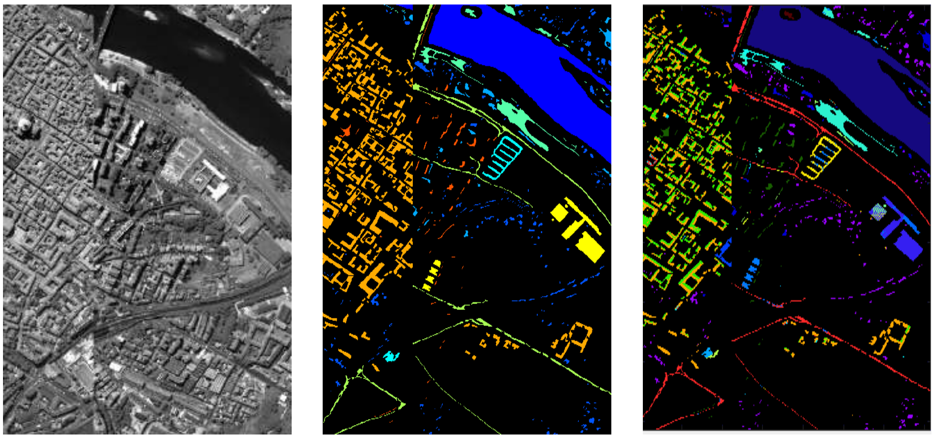


Figure 4.8: On the left there is the Pavia Center dataset, in the center it is shown the Pavia Center ground truth, and on the right the unsupervised region segmentation results with the best NMI achieved. The performance on the metrics are $ARI = 0.88$*,* $NMI = 0.87$ *and* $F1 = 0.90$.

Table 4.4: The table shows the performance on the Pavia University dataset using different combinations of methods and inputs. In the mean-shift methods, the table reports the results on each metric with a fixed bandwidth, that has been selected considering the best results achieved on NMI.

Clustering method	Input Information	PCA	ARI	NMI	F1
K-means	Spectral	No	0.31	0.57	0.60
K-means	Spectral	Yes	0.32	0.58	0.65
K-means	Superpixels	No	0.32	0.54	0.63
K-means	Superpixels	Yes	0.32	0.52	0.63
K-means	Spectral + Superpixels	No	0.33	0.62	0.65
K-means	Spectral + Superpixels	Yes	0.33	0.56	0.62
Mean-shift	Spectral	No	0.47	0.58	0.74
Mean-shift	Spectral	Yes	0.47	0.58	0.74
Mean-shift	Superpixels	No	0.42	0.49	0.74
Mean-shift	Superpixels	Yes	0.38	0.44	0.71
Mean-shift	Spectral + Superpixels	No	**0.59**	**0.72**	**0.84**
Mean-shift	Spectral + Superpixels	Yes	0.57	0.69	0.82

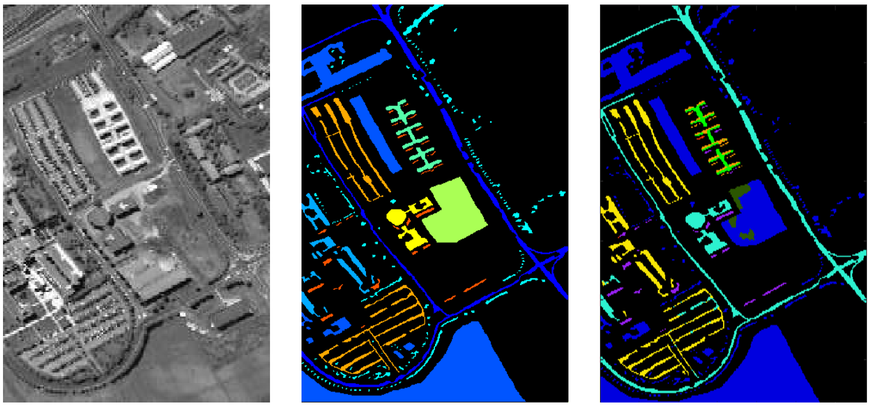


Figure 4.9: On the left there is the Pavia University dataset, in the center it is shown the Pavia University ground truth, and on the right the unsupervised region segmentation results with the best NMI achieved. The performance on the metrics are $ARI = 0.59$*,* $NMI = 0.72$ *and* $F1 = 0.84$.

Table 4.5: The table shows the performance on the Salinas dataset using different combinations of methods and inputs. In the mean-shift methods, the table reports the results on each metric with a fixed bandwidth, that has been selected considering the best results achieved on NMI.

Clustering method	Input Information	PCA	ARI	NMI	F1
K-means	Spectral	No	0.56	0.80	0.74
K-means	Spectral	Yes	0.59	0.81	0.77
K-means	Superpixels	No	0.60	0.81	0.77
K-means	Superpixels	Yes	0.66	0.84	0.81
K-means	Spectral + Superpixels	No	0.62	0.82	0.79
K-means	Spectral + Superpixels	Yes	0.56	0.82	0.76
Mean-shift	Spectral	No	0.69	0.88	0.84
Mean-shift	Spectral	Yes	0.69	0.87	0.84
Mean-shift	Superpixels	No	0.70	0.86	0.84
Mean-shift	Superpixels	Yes	0.71	0.87	0.84
Mean-shift	Spectral + Superpixels	No	**0.85**	**0.91**	**0.90**
Mean-shift	Spectral + Superpixels	Yes	0.82	**0.91**	0.89

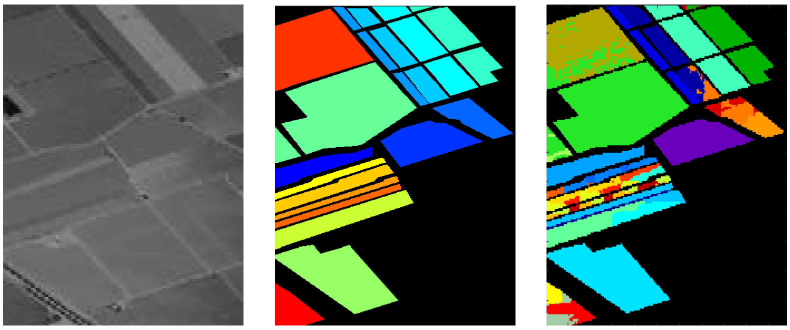


Figure 4.10: On the left there is the Salinas dataset, in the center it is shown the Salinas ground truth, and on the right the unsupervised region segmentation results with the best NMI achieved. The performance on the metrics are $ARI = 0.85$, $NMI = 0.91$ and $F1 = 0.90$.

allows the reduction in dimensionality of the feature vectors but has not shown coherent improvements on all the datasets.

Results reported above have been obtained by heuristically tuning the following parameters:

- the bandwidth of mean-shift
- the weight m of the spatial information
- the weight m_{clust} of the spectral similarity information

The bandwidth of mean-shift controls the number of classes that are extracted by the clustering algorithm. The higher is the bandwidth, the lower the number of clusters.

Table 4.6: The table shows the performance on the SalinasA dataset using different combinations of methods and inputs. In the mean-shift methods, the table reports the results on each metric with a fixed bandwidth, that has been selected considering the best results achieved on NMI.

Clustering method	Input Information	PCA	ARI	NMI	F1
K-means	Spectral	No	0.80	0.90	0.89
K-means	Spectral	Yes	0.80	0.90	0.89
K-means	Superpixels	No	0.82	0.90	0.90
K-means	Superpixels	Yes	0.58	0.79	0.77
K-means	Spectral + Superpixels	No	0.82	0.90	0.90
K-means	Spectral + Superpixels	Yes	0.82	0.90	0.90
Mean-shift	Spectral	No	0.73	0.84	0.83
Mean-shift	Spectral	Yes	0.73	0.84	0.83
Mean-shift	Superpixels	No	0.69	0.80	0.83
Mean-shift	Superpixels	Yes	0.72	0.83	0.85
Mean-shift	Spectral + Superpixels	No	**0.90**	**0.95**	**0.95**
Mean-shift	Spectral + Superpixels	Yes	**0.90**	**0.95**	**0.95**

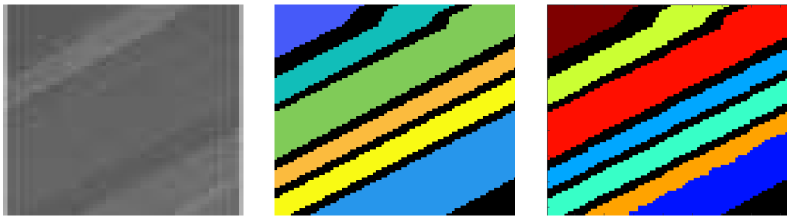


Figure 4.11: On the left there is the SalinasA dataset, in the center it is shown the SalinasA ground truth, and on the right the unsupervised region segmentation results with the best NMI achieved. The performance on the metrics are $ARI = 0.90$, $NMI = 0.95$ and $F1 = 0.95$.

The parameter m modulates the compactness of superpixels, so lower values are more suitable for high-resolution images, while m_{clust} is related to smoothing and therefore to noise reduction.

The number K of superpixels depends mostly on the dimensions of the hyperspectral image, with K that increases in correspondence with larger and cluttered images. It has been empirically found out that the optimal value for K is given by the approximation to the nearest hundred of the ratio between the smallest image dimension and a scaling parameter c_1:

$$K = \lceil \frac{min(H, W)}{\alpha \cdot 100} \rceil \cdot 100 \tag{4.12}$$

where H is the height, W is the width of the image and α is the scaling factor that values 60. K ranges from a minimum of 300 and a maximum of 2000 superpixels.

A variant of the proposed method that exploits bandwidth values automatically determined by following the procedure proposed by Pedregosa et al. [191] has been experimented with. This version of the proposed method is referred to as `OUR_BW`, while the version of the proposed method with all the parameters heuristically set is referred to as `OUR`.

Table 4.7: Comparison of the proposed method with other methods in the state of the art using NMI and ARI. The table shows the ARI and NMI values achieved by every method on the datasets of Pavia Center, Pavia University, Salinas, and SalinasA.

	Pavia Center	Pavia Univ.	Salinas	SalinasA	**Average**	Require number
	NMI - ARI	NMI - ARI	NMI - ARI	NMI - ARI	NMI - ARI	of classes
K-means [80]	0.85 - 0.82	0.40 - 0.63	0.63 - 0.83	0.67 - 0.78	0.64 - 0.77	Yes
GMM [80]	0.77 - 0.74	0.29 - 0.53	0.53 - 0.79	0.78 - 0.87	0.59 - 0.73	Yes
HNMF [57]	0.85 - 0.77	0.38 - 0.57	0.53 - 0.79	0.78 - 0.87	0.64 - 0.75	Yes
SMCE [81]	0.80 - 0.77	0.31 - 0.56	0.57 - 0.78	0.76 - 0.81	0.61 - 0.73	Yes
DLSS [19]	0.52 - 0.42	0.49 - 0.57	0.37 - 0.39	0.63 - 0.81	0.55 - 0.50	Yes
3D-CAE [69]	0.96 - 0.86	0.36 - 0.59	0.67 - 0.85	0.77 - 0.87	0.69 - 0.79	Yes
DEC [79]	0.83 - 0.80	0.41 - 0.67	0.57 - 0.80	0.78 - 0.87	0.65 - 0.79	Yes
BDEC [80]	**0.97** - **0.91**	**0.60** - 0.70	0.68 - 0.87	0.81 - 0.87	0.77 - 0.84	Yes
OUR_BW	0.81 - 0.80	0.53 - 0.70	0.67 - 0.87	0.82 - 0.92	0.71 - 0.82	**No**
OUR	0.88 - 0.87	0.59 - **0.72**	**0.85** - **0.91**	**0.90** - **0.95**	**0.81** - **0.86**	**No**

Finally, in table 4.7 the ARI and NMI performance of the two variants of the proposed method are compared with other unsupervised segmentation methods in the state of the art [80]. All the state-of-the-art methods, whose results are reported in table 4.7, require the tuning of some parameters, some form of a-priori knowledge or a training process.

The proposed method does not need to know the exact number of classes to be identified and does not use any external knowledge about the image content and, more importantly, it is a handcrafted method that does not require the training of a specific model.

This proposal achieves, on average, the best results in terms of NMI and ARI on the considered datasets. In particular, it outperforms all the other methods on Salinas, SalinasA, and Pavia University, while it achieves lower performance on Pavia Center. The algorithm version with automatic bandwidth achieves on average comparable results with respect to other methods.

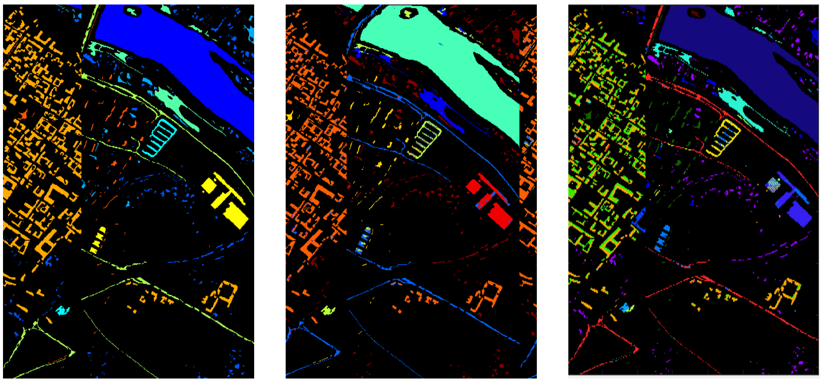
Figures 4.12, 4.13, 4.14, and 4.15 show a comparison between the segmentation results achieved by our method and BDEC technique [80]. The results of BDEC have been retrieved by using the code available from the respective article and by reconstructing the segmented images.

4.5 Robustness to Noise

Taking inspiration from Nalepa et al. [194], the robustness of our method to several types of noise has been assessed. For each dataset, a new hyperspectral image $I^{'}$ has been defined. This image is the result of adding noise N to the hyperspectral images I as follows: $I^{'} = I + N$.

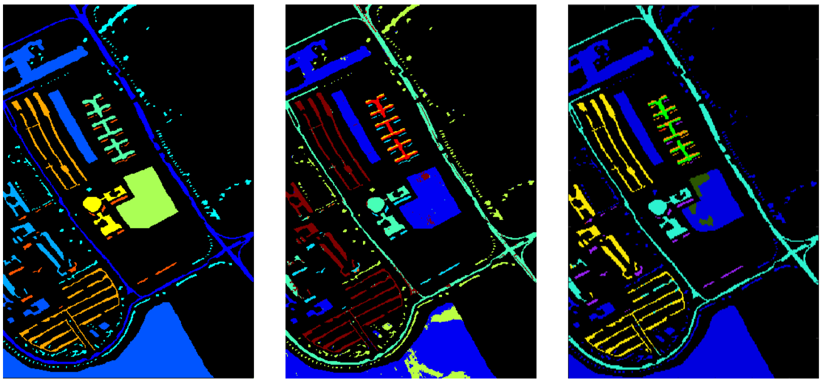
The considered types of noise are:

- Gaussian noise



(a) Ground-truth (b) BDEC (c) Our

Figure 4.12: On the left there is the Pavia Center ground truth, in the center the results of the BDEC method, and on the right the results achieved by our technique. The performance of BDEC are $ARI = 0.97$ and $NMI = 0.91$. The performance of Our method are $ARI = 0.88$ and $NMI = 0.87$.



(a) Ground-truth (b) BDEC (c) Our

Figure 4.13: On the left there is the Pavia University ground truth, in the center the results of the BDEC method, and on the right the results achieved by our technique. The performance of BDEC are $ARI = 0.60$ and $NMI = 0.70$. The performance of Our method are $ARI = 0.59$ and $NMI = 0.72$.

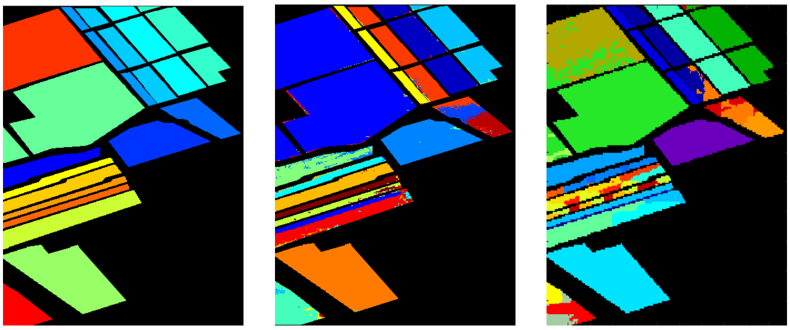
- Impulsive noise (salt & pepper)
- Poisson noise

4.5.1 Gaussian noise

The Gaussian noise simulates thermal and quantization disturbances [194]. The noise signal is defined by a normal distribution probability density function. The probability p for a variable x, with mean μ and the variance σ is defined as:

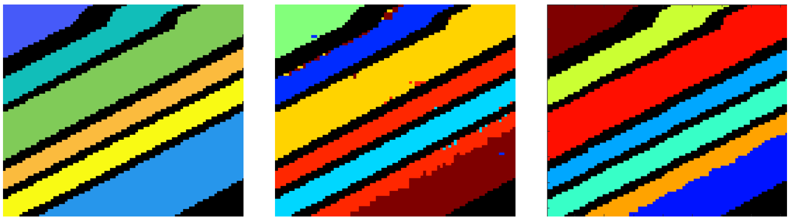
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{4.13}$$

In the experiments, noise has been applied to the 10% of the pixels of the image and, for each of the evaluations, the mean $\mu = 0$ and different values of variance σ: 0, 0.01, 0.05, 0.1, 0.25, and 0.5 have been considered. Table 4.8 reports the results of this investigation.



(a) Ground-truth (b) BDEC (c) Our

Figure 4.14: On the left there is the Salinas ground truth, in the center the results of the BDEC method, and on the right the results achieved by our technique. The performance of BDEC are $ARI = 0.68$ *and* $NMI = 0.87$*. The performance of Our method are* $ARI = 0.85$ *and* $NMI = 0.91$*.*



(a) Ground-truth (b) BDEC (c) Our

Figure 4.15: On the left there is the SalinasA ground truth, in the center the results of the BDEC method, and on the right the results achieved by our technique. he performance of BDEC are $ARI = 0.81$ *and* $NMI = 0.87$*. The performance of Our method are* $ARI = 0.90$ *and* $NMI = 0.95$*.*

As expected, the results show a reduction in performance when the noise is particularly disturbing. This is particularly true in the case of Salinas and SalinasA datasets when a high variance is considered. Overall, the results show that the proposed method is robust to Gaussian noise signal when the original signal is not too deteriorated.

4.5.2 Impulsive noise

The impulsive noise represents an error in the acquisition of data, where a pixel remained saturated (”white” pixel) or where the data for a certain pixel is lost (”black” pixels) [194].

In the experiments, salt & pepper noise has been applied with density on the image. Specifically, 0, 0.01, 0.05, 0.1, 0.25, and 0.5 have been considered as the percentage of the pixels affected by noise. The table 4.9 shows the results for every density and dataset.

Results show that this kind of noise does not impact significantly on the overall performance. Oscillatory performance is justified by the stochasticity of the entire process.

Table 4.8: The table shows the robustness of the proposed method to the addition of Gaussian Noise for each of the datasets with different variances.

Variance	Pavia Center	Pavia Univ.	Salinas	SalinasA	**Average**
	ARI - NMI	ARI - NMI	ARI - NMI	ARI - NMI	ARI - NMI
0	0.81 - 0.80	0.53 - 0.70	0.67 - 0.87	0.82 - 0.92	0.71 - 0.82
0.01	0.81 - 0.78	0.53 - 0.69	0.54 - 0.79	0.90 - 0.95	0.70 - 0.80
0.05	0.81 - 0.78	0.56 - 0.68	0.48 - 0.74	0.82 - 0.90	0.67 - 0.78
0.1	0.84 - 0.81	0.56 - 0.65	0.46 - 0.71	0.73 - 0.84	0.65 - 0.75
0.25	0.84 - 0.81	0.55 - 0.65	0.45 - 0.71	0.34 - 0.56	0.55 - 0.68
0.5	0.83 - 0.80	0.54 - 0.64	0.38 - 0.62	0.24 - 0.51	0.50 - 0.64

Table 4.9: The table shows the robustness of our method to the addition of Impulsive Noise for each of the datasets with different densities of pixels.

Pixel density	Pavia Center	Pavia Univ.	Salinas	SalinasA	**Average**
	ARI - NMI	ARI - NMI	ARI - NMI	ARI - NMI	ARI - NMI
0	0.81 - 0.80	0.53 - 0.70	0.67 - 0.87	0.82 - 0.92	0.71 - 0.82
0.01	0.81 - 0.79	0.53 - 0.70	0.66 - 0.87	0.75 - 0.90	0.69 - 0.82
0.05	0.81 - 0.80	0.52 - 0.68	0.66 - 0.86	0.82 - 0.92	0.70 - 0.82
0.1	0.80 - 0.78	0.52 - 0.68	0.66 - 0.86	0.90 - 0.95	0.72 - 0.82
0.25	0.81 - 0.78	0.50 - 0.67	0.58 - 0.82	0.90 - 0.95	0.70 - 0.81
0.5	0.80 - 0.78	0.57 - 0.69	0.52 - 0.77	0.83 - 0.90	0.66 - 0.79

4.5.3 Poisson noise

The Poisson noise models a signal-dependent photon noise [194]. The signal is defined as the following probability density function p for a variable x:

$$p(x) = \frac{e^{\lambda}\lambda^{x}}{x!} \tag{4.14}$$

where λ represents the expected average value, which we considered to be 5.5 for all the experiments.

In the table 4.10, the results for every dataset with and without noise are shown.

The results show that our method is robust to the presence of Poisson noise on each dataset.

4.6 Computational time and complexity

In this section, the real-time computation for all of the datasets, and the complexity of the entire method are reported. The time complexity of the whole method is $O(n^2)$ where n is the number of pixels of the image. Table 4.11 shows the time in seconds for each part of our method: Augmented H-SLIC, Unsupervised Segmentation, and the total algorithm. Note that, on average, the time required by the augmented H-SLIC is roughly 57% of the total amount of time required for running the algorithm. Note also that the computational

Table 4.10: The table shows the robustness of our method to the addition of Poisson Noise for each of the datasets.

Poisson noise	Pavia Center ARI - NMI	Pavia Univ. ARI - NMI	Salinas ARI - NMI	SalinasA ARI - NMI	**Average** ARI - NMI
No	0.81 - 0.80	0.53 - 0.70	0.67 - 0.87	0.82 - 0.92	0.71 - 0.82
Yes	0.81 - 0.80	0.52 - 0.69	0.66 - 0.87	0.80 - 0.92	0.70 - 0.82

Table 4.11: The table shows the time computation in seconds for Pavia Center, Pavia University, Salinas, and SalinasA datasets, for each step and the entire pipeline.

	Pavia Center	Pavia Univ.	Salinas	SalinasA	**Average**
Augmented H-SLIC	1097.01s	343.16s	388.67s	43.44s	468.07s
Unsupervised Segmentation	835.57s	218.53s	303.06s	19.04s	344.05s
Total	1944.61s	564.87s	693.88s	62.94s	816.58s

time depends on the size of the image, for example, the execution of the algorithm on the SalinasA dataset requires about $63s$ which is eleven times lower than the Salinas dataset.

4.7 Advatanges of the proposed method

With the increase of available data from earth observation drones and satellites, it is very important to reduce the effort on segmenting/labeling remote sensing images. To this end, techniques that, different from data-hungry methods such as deep learning, do not rely on large training sets and do not require a-priori knowledge and/or the number of classes to be segmented are necessary to increase the availability of labeled data. In this study, a method based on hand-crafted features that satisfy the above requirements has been presented. The method has been experimented on four different datasets thus proving its effectiveness in comparison with methods in the state of the art that in contrast may not satisfy all the above requirements.

A further feature of the proposed method is that it could be easily extended to deal with additional information obtained by other types of sensors. It is also worth to be investigated if a-priori knowledge about the image content can be exploited to improve the results, in particular in urban scenes.

Chapter 5

Multimodal Supervised Semantic Segmentation

This chapter presents a combination of works relative to *Ticino: A Multi-Modal Remote Sensing Dataset for Semantic Segmentation* [21], available at *SSRN* (Social Science Research Network), and *Multimodal Fusion Methods with Vision Transformers for Remote Sensing Semantic Segmentation* [22] that has been presented at the *WHISPERS 2023* conference (13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing) [1].

Land is the foundational element of human activities, on which all socio-economic dynamics are developed. It represents the fundamental resource for the survival of urban and rural residents [1]. In recent decades, population growth and the consequent increase in human activities have resulted in intensive land use [195]. However, the distribution of the population is affected by various human and time-varying factors that cause an uncontrolled exploitation of land resources in densely populated areas [196].

Monitoring the land coverage is therefore a duty of primary importance. A rapid evaluation of the soil's condition enables prompt mitigation of these consequences through the implementation of suitable strategies. Furthermore, the collection over time of land use and land cover enables to capture, model and predict past, present, and future dynamics of human activities [197]. For a precise land cover assessment, it is required to analyze images of the territory, manually or with the help of computer vision techniques [198].

The predominant computer vision technique is semantic segmentation, which is an effective technique used in many applications such as autonomous driving, robot navigation, industrial inspection, saliency object detection, agriculture sciences, medical imaging analysis, remote sensing, etc [199]. This technique classifies each individual pixel in the image. The output is then a map having the same spatial extent as the input image, where pixels are grouped into areas belonging to the same semantic class

[1] WHISPERS conference: `https://www.ieee-whispers.com/`

*Table 5.1: Comparison between state-of-the-art datasets for RS semantic segmentation and the Ticino dataset. Note that two versions of the dataset are presented. One in the original scale and one at higher resolution obtained through the cleaning procedure and the pansharpening processing described in section 5.2. (*computed using the other information in the table)*

Dataset	Sensor	Modalities	Area [km^2]	# images	Image size	Res. [m/pixel]	# bands	# classes
Deepglobe [83]	airborne	RGB	1716.9	1156	2448x2448	0.50	3	7
TorontoCity [84]	airborne	RGB/LIDAR	712.5	-	-	0.10	3/1	3
SpaceNet variant [85]	satellite	RGB	3254*	401755	300x300	0.3	3	2
INRIA [86]	airborne	RGB	810	360	1500x1550	0.30	3	2
Urban dataset [88]	airborne	RGB	3.46*	9	1000x1000	0.62	3	3
Coffee dataset [88]	airborne	NIR-RG	56.25*	9	1000x1000	2.50	3	3
Zurich Summer [89]	satellite	MS NIR-RGB	8.56	20	1000x1150	0.61	4	8
Indian Pines [25]	airborne	HS VNIR-SWIR	0.29*	1	145x145	3.70	200	16
Salinas [25]	airborne	HS VNIR-SWIR	1.52*	1	512x217	3.70	204	16
SalinasA [25]	airborne	HS VNIR-SWIR	0.10*	1	86x83	3.70	204	6
Pavia Center [25]	airborne	HS Visible	2.03*	1	1096x1096	1.30	102	9
Pavia University [25]	airborne	HS Visible	0.63*	1	610x610	1.30	103	9
SpaceNet [87, 90, 91]	satellite	PAN	3011	24586	650x650	0.31	1	2
		MS VNIR				1.24 (orig.)	8	
		MS SWIR				1.24 (orig.)	8	
ISPRS Potsdam [92, 90]	airborne	MS IR-RGB	3.42*	38	6000x6000	0.05	4	6
		PAN				0.05	1	
		DSM				0.05	1	
ISPRS Vaihingen [92, 90]	airborne	MS IR-RGB	1.34*	33	2500x2000	0.09	4	6
		PAN				0.09	1	
		DSM				0.09	1	
DSTL [93]	airborne	RGB	1	57	-	0.50	3	10
		PAN				0.31	1	
		MS VNIR				1.24	8	
		MS SWIR				7.50	8	
Ticino/Our	**satellite**	**RGB**	**1331.721**	**1502**	**256x362**	**1.86-2.64**	**3**	**8/10**
		PAN			**96x192**	**5**	**1**	
		HS VNIR			**16x32 (96x192)**	**30 (5)**	**63 (60)**	
		HS SWIR			**16x32 (96x192)**	**30 (5)**	**171 (122)**	
		DTM			**101x203**	**5**	**1**	

[90]. In remote sensing, semantic segmentation is incorporated in multiple fields such as precision farming, environmental monitoring, spatial planning enforcement, management of ecosystem-oriented natural resources such as food management, nature conservation, and many other important applications [2, 200, 201, 202, 203].

In line with other computer vision tasks, also semantic segmentation faced a step forward with the advent of deep learning techniques. In particular, two neural architectures are used: Convolutional Neural Networks [204, 100, 101, 102, 103] and Visual Transformers [121, 122, 120, 123, 124], which are more effective and efficient than traditional computer vision methods [90]. In remote sensing, some hybrid versions have also been used to achieve good performance in semantic segmentation [125, 126, 127, 128, 129, 130]. However, these methods are data-hungry and necessitate large datasets for training.

In the context of remote sensing, the fusion of different modalities can provide complementary information and improve the accuracy of image segmentation [205]. Apart from RGB [88, 83, 85, 86] and panchromatic images, which are readily available at a high-spatial resolution, the hyperspectral signal has shown great discriminative power than any other type of signals for identifying different materials [10]. However, hyperspectral devices have less spatial resolution in favor of a major number of acquired bands. The trade-off between resolution and the number of bands is due to higher production, computational, and management costs of hyperspectral devices with respect to RGB ones [10]. Due to these problems, the majority of single- [89] and multimodal datasets [93, 87, 91, 92] in the state of the art present only multispectral information, which is more discriminative than RGB but not as powerful as the hyperspectral one. The few datasets including hyperspectral information are generally composed of small single images that do not contain enough environmental variety and data to train deep-learning-based methods [25], leading to poor generalization capabilities. Nonetheless, in the literature, remote sensing fusion works have seen an exponential increase in recent years, focusing on the homogenous and heterogeneous fusion of complementary information such as spatio-temporal fusion and pansharpening, demonstrating the importance of pushing the research in this direction [94, 29, 133].

For these reasons, this research presents the Ticino dataset, a novel multimodal remote sensing dataset specifically tailored for semantic segmentation tasks. It covers an area of about 1332 km^2 and it incorporates five distinct modalities: RGB, Digital Terrain Model, Panchromatic, and Hyperspectral images, encompassing the visual-near infrared and short wavelength infrared portions of the electromagnetic spectrum. The RGB modality offers valuable spatial information, the Hyperspectral components contribute to effective material discrimination and finally, the Digital Terrain Model enhances our understanding of the soil morphology. Notably, the dataset includes labeled data for Land Cover and Soil Agricultural Use. To the best of the author's knowledge, the Ticino dataset is the largest, the most various (including over 230 spectral bands), multimodal dataset in the state of the art for RS semantic segmentation. Table 5.1 highlights the main characteristics of the Ticino dataset in comparison with semantic segmentation datasets in the state of the art. Among multimodal datasets, the Ticino dataset is the only one including the hyperspectral signal. Furthermore, the proposed dataset covers the largest area, thus including much more heterogeneity of the land in comparison with the other hyperspectral datasets. Finally, the Ticino dataset is the one with the highest number of modalities.

To facilitate future investigations, two comparative analyses were conducted, evaluating single modality and multimodality deep learning techniques. The first of these investigations is based on CNN models. The purposes of this analysis are, firstly, to provide a benchmark for future research with this dataset and, secondly, to compare single and multimodality approaches, as well as early and middle fusion strategies. The empirical findings of this analysis clearly demonstrate the superiority of multimodality over single modality methods, with middle fusion exhibiting the most substantial performance improvement. The second investigation, instead, is based on Transformers and focuses more on better understanding which fusion techniques are better to exploit the advantages of multimodality and yield better performance. For this purpose, six different fusion techniques were tested, dividing them into early, middle, and late fusion methodologies. This analysis remarks on the better achievements of multimodality and, at the same time, demonstrates the importance of using the correct techniques when the fusion is applied.

The main contributions of this research are:

- a multimodal Remote Sensing dataset that combines RGB, Hyperspectral, and Digital Terrain Model, having both high spatial and high spectral resolutions;
- a baseline that compares single- vs. multimodality deep learning techniques as well as early and middle data fusion techniques;
- a comprehensive analysis of a variety of multimodal fusion techniques based on Transformers.

5.1 Ticino Dataset

The Ticino multimodal satellite dataset has been collected from different sources of information. Specifically:

1. RGB data from Microsoft Bing Maps [206] (see Figure 5.1(a));
2. panchromatic and hyperspectral data from ASI PRISMA [6] (see Figure 5.1(b), 5.1(c) and 5.1(d));
3. digital terrain model of the area considered from Geoportal of Lombardia Region [207] (see Figure 5.1(e)).

The dataset also includes two different pixel-level labelings for semantic segmentation:

1. Land Cover collected from OpenStreetMaps [208] and Italian Agenzie delle Entrate [209] (see Figure 5.1(f));
2. Soil Agricultural Use collected from the Geoportal of Lombardia Region [207] (see Figure 5.1(g)).

The proposed dataset considers a territory around the Ticino river in the south of Milan and has an extension of 1332 km^2. This area has been chosen for its heterogeneity in terms of terrain composition and geomorphological variety. To support data-driven methods such as deep learning, the original dataset has been divided into 1808 smaller tiles. Among them, 306 have been discarded as they presented a number of labeled pixels inferior to 1%. The final dataset is therefore composed of 1502 georeferenced tiles. Each tile consists of five data sources and two pixel-level labelings. Figures 5.1(a-g) show the original images. Figures 5.1(h-n) show two examples of tiles extracted from the dataset. Figures 5.1(o-u) show the same tiles after a post-processing operation that involved pansharpening [20]

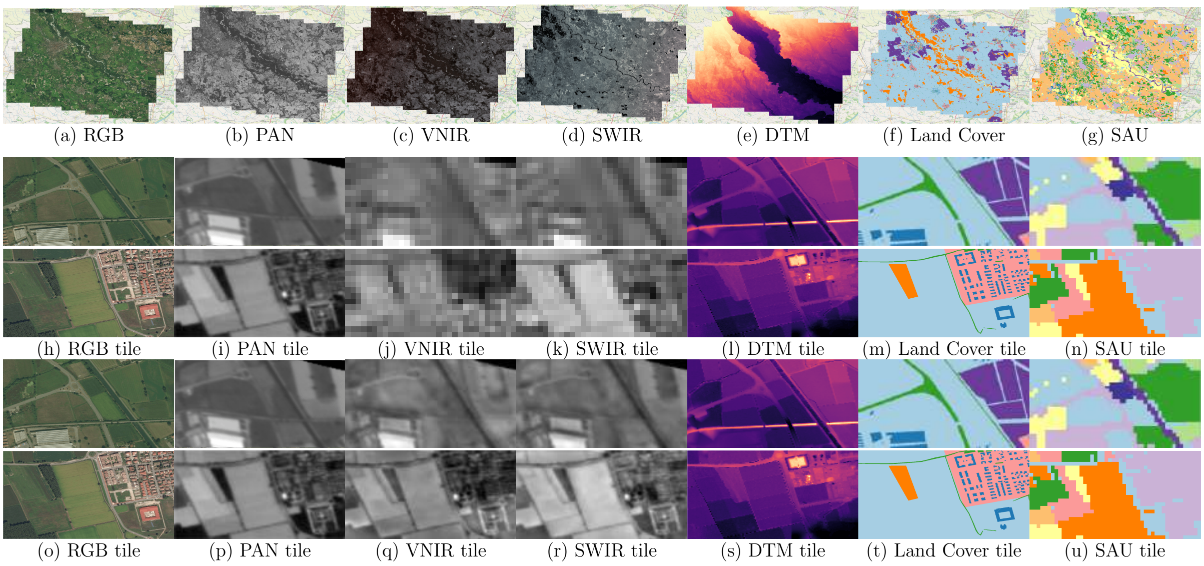


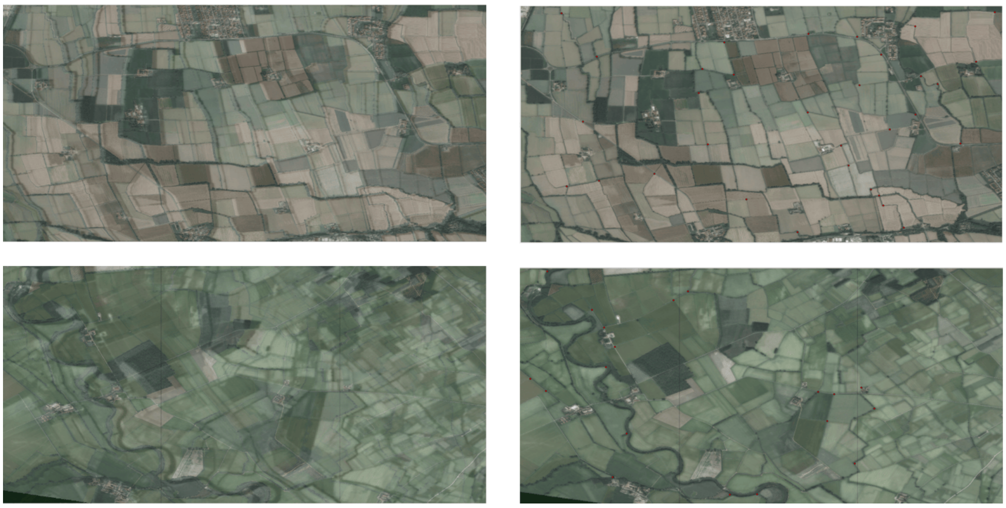
Figure 5.1: Visual representations of each modality and labeling of the entire Ticino dataset (from (a) to (g)) and two examples of tiles (one tile for each row) with the corresponding multimodalities (from (h) to (l)) and labelings ((m) and (n)).

and increased the spatial resolution of the hyperspectral data with the auxilium of the panchromatic information. The dataset has been split into training, validation, and test in percentages of 70%, 15%, and 15%, resulting in 1051 images for training, 225 for validation, and 226 for testing.

RGB data Figure 5.1(a) shows the RGB data included in the dataset. It has been collected from the Microsoft Bing Map service [206] through an open-source tool[2]. These images present a different horizontal and vertical resolution. Specifically, they have a spatial resolution of 1.86 m/px for the vertical dimension and 2.64 m/px for the horizontal one. The RGB source is the data with the highest spatial resolution in the dataset. Each RGB image tile has a dimension of about 256x362 pixels.

Panchromatic data Figure 5.1(b) shows the panchromatic (PAN) data collected from the ASI PRISMA satellite [6]. PAN is a grey-level image in the visible part of the spectrum (400-700 nm). It has the highest spatial resolution of the dataset, namely 5m/px. The original PRISMA and RGB data from Microsoft Bing presented a problem of georeference disalignment. The alignment of the two sources has been done with an interactive approach that involved the selection of more than 700 correspondent pairs of Ground Control Points between the RGB and PAN images, and the following estimation of a Thin Plate Spline Transformation for the geometric correction. The selection and the transformation were applied using QGIS Desktop software [4]. Figures 5.2(a) and (b) show the result of the alignment procedure. In the figures, two crops, considering RGB and panchromatic

[2] `https://github.com/dakshaau/map_tile_download`



(a) Disaligned (b) Aligned

Figure 5.2: Disalignment of RGB, Panchromatic, and hyperspectral data. The figure shows two RGB and panchromatic crops overlapped before (left) and after (right) the alignment operations.

modalities, are overlapped to show the difference between before (5.2(a)) and after (5.2(b)) the alignment. The final PAN tiles have a resolution of about 96x192 pixels.

Hyperspectral data Visual and Near-Infrared (VNIR) and Short-Wave Infrared (SWIR) cubes (figures 5.1(c) and (d)) present a resolution of 30m/px (the lowest spatial resolution of the dataset) and a spectral resolution of less than 12 nm. This data has been collected from ASI PRISMA satellite [6] with the level-2D pre-processing, which is the highest level distributed and solves most of the acquisition problems related to the atmosphere, co-registration, etc. The VNIR data includes the spectral information of the visible and near-infrared parts of the spectrum, from 400 to 1010 nm. The VNIR cubes present 63 bands out of the original 66, as three bands did not contain valuable information. The SWIR component of the dataset represents the information in the short wavelength infrared part of the spectrum, from 920 to 2500 nm, with a portion of the spectrum that overlaps the VNIR information. The SWIR cubes contain 173 bands, but even in this case, the last two have been discarded due to the absence of valuable information. For each sample in the dataset, the hyperspectral cubes are image tiles of around 16x32 pixels. The same alignment transformation applied on the PAN image has been applied to align the VNIR and SWIR data. Moreover, a second version is presented of the dataset where the hyperspectral cubes have been enhanced to reach the same spatial resolution of the PAN image using a pansharpening algorithm detailed in section 5.2. The resulting hyperspectral images are at a spatial resolution of about 96x192 pixels.

Digital Terrain Model data The last source included in the dataset is the Digital Terrain Model (DTM). As visible in Figure 5.1(e), the DTM represents a topographic

Table 5.2: Land Cover classes of the presented Ticino dataset and image cardinality per class.

Classes	**Id**	**# images**
Background	0	1497
Building	1	1242
Road	2	1326
Residential	3	555
Industrial	4	216
Forest	5	675
Farmland	6	443
Water	7	169

model of the bare Earth. It contains the elevation data of the terrain in a rectangular grid. It has been collected from the geoportal of the Lombardia region [207]. The DTM includes the urban and extra-urban areas. The model has been obtained from the geoportal by combining and harmonizing different sources of the data, removing possible anomalies, and finally extracting a Triangular Irregular Network model, achieving the final DTM model of the Lombardia region with a resolution of 5m/px [207]. The DTM used in this dataset presents image tiles of about 101x203 pixels and an elevation that ranges from 51.86 to 124.75 meters.

Land Cover Labeling Figure 5.1(f) shows the Land Cover segmentation. In the same way as the RGB data, the segmentation has different vertical and horizontal spatial resolutions, respectively equal to 0.68 and 0.96m/px. The final labeling, obtained after a refining and merging process described in section 5.2, consists of information from Open Street Map (OSM) [208], the Italian Agenzia delle Entrate [209], and manually added labeling. The final version includes 8 classes: *Background*, *Building*, *Road*, *Residential*, *Industrial*, *Forest*, *Farmland*, and *Water*. The *Background* class represents unlabeled pixels. Table 5.2 shows the image per-class cardinality along with the class name and identification number. The class distribution is slightly unbalanced, ranging from 169 images for the class *Water* to 1242 for the class *Building*. Moreover, Figure 5.3 offers a deeper analysis of the Land Cover labeling. The first row shows the number of pixels belonging to each class, while the second row the number of pixels per label for all the three sets in which the dataset has been divided dataset, namely the training (a), the validation (b), and the test (c) set.

Soil Agricultural Use Labeling (SAU) Figure 5.1(g) shows the SAU labeling. This segmentation has a resolution of 20m/px [210] and it has been collected from the Geoportal of Lombardia region [207]. The labeling, after the refinements described in section 5.2, includes 10 classes: *Background*, *Other agricultural crops*, *Forage crops*, *Corn*, *Industrial*

Table 5.3: Soil Agricultural Use classes of the presented Ticino dataset and image cardinality per class.

Classes	**Id**	**# images**
Background	0	1475
Other agricultural crops	1	380
Forage crops	2	918
Corn	3	1029
Industrial plants	4	669
Rice	5	1323
Seeds	6	177
Man-made areas	7	1175
Water bodies	8	337
Natural vegetation	9	1315

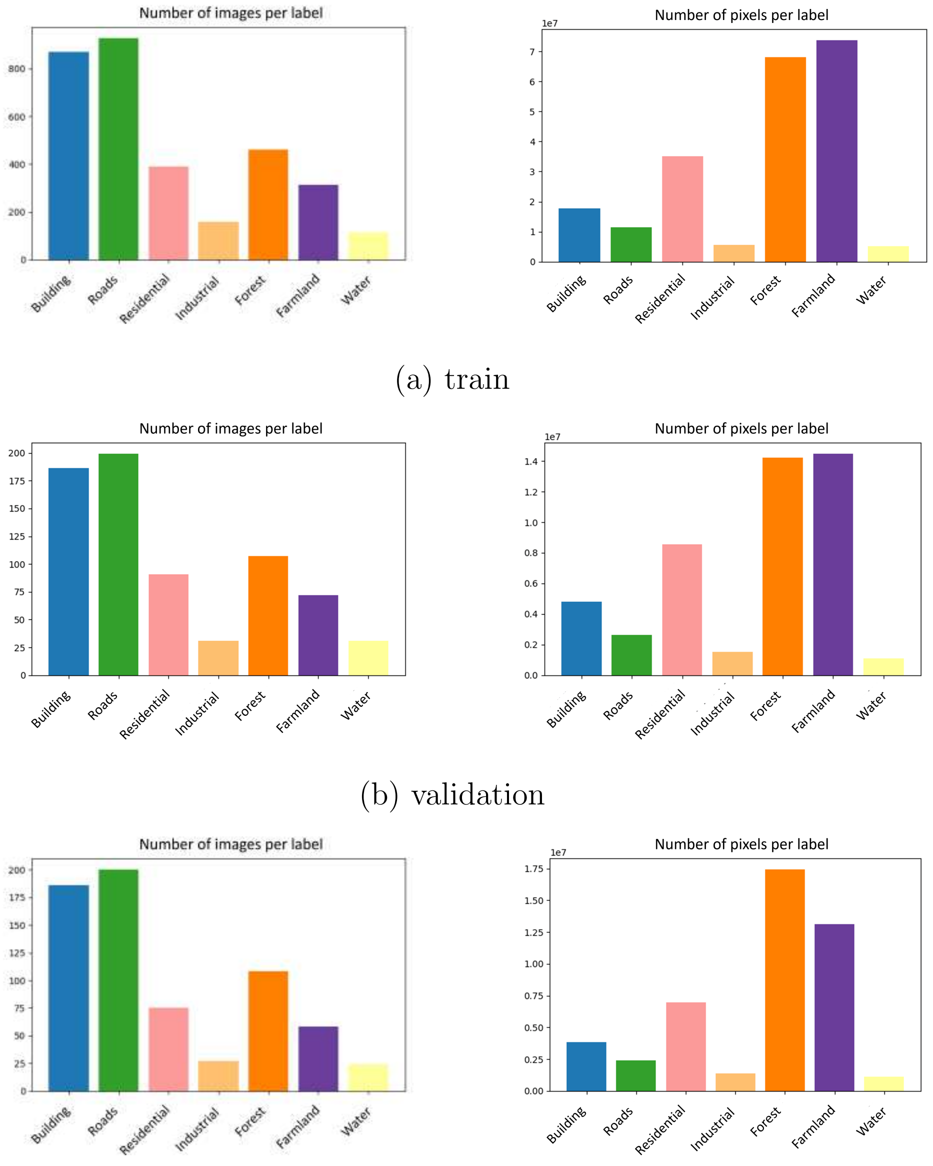
plants, *Rice*, *Seeds*, *Man-made areas*, *Water bodies*, and *Natural vegetation. Other agricultural crops* class indicates the not labeled farmlands and provides discrimination from the natural vegetation that instead describes forest, trees, and vegetation areas. Table 5.3 shows the image per-class cardinality of the Soil Agricultural Use, even in this case, along with the class name and identification number. As for Land Cover, the class distribution is slightly unbalanced ranging from 177 for the class *Seed* to 1323 for the class *Rice.* Finally, as before, a deeper analysis of the SAU distribution is proposed in Figure 5.4. The first row represents the number of pixels belonging to each class, while the second row the number of pixels per label for all three sets: training (a), the validation (b), and the test (c) set.

5.1.1 Refinement of the original labeling

This subsection will describe the refinement process of the original labelings to achieve the two final ground truths for semantic segmentation.

The final dataset has been collected by merging information from Open Street Map [208] and the Italian Agenzie delle Entrate [209], augmenting them with the creation of the *Water* labeling. As described in section 5.1, the dataset consists of 8 classes: *Background*, *Building*, *Road*, *Residential*, *Industrial*, *Forest*, *Farmland*, and *Water.*

Background, *Residential*, *Park*, *Industrial*, and *Forest* originally derived from the OSM labeling [208]. The original OSM segmentation includes 22 classes: *Background*, *Buildings*, *Forest*, *Residential*, *Farmland*, *Parking*, *Industrial*, *Stadium*, *Meadow*, *Pond*, *Park*, *Square*, *Harbour*, *Airport*, *Bridge*, *Beach*, *Industrial harbour*, *Baseball*, *Desert*, *Rock*, *Glacier*, and *River.* After having divided the area under investigation into 1808 tiles, the classes with low representations in terms of the number of image samples have been discarded: *Harbour*, *Airport*, *Bridge*, *Beach*, *Industrial harbour*, *Baseball*, *Desert*, *Rock*, *Glacier*, and *River.* As a consequence, 306 samples have been discarded because they mainly included the *Background* class.



(a) train

(b) validation

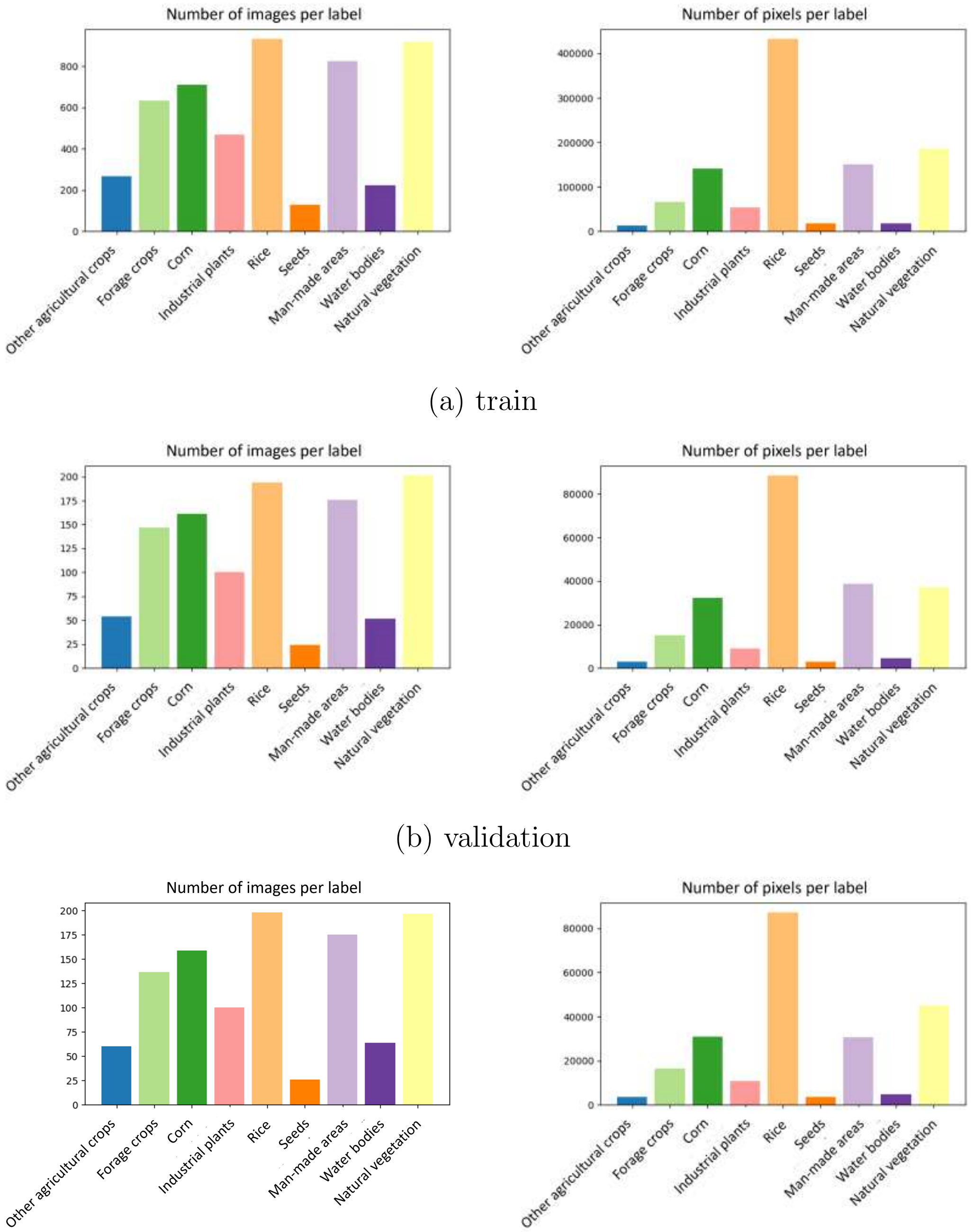
(c) test

Figure 5.3: Distribution of the Land Cover split of the dataset in training (left), validation (center), and test (right) sets. The first row represents the number of images per class (without Background). The second row represents the number of pixels per class (without Background).

Building and *Road* labelings have been collected by the Italian Agenzie delle Entrate [209]. The former has been inserted in the dataset as a substitute for the *Building* labeling of OSM because it is more accurate and complete in the area considered, while the latter was not present in the original OSM labeling.

Finally, *Water* is a combination of the *Pond* segmentation provided by OSM and a manual labeling provided by the authors of the Ticino River.

The original Soil Agricultural Use labeling has been acquired from the Geoportal of Lombardia region [207] and consisted of the following 22 classes: *Background*, *Other agricultural crops*, *Other cereals*, *Beet*, *Forests and tree crops*, *Nursery crops*, *Horticultural*



(a) train

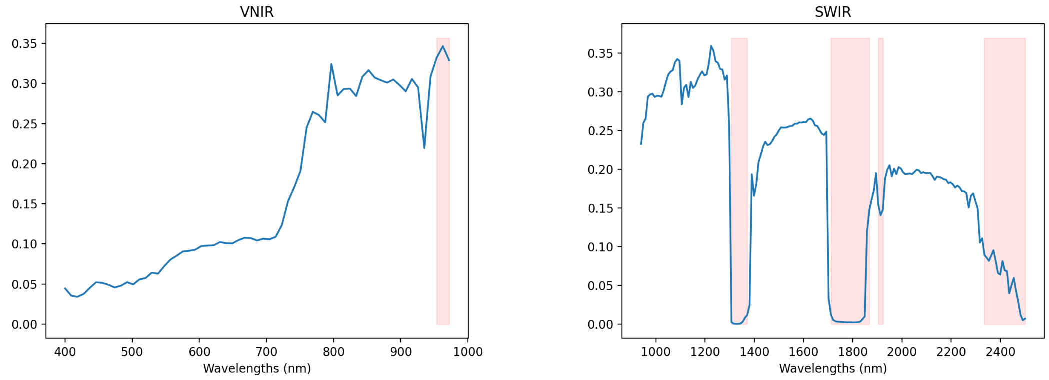
(b) validation

(c) test

Figure 5.4: Distribution of the Soil Agricultural Use split of the dataset in training (left), validation (center), and test (right) sets. The first row represents the number of images per class (without Background). The second row represents the number of pixels per class (without Background).

crops, *Forage crops*, *Fruit crops*, *Corn*, *Olive tree*, *Industrial plants*, *Rice*, *Seeds*, *Tainted and uncultivated*, *Fallow land*, *Vine*, *Man-made areas*, *Natural barren areas*, *Water bodies*, *Unclassifiable agricultural land*, and *Natural vegetation*.

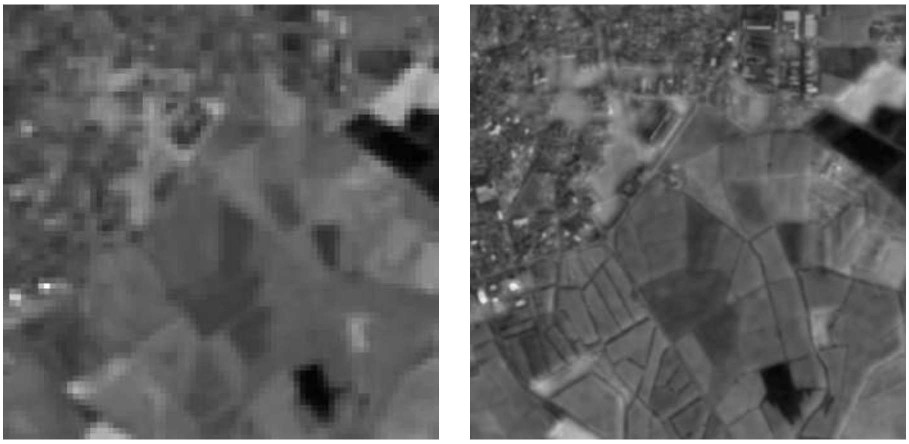
Other cereals, *Floriculture crops*, *Horticultural crops*, *Fruit crops*, *Vine*, *Beet*, and *Olive-tree* labels have been removed due to the low representation in the area considered. While *Forest and tree crops* and *Natural barren areas* have been respectively joined with



(a) VNIR

(b) SWIR

Figure 5.5: Visual representation of the cleaning of the corrupted bands (first step of the pre-processing). The mean signature on each band with the removed bands from VNIR (left) and SWIR (right) pointed out in red. The figures show that the removed bands correspond to the overlapping band between the two modalities and the water absorption wavelengths (where the signal is almost zeroed out).



(a) Original image

(b) Pansharpened image

Figure 5.6: Visual representation of the Pansharpening results (second step of the pre-processing). Comparison between the original hyperspectral image (band 50) on the left, and the hyperspectral pansharpened image (band 50) on the right after GSA algorithm.

the *Natural vegetation* and *Water bodies* as they have a similar semantic meaning. Finally, *Unclassifiable agricultural land*, *Tares and uncultivated*, and *Fallow land* were merged with the *Background* class because the semantic meaning was not clearly defined. The final labeling resulting from the cleaning process includes 10 classes as follows (see also section 5.1): *Background*, *Other agricultural crops*, *Forage crops*, *Corn*, *Industrial plants*, *Rice*, *Seeds*, *Man-made areas*, *Water bodies*, and *Natural vegetation*.

5.1.2 Data Pre-processing

As described in section 5.1, a pre-processing has been applied in order to remove corrupted bands from the HS component and to enhance its spatial resolution from 30m/px to 5m/px.

Before applying the pre-processing procedure, the HS bands with zero information have been eliminated, thus obtaining 63 bands for the VNIR and 171 bands for SWIR.

As described in 3.1.1, following Zini et al. [20], as a first step of the cleaning procedure of VNIR and SWIR sources, the corrupted bands have been identified using the information about invalid pixels from the PRISMA documentation [6, 205]. Each PRISMA image comes with correspondent information regarding the validity of each pixel in each band. A pixel is not valid if a problem occurs during the acquisition phase or the PRISMA pre-processing. For each band, the number of invalid pixels has been computed. Then, bands presenting a number of invalid pixels above a threshold, empirically fixed to 0.001%, are removed. The final part of VNIR is discarded. The removed bands from SWIR mainly correspond to the water absorption part of the spectrum where the information is almost zeroed out. As a second step, a visual inspection of each band led to the removal of the 39th band of the SWIR component due to the presence of visual artifacts. To better visualize the effect of the cleaning procedure, Figure 5.5 highlights in red the corrupted bands that were removed from the VNIR (Figure 5.5(a)) and SWIR (Figure 5.5(b)) signals. The resulting dataset used in the experiments for all the settings and configurations is:

- RGB: 3 channels;
- PAN: 1 channel;
- VNIR: 60 channels (cleaned from corrupted bands);
- SWIR: 122 channels (cleaned from corrupted bands and visual artifacts);
- DTM: 1 channel.

To take advantage of PRISMA data, a pansharpening operation has been used to improve the spatial resolution of VNIR and SWIR. PRISMA satellite provides a panchromatic image (PAN) and two hyperspectral cubes for VNIR and SWIR information captured at the same time. Following the results of Loncan at el. in 2015 [29] and Vivone et al. in 2022 on hyperspectral and PRISMA data pansharpening [133], the Gram-Schmidt Adaptive (GSA) [211] algorithm has been selected. The pansharpening has been applied on VNIR and SWIR concatenated in a single hyperspectral data. The final result (HS↑) is a hyperspectral cube corresponding to the fusion of the spectral information (VNIR and SWIR) and the spatial information from the PAN data. Figure 5.6 shows the 50th band of the hyperspectral signal in its original form (Figure 5.6(a)) and the same band after the pansharpening operation through the GSA algorithm (Figure 5.6(b)). The output has a spatial resolution of 5m/px (same as PAN) and a total of 182 bands that correspond to the VNIR and SWIR channels concatenated after the cleaning phase.

The final version of the dataset used in the experiments consisted of these modalities obtained by fusing PAN with VNIR and SWIR:

- RGB with 3 bands and a resolution of 1.86/2.64 meters per pixel;
- Hyperspectral with 182 bands and a resolution of 5 meters per pixel (HS↑);
- Digital Terrain Model (DTM) with 1 band and a resolution of 5 meters per pixel.

5.2 Methods

As mentioned above, two comparative analyses have been conducted. One, concerning CNN models focuses on providing a benchmark for future research and to demonstrate the superiority of multimodality. One, considering Transformer and different fusion methods to better understand which one is the best fusion strategy for the data considered. This section is divided into subsections that describe each analysis.

5.2.1 CNNs

In these experiments, different combinations and techniques of fusion have been considered for the modalities.

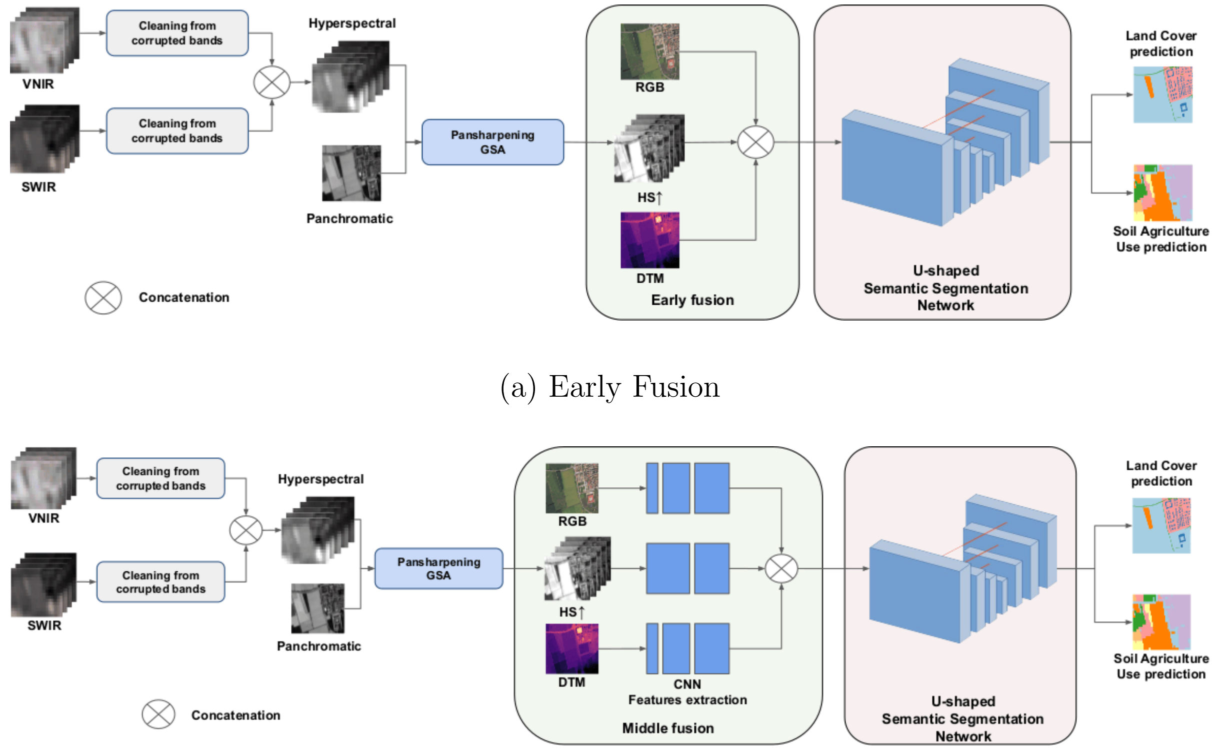
For each configuration, the same neural network model has been tested, consisting of a U-shaped architecture with a Residual Network of size 18 (Resnet18) backbone, using the Segmentation Models Pytorch framework [3].

For every test, the same settings of the learning rate, data augmentation, and normalization have been considered using the Albumentations library [212]. To train the models, a setup with 400 epochs, Adam optimizer with learning rate $1e-04$, and a StepLR scheduler with step_size of 30 and gamma of 0.85 has been used. The data augmentation for the training consisted of RGB normalization, HS↑ normalization, DTM normalization, Resize to 256 × 352, Random Crop of 256 × 256, Random Rotation between -180 and 180, Horizontal and Vertical Flip, and Transpose transformations. The validation and test data augmentation consider only RGB normalization, HS↑ normalization, DTM normalization, and image resize to 256 × 352.

Every modality has been normalized between 0 and 1 considering the max and min values of all the training dataset for each source independently. They have been also standardized with mean 0 and standard deviation 1, using the equation (5.1) [212] and always considering the mean and std of each modality and each channel of all the training set independently:

$$norm\_img = \frac{img - (mean \cdot max\_pixel\_value)}{std \cdot max\_pixel\_value} \tag{5.1}$$

[3] https://github.com/qubvel/segmentation_models.pytorch



(a) Early Fusion

(b) Middle Fusion

Figure 5.7: CNN experiments pipelines that represent the pre-processing, the fusion, and the segmentation. The two images show (a) the early and (b) the middle fusion techniques.

Figure 5.7 shows the complete procedures for both early and middle fusion. Both start with a pre-process to clean hyperspectral data from corrupted bands and improve their spatial resolution using the Gram-Schmidt Adaptive (GSA) pansharpening technique that fuses hyperspectral and panchromatic data.

Early fusion As shown in Figure 5.7(a), the pipeline for data-level fusion experiments consists of naively concatenating all the modalities together before using them as input of the U-shaped model.

The different combinations of modalities described above have 3 bands for RGB, 182 for HS↑, 185 for (RGB + HS↑), and 186 for (RGB + HS↑ + DTM). The definition of the U-shaped architecture and the layers of ResNet18 remained the same for all the experiments apart from the input layer which is changed according to the dimension of the input.

Middle fusion In the middle fusion approach, as shown in Figure 5.7(b), the different modalities are firstly processed independently to extract high-level features from each of them and later concatenated the features in order to create the input for the U-shaped architecture.

Table 5.4: Middle fusion Module for the extraction of features from each modality.

Modality	Layer	Description	Padding
RGB	Conv2d	$3 \times 3 \times 16$	2×2
	ReLU		
	Conv2d	$3 \times 3 \times 32$	2×2
	ReLU		
	Conv2d	$3 \times 3 \times 64$	2×2
	ReLU		
Hyperspectral (HS↑)	Conv2d	$3 \times 3 \times 128$	2×2
	ReLU		
	Conv2d	$3 \times 3 \times 64$	2×2
	ReLU		
DTM	Conv2d	$3 \times 3 \times 16$	2×2
	ReLU		
	Conv2d	$3 \times 3 \times 32$	2×2
	ReLU		
	Conv2d	$3 \times 3 \times 64$	2×2
	ReLU		

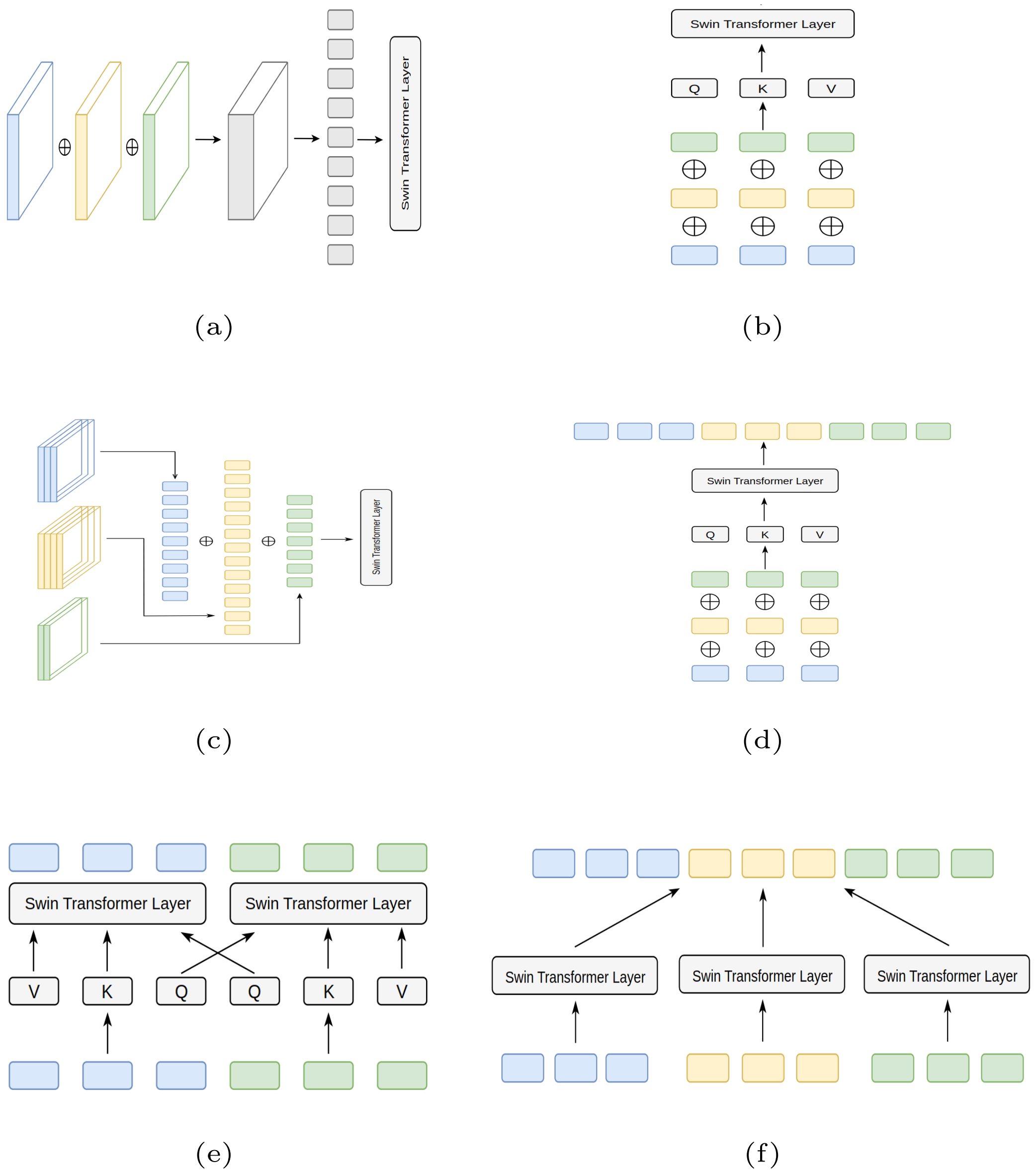
For each modality, the feature extraction module consists of convolutional and ReLU layers that use padding to maintain the same width and height of the U-shaped architecture. In the RGB and DTM cases, 3 convolutional layers increase the number of channels and extract the features. In the hyperspectral case, 2 convolutional layers are not only used to extract features but also to optimally reduce the number of channels of the starting hyperspectral inputs, thus overcoming the problem of the curse of dimensionality [10]. Table 5.4 summarizes the middle fusion module used for the extraction of the features in all modalities. After the application of the convolutional layers, all of the modalities share the same amount of feature maps to balance their importance during the training and are concatenated to become the input for the U-shaped architecture.

5.2.2 Transformers

Transformers present two critical challenges: 1) as CNNs they need large amounts of data, and 2) they are characterized by high computational complexity due to the quadratic nature of the self-attention mechanism that characterizes them. To address these concerns, the Shifted-Window Transformer (Swin) was introduced to specifically resolve issues related to computational complexity [120], while data-efficient transformers were proposed to mitigate the demands for extensive training data [213].

A comprehensive analysis of multimodal fusion methods for semantic segmentation of RS images based on the use of Swin-UperNet transformers [120] is presented with different fusion techniques that have been studied and adapted to the characteristics of the pansharpened version of Ticino. Therefore, all the methods were adapted to three modalities.

All the experimented fusion techniques are based on a U-shaped neural architecture



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.8: Fusion methods schemes considered in this work: (a) Early concatenation; (b) Token Patch Embedding; (c) Channel Patch Embedding; (d) Token Fusion at Attention Level; (e) Cross-Attention; and (f) Late Concatenation.

composed of an encoder and a decoder module. The encoder is a hierarchical shifted window-based vision transformer (Swin) [120], while the decoder is a UperNet with skip connections [214], which is a powerful semantic segmentation model known for its effectiveness in capturing intricate spatial relationships and high-level context. Six multimodal fusion techniques have been deployed and compared:

1. *Early Concatenation* (EC);

2. *Token Patch Embedding* (TPE);

3. *Channel Patch Embedding* (CPE);

4. *Token Fusion at Attention Level* (TFA);
5. *Cross-Attention* (CA);
6. *Late Concatenation* (LC).

A generalized schematic representation of these fusion methods can be seen in Figure 5.8. They are categorized into three classes based on where the fusion occurs: early fusion at the input level Figure 5.8(a), middle fusion at an intermediate point within the encoder Figure 5.8(b)-(e) and late fusion after the encoder's processing Figure 5.8(f). The investigated methodologies can vary considerably in terms of complexity, performance capabilities and computation requirements. For both *Early Concatenation* and *Late Concatenation* methods, modifications to the Swin-UperNet have been required to accommodate all three modalities presented in the dataset used in the experiments. In the case of the middle fusion methods, suitable strategies for integrating these three modalities had to be devised, drawing inspiration from prior research in multimodal fusion. The investigated methodologies have different benefits and can vary considerably in terms of complexity, performance capabilities and computation requirements.

Swin-based encoder The encoder is based on the canonical Swin transfomer architecture [120], consisting of 4 stages $\{S_i\}_{i=1}^{4}$. Each Stage, apart from the first one, is characterized by a Patch Merging module and a Swin Transformer Block (STB). Each Block includes at least a pair of consecutive Window Multi-head Self Attention (W-MSA) and Shifted Window Multi-head Self Attention (SW-MSA) modules. The first stage S_1 consists of a Linear Embedding layer and an STB. At the beginning, the image is divided into N patches $\{p_i\}_{i=1}^{N}$ that are then introduced into the first Stage S_1. Here, each patch p_i is projected by the Linear Embedding ($embedding()$) layer into a token z_i. All tokens $Z = \{z_i\}_{i=1}^{N}$ enter into the STB and consequently in the self-attention modules that extract the new tokens and give them to the stage S_2. Each stage, from the second to the last, starts with the Patch Merging module that reduces the number of patches grouping them 2 by 2 and then giving them to the STB. Given the U-shape of the encoder-decoder model, intermediate representations produced after each stage of the Swin encoder are subsequently fed into the symmetric UperNet decoder using skip connections.

Fusion techniques Let's consider the case of fusing three modalities $\{X_i\}_{i=1,2,3}$ (RGB, HS and DTM in this analysis). Z_i denotes the respective set of token embeddings of the modality X_i and Z the input of the STB derived by the previous operations.

Early fusion. The simplest fusion strategy is the *Early Concatenation* (EC), where the images from multiple modalities are concatenated ($concat()$) at input level on channel dimension and then processed by one Swin-based Encoder:

$$X_{(1,2,3)} = concat(X_1, X_2, X_3)$$

$$Z = embedding(X_{(1,2,3)}).$$

Middle fusion. A middle fusion solution is the *Token Patch Embedding* (TPE) concatenation in which the token embedding sequences from multiple modalities are concatenated and fed into the Swin Transformer layers of the first STB [136]:

$$Z_i = embedding(X_i) \; with \; i = 1, 2, 3$$

$$Z = concat(Z_1, Z_2, Z_3).$$

The idea here is that with multimodal data, all the positions of tokens from different modalities can be treated as a single sequence. By doing so, the context of one modality can be effectively used to encode the positions of tokens from other modalities. However, it's important to note that this approach can lead to longer sequences after concatenation, which in turn increases the computational complexity.

Another middle fusion method is the *Channel Patch Embedding* (CPE) [215], which involves generating individual token embeddings for each channel within every modality. These embeddings are then concatenated and fed as input to the first STB. For example for hyperspectral data, this would correspond to the individual spectral bands, while for RGB data, to the different color channels. Formally:

$$Z_{i,j} = embedding(X_{i,j})$$

where i is for the modality and j for the channel of the modality. Then, for $i = 1, 2, 3$ and each channel j:

$$Z = concat(Z_{i,j}).$$

The *Token Fusion at Attention Level* (TFA) method involves processing the three modalities separately within three distinct Swin Transformer encoders, alternating one and three streams throughout the process. It has been designed by us as a variant of the *Token Patch Embedding* where the concatenation is done at the token level at each stage. Before computing W-MSA and before SW-MSA in each transformer block, the tokens generated by the three modalities up to that point are concatenated, allowing for joint attention computation. After attention computation, the outputs are divided ($split()$) and processed individually by the three encoders until the next attention module. In this particular case, let's also consider Y_i^l as the tokens of the i-th modality at stage l (in the first stage it will be equal to Z_i) and Y^l as the input of the Transformer Block in S_l. For each stage l the operations are as follows:

$$Y^l = concat(Y_1^l, Y_2^l, Y_3^l)$$

$$Y_{1,w}^l, Y_{2,w}^l, Y_{3,w}^l = split(WMSA(Y^l))$$
$$Y_w^l = concat(Y_{1,w}^l, Y_{2,w}^l, Y_{3,w}^l)$$
$$Y_1^{l+1}, Y_2^{l+1}, Y_3^{l+1} = split(SWMSA(Y_w^l)).$$

These operations are computed for every self-attention operation in every STB. Every Y^l is fed into the decoder through skip connections.

Cross Attention (CA) is a method used in two-stream Transformers [140], to facilitate cross-modal interactions by exchanging query embeddings between modalities. In this case, the third modality has been leveraged following the idea outlined by Dufter et al. [216], and utilizing it as positional embedding. Considering Q_i, K_i, V_i, the query, key and value of the canonical self-attention technique for the i-th modality and MSA as self-attention operator valid for both W-MSA and SW-MSA, the cross attention between only X_1 and X_2 is computed as:

$$\begin{cases} M_1 = MSA(Q_2,K_1,V_1) \\ M_2 = MSA(Q_1,K_2,V_2) \end{cases}$$

where M_1 and M_2 are the token outputs for the first stream of modality 1 and the second stream of modality 2. Cross-attention allows for cross-modal interactions, highlighting the importance of considering self-attention within each modality for a more comprehensive understanding.

Late fusion. *Late Concatenation* (LC) works in a multi-stream mode. It involves processing the three modalities separately in three distinct Swin Transformer encoders. The output of each stage is then concatenated on the channel dimension and into the UperNet decoder. Formally, let's consider the output at each stage l for each i-th modality as O_i^l:

$$O^j = concat(O_1^l, O_2^l, O_3^l).$$

Each O^j is then used in the skip connection with the correspondent layer of the UperNet decoder.

All fusion methods were implemented using three modalities: RGB, HS↑, and DTM except for Cross-Attention in which RGB and HS↑ were employed as main modalities and DTM as positional embedding.

Due to the high computational requirements of Swin-UperNet models and to adapt HS information to some of the fusion methodologies, for these experiments, a Principal Component Analysis (PCA) has been applied to the pansharpened image. The PCA also helped to address the curse of dimensionality problem, typical of HS data, extracting spectrally homogeneous regions. The first four principal components were retained, accounting for 99% of the variation and resulting in a revised HS* with four spectral bands.

Before training, a data augmentation strategy based on the Albumentations library [212] was also applied, similarly to the CNNs experiments. It includes random cropping, rotation, and horizontal and vertical flipping to images resized to 256x256 pixels and normalized. All models employed a Swin encoder configuration with a patch size of 4, a window size of 7, and a depth specified as 2, 2, 6, 2 along with attention heads set to 3, 6, 12, 24 and expansion layer. Due to limitations in computational resources, the embedding dimension for each Swin Transformer was adjusted accordingly: 96 for Early Concatenation, Token Patch Embedding and Cross-Attention, 48 for Token Fusion at the Attention Level and Late Concatenation, and 24 for the Channel Patch Embedding method. All models were subject to a stochastic depth regularization of 0.3. For training, Adam optimizer was employed and trained for 250 epochs with an initial learning rate of 10^{-3} and weight decay 10^{-4}. A learning rate scheduler was also applied to reduce it. The cross-entropy loss was used for training. All experiments were run on NVIDIA GTX 1070 GPU with 8GB of RAM.

5.3 Experiments and results

This section will be divided into two parts, the first dedicated to the CNNs experiments and second one to the Transformers.

5.3.1 CNN results

The compared configurations are:

- single modality: RGB or HS↑;
- multimodalities: (RGB + HS↑) or (RGB + HS↑ + DTM).

The two fusion techniques that have been tested for each of the possible combinations are:

- early fusion;
- middle fusion.

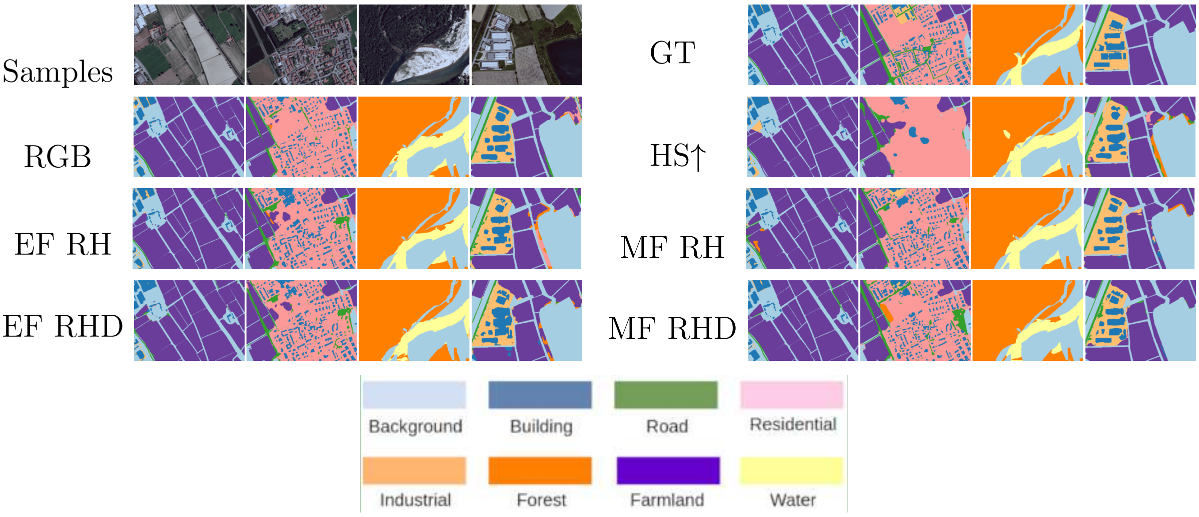


Figure 5.9: Visual prediction of Land Cover segmentation for all the approaches. EF and MF are respectively for Early and Middle Fusion, while RH and RHD are respectively for the combinations (RGB + HS↑) and (RGB + HS↑ + DTM).

The evaluation and comparison of the experiments are based on Accuracy (Acc), mean Intersection over Union (mIoU), and Precision. All of them have been computed by considering the average performance on single classes. For the evaluation, the torchmetrics library has been used, setting the operation to purposely ignore the background. The algorithms have been modified from the original procedure due to an error that has been found during the work on this thesis. In particular, when the parameter *ignore_index* was used in a multiclass metric, it zeros out the metric on the class to ignore but it still included the value in the average computation, counting the ignore index in the mean computation and leading to wrong results[4]. A comprehensive evaluation of single classes is also reported, always based on the same metrics.

Land Cover Evaluation Table 5.5 displays the overall and class-specific results for Land Cover achieved across different configurations. Combining all modalities using middle fusion yields the best performance across evaluation metrics. The results highlight how performance depends on both the modalities involved and the chosen fusion method.

An in-depth analysis shows that HS-only is worse than RGB-only modality. This behavior is probably due to the fact that the HS cube is at a lower resolution than RGB images, thus causing a loss of finer details in the segmentation process.

Early fusion approaches achieve comparable results to RGB alone but fail to outperform it. Conversely, middle fusion outperforms the RGB-only architecture in both experiments, demonstrating the usefulness of HS and DTM information in the Land Cover scenario and the importance of datasets that allow multimodal approaches. In particular, both middle fusion setups (RGB+HS) and (RGB+HS↑+DTM), which are very similar in terms

[4] `https://github.com/Lightning-AI/torchmetrics/issues/1692`

Table 5.5: Land Cover overall and single classes results of every experiment configuration with CNNs, divided by modalities combination and fusion techniques. Bold values represent the best performance obtained on the rows.

		Land Cover					
		No fusion		Early fusion		Middle fusion	
Class	Metric	RGB	HS↑	(RGB + HS↑)	(RGB + HS↑ + DTM)	(RGB + HS↑)	(RGB + HS↑ + DTM)
Building	Acc	0.62	0.39	0.64	0.63	**0.75**	0.69
	IoU	0.50	0.31	0.49	0.48	**0.54**	0.53
	Precision	**0.71**	0.60	0.68	0.67	0.66	0.69
Road	Acc	0.52	0.29	0.45	0.41	**0.57**	0.55
	IoU	**0.42**	0.23	0.34	0.33	0.41	0.41
	Precision	**0.69**	0.52	0.58	0.62	0.59	0.61
Residential	Acc	0.85	**0.87**	0.82	0.85	0.75	0.80
	IoU	**0.64**	0.57	0.62	0.58	0.63	**0.64**
	Precision	0.72	0.62	0.72	0.64	**0.79**	0.76
Industrial	Acc	0.64	0.52	0.62	0.47	**0.72**	0.67
	IoU	0.50	0.40	0.47	0.41	**0.55**	0.54
	Precision	0.70	0.64	0.65	**0.75**	0.70	0.74
Forest	Acc	0.92	0.90	0.92	0.89	0.95	**0.96**
	IoU	0.87	0.85	0.88	0.86	0.90	**0.92**
	Precision	0.94	0.93	0.95	0.96	0.95	**0.96**
Farmland	Acc	0.93	0.91	0.93	**0.95**	0.93	**0.95**
	IoU	0.85	0.82	0.86	0.88	0.87	**0.90**
	Precision	0.91	0.89	0.91	0.92	0.94	**0.95**
Water	Acc	0.79	0.86	0.87	0.85	**0.89**	0.88
	IoU	0.65	0.72	**0.74**	0.73	**0.74**	0.73
	Precision	0.79	0.82	0.83	**0.85**	0.81	0.81
Overall	Acc	0.75	0.68	0.75	0.72	**0.79**	0.78
	IoU	0.63	0.56	0.63	0.61	0.66	**0.67**
	Precision	0.78	0.72	0.76	0.77	0.78	**0.79**

of performance, outperform RGB-only by about 4%, 4%, 1% in terms of Accuracy, mIoU, and Precision respectively. Middle fusion setups outperform also HS-only by about 11%, 11%, and 5% in terms of Accuracy, mIoU, and Precision respectively.

The performance of each class is also analyzed. Focusing on mIoU metrics, only *Residential* and *Road* classes perform better using only RGB. These classes, especially Road, benefit from high spatial resolution information. For the Road class, spatial resolution degrades with HS-only and early fusion approaches. In HS-only, RGB cannot compensate for the loss of spatial information, yielding the worst results. In early fusion methods, RGB partially compensates for the missing information, but the shallow fusion methodology prevents exploiting the best characteristics of each source. The mIoU performance in middle fusion approaches is perfectly comparable to the RGB-only setting, supporting the hypothesis. Moreover, middle fusion achieves the best accuracy for the *Road* class, as the mixed features aid in discrimination. For *Residential* labeling, spatial resolution is less fundamental than for *Road* due to less fine-grained labeling, resulting in comparable performances across all configurations. Middle fusion of all modalities achieves the same mIoU evaluation as RGB-only. Other labelings benefit from involving

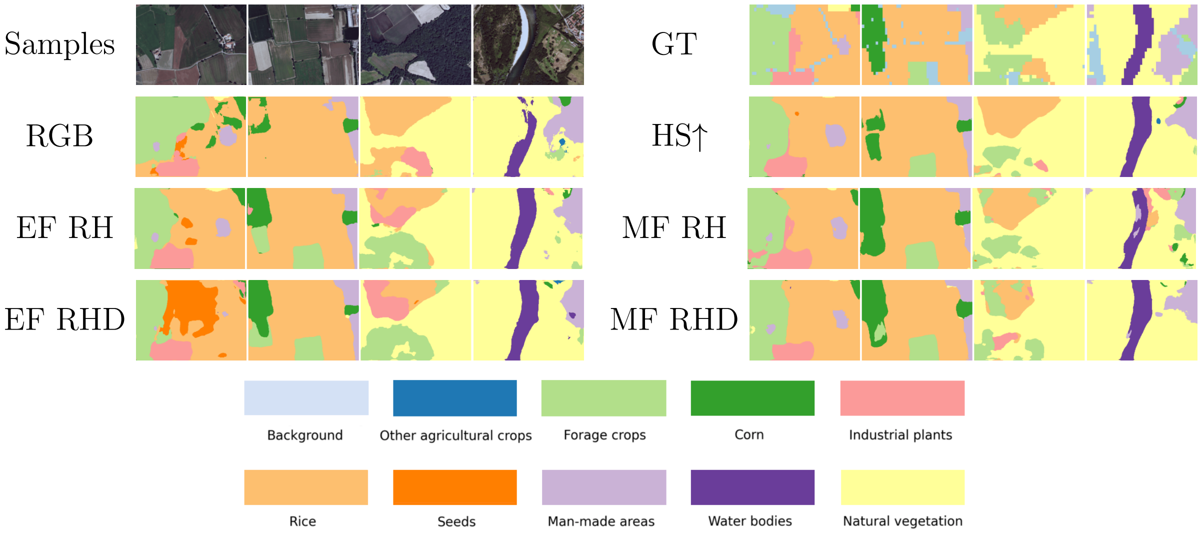


Figure 5.10: Visual prediction of Soil Agricultural Use segmentation for all the approaches. EF and MF are respectively for Early and Middle Fusion, while RH and RHD are respectively for the combinations (RGB + HS↑) and (RGB + HS↑ + DTM).

HS and DTM data. The improvement is evident in middle fusion, where the best mIoU scores are obtained. Comparing middle fusion configurations with and without DTM, results are comparable. However, it is notable that *Forest* and *Farmland* perform better for all metrics when DTM is involved, indicating the usefulness of terrain elevation in discriminating between these classes.

Finally, visual results for all of the approaches are shown in figure 5.9. Focusing on the best overall model (middle fusion with all modalities), it accurately classifies all labels, from fine-grained *Road* and *Building* to coarse-grained *Residential*, *Farmland*, and *Industrial*. Notably, the model recognizes a forest area located in the RGB image but not in the labeling, demonstrating good performance even with noisy labels.

To summarize, shallowly combining RGB, HS, and DTM data for Land Cover semantic segmentation is insufficient to surpass the usage of RGB images alone. However, as demonstrated in middle fusion, a proper combination of multimodalities yields better performance than all other configurations, emphasizing once again the importance of selecting the correct fusion methodology. Furthermore, specific classes benefit more from specific modalities in combination with others. Therefore, to further improve performance in semantic segmentation, modalities, and their combinations should be selected based on fusion methods and the class being discriminated.

Soil Agricultural Use Evaluation Table 5.6 presents the overall results for Soil Agricultural Use, highlighting the poor performance of the RGB-only approach and the beneficial impact of HS modality in class discrimination. Due to the lower resolution of SAU labeling, more focus is dedicated to the Accuracy rather than the mIoU. Early

Table 5.6: Soil Agricultural Use overall and single classes results of every experiment configuration with CNNs, divided by modalities combination and fusion techniques. Bold values represent the best performance obtained on the rows.

Soil Agricultural Use							
		No fusion		Early fusion		Middle fusion	
Class	Metric	RGB	HS↑	(RGB + HS↑)	(RGB + HS↑ + DTM)	(RGB + HS↑)	(RGB + HS↑ + DTM)
Other agricultural crops	Acc	0.26	0.31	0.34	0.36	0.42	**0.46**
	IoU	0.17	0.26	0.27	0.26	0.29	**0.32**
	Precision	0.34	0.63	0.55	0.48	0.50	**0.52**
Forage crops	Acc	0.14	0.32	0.32	**0.40**	0.35	0.37
	IoU	0.11	0.24	0.23	0.26	0.24	**0.27**
	Precision	0.34	0.52	0.45	0.42	0.44	**0.48**
Corn	Acc	**0.51**	0.45	0.48	0.47	0.49	**0.51**
	IoU	0.31	0.31	0.32	0.33	0.34	**0.36**
	Precision	0.45	0.50	0.49	**0.54**	0.52	**0.54**
Industrial plants	Acc	0.17	0.31	0.34	0.31	**0.46**	0.38
	IoU	0.09	0.18	0.19	0.19	**0.27**	0.21
	Precision	0.16	0.29	0.30	0.33	**0.39**	0.33
Rice	Acc	0.74	**0.81**	0.78	0.80	**0.81**	0.80
	IoU	0.57	0.64	0.63	0.65	**0.68**	**0.68**
	Precision	0.72	0.75	0.77	0.77	**0.81**	**0.81**
Seeds	Acc	0.01	0.06	0.12	0.17	0.21	**0.25**
	IoU	0.00	0.04	0.07	0.10	0.16	**0.20**
	Precision	0.02	0.14	0.12	0.20	0.39	**0.53**
Man-made areas	Acc	0.89	0.89	0.89	0.89	**0.90**	**0.90**
	IoU	0.77	0.76	**0.78**	0.76	0.77	0.77
	Precision	0.85	0.83	**0.86**	0.83	0.84	0.84
Water bodies	Acc	0.56	0.72	0.70	**0.75**	0.66	0.69
	IoU	0.46	0.55	**0.56**	0.57	0.55	**0.56**
	Precision	0.72	0.69	0.74	0.70	**0.77**	0.75
Natural vegetation	Acc	0.82	0.83	0.83	0.80	0.84	**0.85**
	IoU	0.64	**0.67**	**0.67**	0.65	**0.67**	**0.67**
	Precision	0.75	**0.78**	0.77	**0.78**	0.77	0.76
Overall	Acc	0.47	0.54	0.55	0.57	0.59	**0.60**
	IoU	0.35	0.41	0.41	0.42	0.44	**0.45**
	Precision	0.50	0.59	0.58	0.58	0.62	**0.63**

fusion approaches achieve comparable results to the HS-only experiment. The early fusion approach already demonstrated the advantages of using multimodality w.r.t. RGB-only (with an increment of 10%) and HS-only (3%). DTM also contributes to segmentation, yielding improvements in accuracy and mIoU. As observed for Land Cover, the choice of fusion methodology is crucial. Middle fusion approaches showcase the true advantages of a multimodal approach, outperforming RGB-only and HS-only experiments, with the best results obtained by combining all modalities. The difference in performance between RGB-only and middle fusion using HS and DTM is significant, with an increment of about 13%, 10%, and 13% for Accuracy, mIoU, and Precision, respectively. The improvement gained by using middle fusion with HS and DTM w.r.t. HS-only modality is about of 6%, 4%, and 5% for Accuracy, mIoU, and Precision, respectively.

Table 5.6 also reports the segmentation results for each class. All methods outperform the RGB-only approach in class discrimination. *Man-made areas*, *Natural vegetation*, and *Corn* show comparable performance with RGB-only, but slight improvements are observed with the involvement of other modalities. The *Seeds* class, in particular, demonstrates

Table 5.7: Land Cover overall results for each method, dividing single modality and multimodality fusion methods using Transformers. The bold and underlined values represent the best and the second-best performance achieved for each metric, respectively.

	Method	Acc	Pr	mIoU	Macs	Pars
Single	RGB	67.22	72.75	55.71	9.65	39.28
	HS*	58.10	62.71	45.51	9.65	39.28
Multi	Early Conc. (EC)	67.57	73.15	56.06	9.68	39.29
	Tok. Pat. Emb. (TPE)	68.89	64.71	73.95	16.40	60.60
	Cha. Pat. Emb. (CPE)	65.01	71.18	53.85	65.43	241.96
	Tok. Fus. Att. (TFA)	69.13	74.27	57.51	16.14	38.74
	Cross-Att. (CA)	**71.85**	74.72	59.42	37.86	111.61
	Late Conc. (LC)	71.84	**75.31**	**59.69**	16.14	63.29

significant improvements when other modalities are utilized, going from 1% accuracy with RGB-only to 25% accuracy with all modalities and the middle fusion approach. Fusion methodology also plays a crucial role, with middle fusion generally yielding better improvements over early fusion.

Visual results for each combination and fusion technique are reported in figure 5.10. The segmentations achieved by the best approach, with middle fusion and all the modalities involved, accurately identify all classes despite the low resolution of SAU labeling.

This investigation demonstrates the usefulness of a multimodal approach, especially for Soil Agricultural Use segmentation. Hyperspectral data and the Digital Terrain Model prove to be even more beneficial in this context than in Land Cover labeling, where RGB alone fails in achieving satisfactory results. Consequently, the availability of comprehensive multimodal datasets is crucial for future research.

5.3.2 Transformer results

This section presents the outcomes of the experiments, which have been assessed and examined using three evaluation metrics averaged on classes: Accuracy (Acc), Precision (Pr) and mean Intersection over Union (mIoU). The computational complexity of each method using the number (Million) of parameters of the neural model (Pars) and the number (Giga) of multiply–accumulate operations (Macs) have also been measured. Table 5.7 reports the results achieved by every tested method, comparing both single modality and multimodality approaches[5]. Figure 5.11 shows examples of visual results from each method. As expected, due to the higher spatial resolution and as demonstrated by the previous experiments, RGB mode performs better among single-mode approaches, achieving superior performance to HS*.

Comparing multimodal and single modality approaches, it is possible to note that,

[5] The results in the table differs from the one in the original paper [22] because the adjustment of the metrics from torchmetrics (described in section 5.3.1) was not applied in the article.

apart from *Channel Patch Embedding*, all the multimodal methods outperform RGB alone. Among these methods, *Cross-Attention* and *Late Concatenation* achieve the best results. They are both comparable, nonetheless, the former is the best on Acc, while the latter reaches better results on Pr and mIoU. Nevertheless, when taking into account Pars and Macs in the analysis, it becomes evident that the *Late Concatenation* method exhibits a significantly lower number of Pars compared to the *Cross-Attention* approach. In contrast, the *Cross-Attention* method substantially increases the complexity of the RGB network by about twice. The same argument is valid for the Macs where the *late Concatenation* has less than half Macs than *Cross-Attention*. Therefore, *Late Concatenation* is considered the best method, outperforming RGB of 4.04%, 2.24% and 3.47% on Acc, Pr and mIoU, respectively. Figure 5.12 shows a comparison of all methods (excluding HS*). Ideally, the best method is the one in the upper right part of the plot with a small circle that indicates the number of Pars, confirming the conclusion that *Late Concatenation* is the method that overall performs better. It is worth noting that *Token Fusion at Attention Level* represents an excellent trade-off between performance and resources used, since it is superior to RGB and, at the same time, has fewer Pars with comparable complexity in terms of Mac (equal to *Late Concatenation*).

To summarize, excluding *Channel Patch Embedding*, five of six multimodal approaches outperformed RGB and consequently any other single modality approach. In particular, compared with RGB, two methods distinguished themselves. The *Token Fusion at Attention Level* revealed to be the best compromise in terms of performance (outperforming RGB) and memory (parameters). The *Late Concatenation* method proved to be the best multimodal method. Even with Transformers, the results demonstrate that a multimodal approach is more efficient in terms of performance while keeping the consumption of resources comparable with single modality methods.

5.4 Usefulness of multimodality in remote sensing semantic segmentation

In this work, the Ticino dataset has been presented, a novel multimodal dataset for RS semantic segmentation, that is crucial in various applications, including environment management and precision farming. The use of multimodal sources of information enhances the segmentation performance and class discrimination, thus the scarcity of existing multimodal datasets poses challenges in RS semantic segmentation. Existing datasets have low cardinality or lack spectral information, limiting the effectiveness of data-hungry deep-learning techniques that require diverse samples for training.

The proposed dataset presents five modalities: RGB, panchromatic, VNIR, SWIR, and DTM and two labelings: the Land Cover with eight classes and the Soil Agricultural

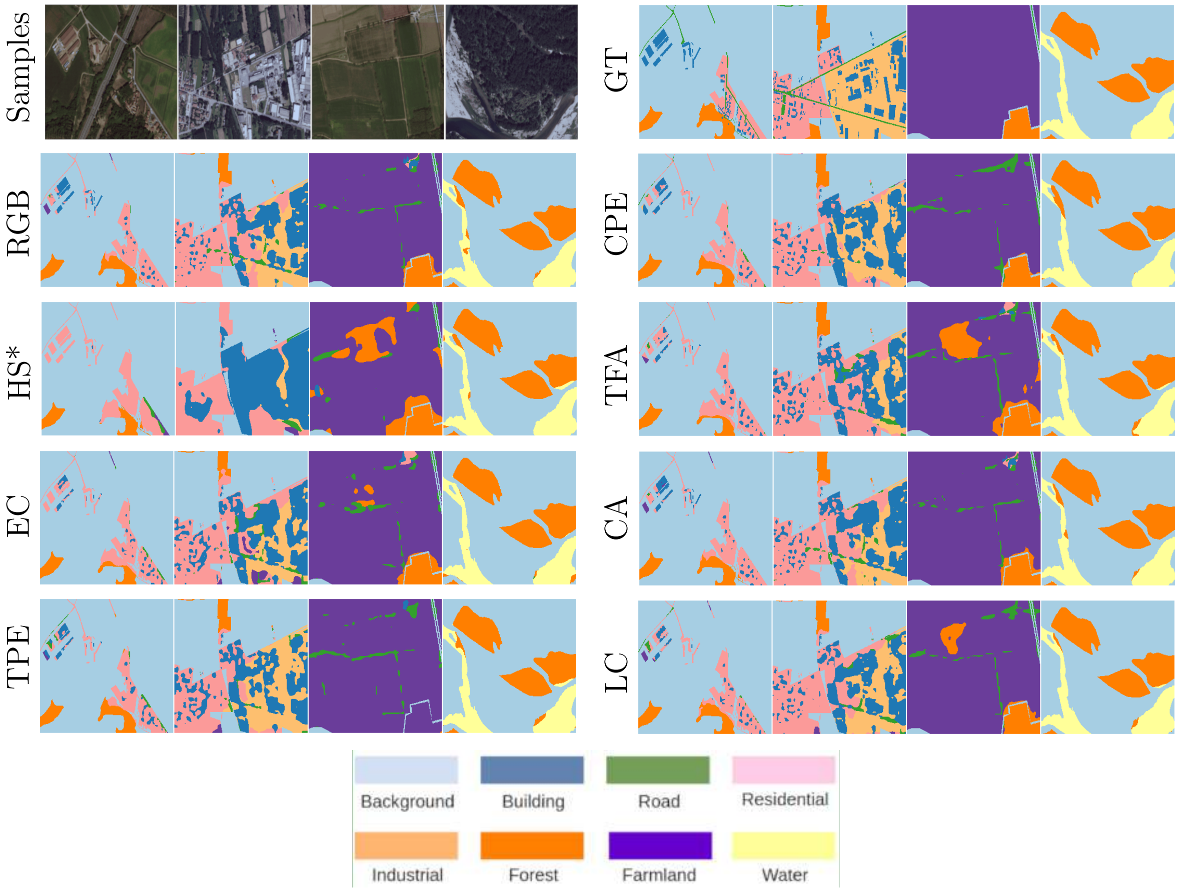


Figure 5.11: RGB samples at top-left row; Ground truth (GT) at top-right row; single modality results: RGB, HS; multimodality results: EC, TPE, CPE, TFA, CA and LC.*

Use with 10 classes. This dataset is the biggest and most diverse dataset for RS semantic segmentation as it includes a high cardinality of images for all the modalities. Specifically, the Ticino dataset provides 1502 tiles and an extension of around 1332 km^2.

Furthermore, the advantages of these modalities have been investigated in two sets of experiments. The first set of experiments was based on CNN models. On the first hand, the scope of this analysis was to understand if the combination of complementary modalities can outperform the use of single RGB modality, and, on the second hand, to provide a baseline for multimodal RS semantic segmentation on the proposed dataset. The results, based on early and middle fusion approaches, show that both Land Cover and especially SAU labelings can benefit from the multimodal approach, with the best setting represented by the combination of all the modalities and middle fusion. The second set of experiments, instead, was based on Transformers, with the aim of properly studying the effect and improvements achieved using different fusion methods. In this case, six fusion methods were tested: *Early Concatenation* (EC), *Token Patch Embedding* (TPE), *Channel Patch Embedding* (CPE), *Token Fusion at Attention Level* (TFA), *Cross-Attention* (CA), and *Late Concatenation* (LC). The results showed once again that multimodality outperforms the use of single modality approaches in terms of performance, without overcomplicating the

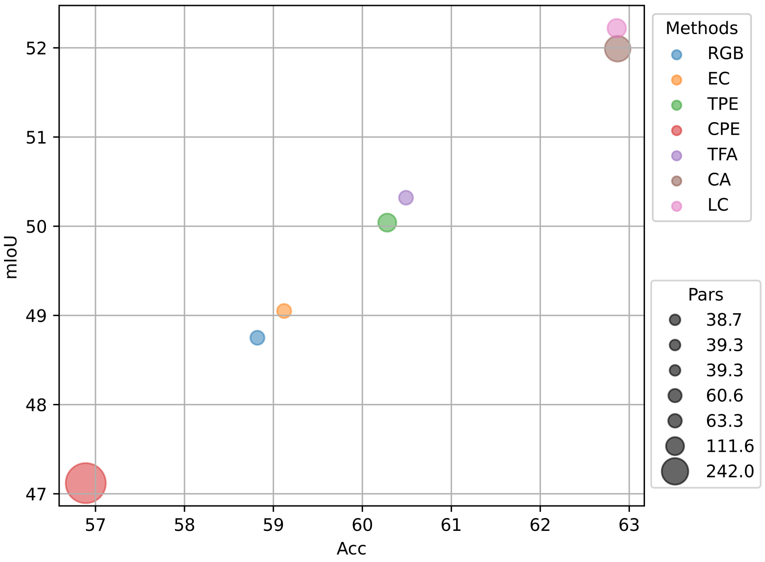


Figure 5.12: Comparison of the performance of the fusion methods based on Acc (x), mIoU (y) and Parameters (area of the circles).

resources needed. Among the tested methods, the *Late Concatenation* demonstrated better performance thanks to the ability to extract high-level features from each heterogenous source independently, and then combine them. The *Token Fusion at Attention Level* showed the best compromise between performance and complexity.

Plenty of challenges connected to semantic segmentation are still open, and this dataset can become the first step in the right direction. This dataset can also help investigate open issues such as hyperspectral pansharpening, dimensionality reduction of high cardinality data, and spatio-temporal fusion of modalities.

As a future work, a further refinement of the labeling has been planned, reducing the noisy labels and balancing the low-represented classes by using the dataset itself for a semi-supervised labelization of the background. The extension of the dataset and the increment of its variability have also been planned. Moreover, this data allows to continue investigating the field of RS semantic segmentation to further exploit the usefulness of the HS and the DTM. This can be achieved by studying specific techniques of fusion that take into consideration the difference between each modality (e.g. time stamp, resolution, etc.) and their relation with the semantic classes.

Chapter 6

Digital Soil Mapping

The investigation described in this chapter has been published in *MDPI Sensors* journal and it refers to the paper *Estimation of Soil Characteristics from Multispectral Sentinel-3 Imagery and DEM Derivatives Using Machine Learning* [23].

Monitoring soil properties is a fundamental aspect of precision agriculture, offering improved resource management [217], enhanced risk assessment and effective land erosion monitoring [218]. Furthermore, soil has the potential for carbon sequestration, which could prove to be a formidable tool in combating climate change in the years ahead [219].

The primary method for the characterization of soil involves manually collecting soil samples, drying them and subsequently performing chemical analyses in a laboratory setting [155]. However, the manual collection of soil samples, along with their corresponding physicochemical characterization, is a time-consuming process that lacks scalability for extensive areas [220]. Different soil properties interact with electromagnetic radiation in diverse ways. As electromagnetic waves strike the Earth's surface, they can be absorbed, transmitted or reflected. The reflection and absorption patterns at different wavelengths provide insights into the composition, structure and properties of the observed materials [221]. More recently, hyperspectral and multispectral soil characterization has emerged as a highly valuable tool for the estimation of soil properties without the need for chemical analyses of the soil samples [222, 223]. Multispectral and hyperspectral remote sensing harness data from multiple narrow and contiguous bands across the electromagnetic spectrum, with each band corresponding to a specific wavelength range. This technology has proven to be immensely valuable for soil characterization due to its ability to detect and analyze various soil properties [224, 221].

Among the aforementioned research papers cited in section 2.4, the work authored by Zhou et al. [146] stands out as one of the most pertinent contributions. This study presents a comprehensive comparison of diverse satellite sensors (including Landsat-8, Sentinel-2 and Sentinel-3), each coupled with varying spatial and temporal resolutions, in an effort to predict the organic carbon content and C:N ratio in Switzerland. The outcomes of this analysis show that the prediction models based on Landsat-8 and Sentinel-2 yielded the most favorable and least favorable results, respectively, in terms of error in the organic carbon estimations. It is worth noting, however, that this investigation also highlighted the potential inherent in models based on Sentinel-3 data. Despite its coarser resolution in comparison to Landsat-8 and Sentinel-2 (300 m versus 30 m and 10–60 m), models utilizing Sentinel-3 exhibited competitive or even superior accuracy. Remarkably, Sentinel-3's advantage lies in its broader spectral coverage, offering 21 bands as opposed to the 7 of Landsat-8 and the 13 of Sentinel-2. This expanded spectral range holds the promise of enhancing the estimations of soil parameters. Moreover, it is important to emphasize that Sentinel-3, while characterized by reduced spatial resolution, represents a relatively novel sensor that remains largely unexplored for machine learning-based soil parameter estimation, as highlighted by Odebiri et al. [225]. One limitation of this study [146] is the extent of the area under investigation, which is limited to the Swiss territory. In fact, the amount of data used in the experiments is limited, making it difficult to use data-hungry methodologies such as machine learning. The size of the area under investigation is very important in terms of demonstrating the generalization capabilities of soil estimation methods on regions not considered during the training phase. In fact, soil properties vary significantly across different regions due to their unique physicochemical properties, resulting from factors such as climate, topography and time [226]. Moreover, terrain features, including slope, aspect and elevation, along with environmental elements like water availability and vegetation, play a crucial role in influencing spectral transmission, which is essential in this application [227].

Starting from these aspects, this study aims to show the effectiveness of soil parameter estimation models over a larger geographical area than Switzerland.

In more detail, in this chapter, different machine learning methodologies are evaluated for the estimation of the multiple soil characteristics of a continent-wide area corresponding to the European region using multispectral Sentinel-3 satellite imagery [5] and DEM [228] derivatives. The soil characteristics' ground truth is obtained from the LUCAS library, which is the largest collection of physicochemical soil properties and corresponding spectral reflections acquired in the laboratory. The LUCAS library includes about 20,000 samples taken from specific geographical locations across the entire European region [155]. With each geographical location of the LUCAS dataset, the corresponding Sentinel-3 multispectral signature and DEM derivative are associated, thus obtaining a large remote sensing dataset to be used for the estimation of multiple soil characteristics. The study area includes the entire European region, comprising an extensive collection of soil samples with remarkable diversity and heterogeneity. The analysis presented in this work provides insights into the potential of machine learning techniques to generalize over a vast geographical area. Nevertheless, given the substantial variations in soil properties across different regions, as mentioned earlier, the validity of these findings for areas beyond Europe should be empirically verified.

The main contributions of this study are the following:

- a multisource remote sensing dataset of the European region by merging multispectral images from Sentinel-3 and DEM derivatives from the European Copernicus mission and the corresponding LUCAS samples;
- a benchmark of several machine learning methods for the estimation of the soil characteristics using multispectral signals, DEM derivatives and a combination of them;
- methods based on an artificial neural network (ANN) capable of predicting all the soil characteristics at the same time;
- an analysis of the importance of each input source (multispectral and DEM) in predicting the soil properties.

6.1 AI Methods for Digital Soil Mapping

Machine learning has revolutionized various scientific disciplines by enabling computers to learn patterns from data and make precise predictions. Several methodologies have been presented in the state of the art, such as ANNs, gradient boosting (GB), random forest (RF), support vector regressor (SVR), etc. Among these methods, ANNs have emerged as a powerful class of data-driven algorithms inspired by the biological neural networks of the human brain. These methods have gained significant popularity due to their ability to deal with complex and high-dimensional data, making them well-suited for a wide range of applications, such as classification and semantic segmentation. This study leverages the potential of ANNs in combination with multispectral Sentinel-3 satellite imagery and DEM derivatives to estimate multiple soil characteristics over a continent-wide area corresponding to the European region. ANNs optimize their performance through a data-driven training procedure that minimizes a loss function by employing interconnected nodes and weighted connections to learn from the data. The use of data-driven ANNs in this research provides an efficient method for the improvement of the precision of soil property estimation, and their application in conjunction with multispectral features and DEM derivatives demonstrates their crucial role in advancing soil characterization using machine learning techniques.

A machine learning method for the estimation of a soil parameter s learns a mapping function $f(\mathbf{u};\Theta)$, with parameters Θ, between a given input $\mathbf{u}$ and a target soil parameter s. The input can be either a multispectral signature $\mathbf{m}$ alone or a combination of multispectral and DEM features $\mathbf{d}$. This method can be extended for the estimation of multiple soil parameters $\mathbf{s} = \{s_1, \ldots, s_P\}$, with P being the number of parameters to be estimated. At the inference time, given the input $\mathbf{u}$, the mapping function $f(\mathbf{u}; \Theta)$ outputs the estimation $\hat{s}$ of a given soil property. In the case of multi-variable estimation, the mapping function outputs the estimation of all the variables at the same time $\hat{\mathbf{s}} = \{\hat{s}_1, \cdots, \hat{s}_P\}$. The goodness of the predictions can be evaluated by comparing s with $\hat{s}$ in the case of single-variable estimation, and by comparing $\mathbf{s}$ with $\hat{\mathbf{s}}$ in the case of multiple-variable estimation. The metrics used for the evaluation are the coefficient of determination R^2, root mean square error (RMSE) and mean absolute error (MAE).

This study considers the following four state-of-the-art methods [146]: ANNs, GB, RF and support vector regressor (SVR). For all these methods, we evaluated the use of two different inputs, namely (i) multispectral information ($\mathbf{m}$); and (ii) a combination of multispectral information ($\mathbf{m}$) and features extracted from the DEM, which we refer to as DEM derivatives ($\mathbf{m}$,$\mathbf{d}$). While for the GB, RF, and SVR, the canonical implementations available from the Python Scikit-Learn library (`https://scikit-learn.org/stable/`, accessed on 12 September 2023) are adopted, in the case of the ANNs, a network architecture for the soil parameter estimation task has been specially engineered. GB, SVR, and RF follow the same configuration used by Zhou et al. [146]. For all of them, the parameters were optimized using the grid search algorithm.

The GB algorithm uses regression trees and gradient optimization as a procedure for the minimization of the loss function. The algorithm consists of training a set of regression trees in sequence. At each step, the residual error of the previous tree is used as a label for the current tree. After the prediction of all the trees, gradient optimization is used to change the weights of each tree, minimizing the loss function. The loss function considered in the experiments is the mean squared error (MSE) (Equation (6.1)).

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2 \tag{6.1}$$

where x_i and y_i are the corresponding pair of the input and output, and N is the number of samples.

SVR, instead, is a machine-learning algorithm that allows for a definition of a tolerance on the accepted error of the model. Based on a kernel function, it allows for the identification of the best hyperplane to fit the data. The kernel function used in the experiments is the radial basis function (RBF) (Equation (6.2)), due to its performance in soil mapping [229].

$$k(x_i, x_j) = exp\Big(-\sigma \|x_i + x_j\|^2\Big) \tag{6.2}$$

where k is the kernel function, x_i and x_j are the input vectors, and σ represents the width

Table 6.1: Structure of the neural network used for the soil variable estimation. N represents 21 if only multispectral features are used and 29 if geomorphological features are added. M is the number of outputs of the neural network, which can be 1 for the single-variable estimation or 12 for the multi-variable estimation.

Stage	Operation	Output size
Pre-processing	Input	N
Encoding	Linear + Hardswish	32
	Linear + Tanhshrink	128
	Linear + Hardswish	32
	Linear	M
	Total parameters	$32N + 8,192 + 32M$

of the RBF, thus regulating the relationship between input and output.

RF uses multiple decision trees to predict the output. Each tree makes its own prediction, and the final values are obtained by merging them together. RF also uses a bootstrap technique to generalize the training and avoid overfitting [230]. Basically, every time a prediction is made during the training, a set of N samples from the dataset is chosen and used for the training step. The remaining samples are used to evaluate the value of the loss function, which is then used during the optimization of the model. In this case also, the loss function chosen is the MSE (Equation (6.1)). The number of trees and the number of samples for the bootstrap procedure were optimized with a grid search algorithm, as was the case for the other tested methods.

The ANNs were implemented using the PyTorch Library (`https://pytorch.org/`, accessed on 12 September 2023). The architecture and hyperparameters of the model were optimized using a grid search method. The optimized architecture of the ANNs is described in Table 6.1. It is composed of five linear layers, including input and output layers. The three hidden layers are separated in turn by Tahnhshrink and Hardswish (h-swish) activation functions [231], which were selected during the optimization of the architecture. The input and output layers of the ANNs were parameterized so that they can be adapted to different types of inputs (($\mathbf{m}$) and ($\mathbf{m}$,$\mathbf{d}$)) and outputs. In fact, the ability of the ANN to predict multiple soil properties at the same time is also investigated, thus having two kinds of outputs: single and multi. To distinguish between the ANN for single-variable estimation and the ANN for multiple-variable estimation, ANN Single and ANN Multi are used as labels. The loss function used to train the ANN models is the MSE (Equation (6.1)). Furthermore, the hyperparameter optimization selected the Adam optimizer with a learning rate of 10^{-2} as the best-performing configuration.

6.1.1 Explainability Investigation

Data-driven methods automatically learn the importance of an input feature in predicting a given output. The more closely correlated the input feature is with the output, the more frequently it will be considered during the learning process. However, this correlation cannot be observed directly and requires an appropriate strategy that depends on the machine learning method adopted for the prediction [232].

The analysis of the feature importance is focused on the RF method, with the aim of understanding which bands of the multispectral signal and which DEM derivatives contribute most significantly to a given soil variable estimation. Due to their remarkable effectiveness and interpretability, RFs have acquired great popularity in the literature. This unique combination establishes them as a potent tool not only for accurate forecasting but also for the facilitation of a comprehensive explanation of outcomes.

Two strategies exist in the state of the art: the mean decrease in impurity (MDI) and the feature permutation (FP) [233, 151]. The former counts the number of times a feature is used to split a tree node, weighted by the number of samples it splits. In the version used in this investigation, the decrease in node impurity is also weighted by the probability of reaching that node. The latter, instead, measures the increase in the prediction error of the model after the values of the features are permuted, breaking the relationship between the features and the true outcome. The impurity-based feature importance cannot scale up well on high-cardinality input features. Therefore, the FP technique has been adopted.

6.2 Multimodal dataset for the estimation of soil parameters

One of the contributions of this work is the creation of a large library of soil properties, spaceborne multispectral data and digital elevation data corresponding to N georeferenced land points. The entire library is defined as $\mathcal{L} = (\mathcal{S}, \mathcal{M}, \mathcal{D})$, with $\mathcal{S} = \{\mathbf{s}_i\}_{n=1}^{N}$, $\mathcal{M} = \{\mathbf{m}\}_{n=1}^{N}$ and $\mathcal{D} = \{\mathbf{d}\}_{n=1}^{N}$ being the soil properties, multispectral and digital elevation model information, respectively.

For each land point n, $\mathbf{s}_n$, $\mathbf{m}_n$ and $\mathbf{d}_n$ are defined as follows:

- $\mathbf{s}_n = \{s_1^n, \cdots, s_P^n\}$ are the $P = 12$ soil properties;
- $\mathbf{m}_n = \{b_1^n, \cdots, b_L^n\}$ are the $L = 21$ bands of the spaceborne multispectral data;
- $\mathbf{d}_n = \{d_1^n, \cdots, d_K^n\}$ are the $K = 8$ features extracted from the digital elevation model information.

The following section will present, first, how the multispectral data are collected and processed, and secondly, how the soil properties have been selected and how the association between the properties and the spectra is created. Finally, how the digital elevation data have been collected and how the features to be associated with the soil properties have been extracted will be described.

6.2.1 Sentinel-3 - Multispectral Data

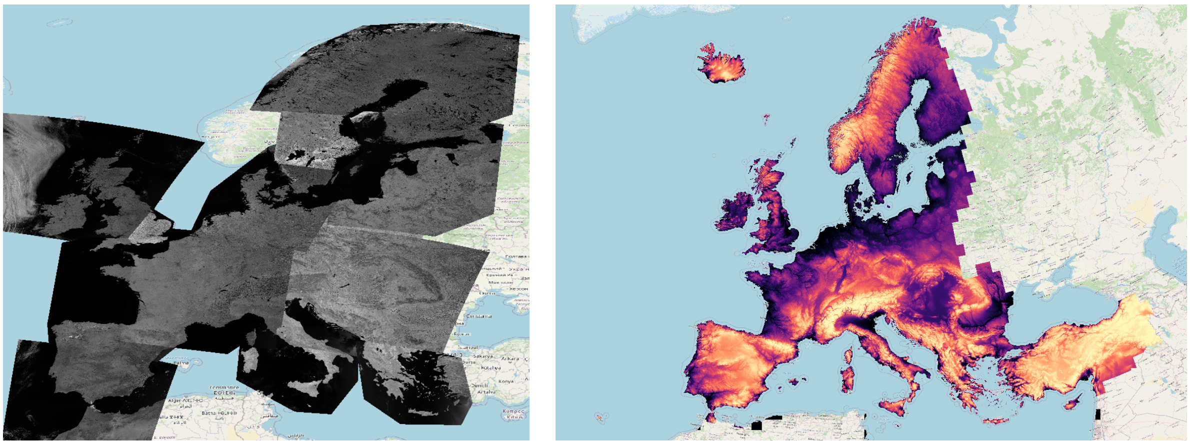
The data has been gathered from the Sentinel-3 satellite mission in order to cover the entire European continent. Sentinel-3 is a multi-instrument mission to measure the sea surface topography, sea and land surface temperature and ocean and land color with a high level of accuracy and reliability. The mission is composed of two satellite platforms: Sentinel-3A, launched on 16 February 2016, and Sentinel-3B, launched on 25 April 2018.

Both satellites are equipped with several different sensing instruments and are collocated in the low Earth orbit. Among all the instruments, the Ocean and Land Colour Instrument (OLCI) is the most interesting for the purpose of this work. OLCI has a spatial resolution of 300 m and is capable of measuring 21 spectral bands, from 400 nm to 1020 nm. It is capable of producing images with a swathe of 1270 km, and is not centered at the nadir, but is tilted 12.6° westwards to mitigate the negative impact of sun glint contamination. OLCI images are processed and distributed through the Copernicus Open Access Hub [234].

For this work, images preprocessed with level 1 were used, which includes top-of-atmosphere (TOA) radiometric measurements, radiometrically corrected, calibrated and spectrally characterized. These images are also quality-controlled and georeferenced (latitude, longitude and altitude). The images cover an area of 1200 km^2 and have a resolution of 300 m. All the images are encapsulated in a Network Common Data Form 4 (NetCDF 4) format [235] and processed through QGIS Desktop (version 3.20.2) software [4].

Images of level 1 were downloaded from the Copernicus Open Access Hub, using the Semi-Automatic Classification Plugin for QGIS (version 7.8.35) [236]. This tool allowed the downloading of images from the hub and computing preprocessing operations specific for Sentinel-3, creating 21 images corresponding to the 21 spectral bands. Atmospheric correction DOS1 (dark object subtraction) was applied [237]. This methodology is one of the most common techniques adopted for such purposes: water, forests and shadows are considered dark objects when their values of reflectance are close to zero. Dark objects are detected automatically when the pixel reflectance value is less than or equal to 1.0%. The assumption is that some pixels within the image receive 0% of solar radiation and the values of radiance corresponding to these pixels registered by the satellite correspond to atmospheric dispersion.

A total of 14 images were downloaded, spanning five years, from 2016 to 2021, with an acquisition time between the summer period, from May to September. Similarly to the practice adopted by Zhou et al. [146], all images were chosen with a cloud coverage inferior to 10% of the acquisition. The images were loaded into the QGIS software as single raster layers, with the same datum, EPSG:4326 World Geodetic System 1984, considered for each. Figure 6.1a shows all the multispectral images collected by Sentinel-3 and properly merged to cover the entire European continent. It should be noted that, for the sake of visualization, all the downloaded images are visualized by showing their spectral average,



(a)

(b)

*Figure 6.1: Representation of the data used in these experiments. (**a**) shows multispectral images collected by Sentinel-3 and properly merged to cover the entire European continent. For the sake of visualization, (**a**) represents the average over all the bands, normalized between 0 and 1. (**b**) represents the digital elevation model acquired from the Copernicus Land Monitoring Service. Greenish colours represent low-elevation values (approx. −214 m), while reddish colours represent high-elevation values (approx. 5105 m).*

and thus, the resulting patchwork effect is only a visual artifact.

6.2.2 LUCAS - Soil Data

To connect soil properties with each multispectral signature, the target variables considered in the LUCAS library are included. LUCAS is a programme carried out by EUROSTAT (the European Statistical Office) that aims to organize harmonized surveys across all the states of the European Union over time [155]. The LUCAS library includes a total of approximately 20,000 samples, each of 0.5 kg of topsoil material. The topsoil sampling locations were selected to be representative of the European landscape features. The selection was based on a stratified random sampling that took into consideration the CORINE land cover 2000, the the Shuttle Radar Topography Mission (SRTM) DEM and derived slope, aspect and curvature [238]. The authors of the dataset decided to exclude areas above 1000 m from the survey due to the challenges associated with accessing and sampling these high-altitude locations. Finally, the LUCAS topsoil sample points exhibit a density of approximately 1 per 199 km^2, which theoretically permits a grid cell size of approximately 14 km [238]. All the dried samples were analyzed for the percentage of coarse fragments, particle size distribution (% clay, silt, and sand content), pH (in $CaCl_2$ and H_2O), organic carbon (g/kg), carbonate content-CaCO$_3$ (g/kg), phosphorous content (mg/kg), total nitrogen content (g/kg), extractable potassium content (mg/kg), CEC (cmol(+)/kg) and hyperspectral reflections, measured in a laboratory environment. A great portion of the data, namely 43% of all samples, was collected from croplands. Figure 6.2 shows violin plot distributions of all the 12 soil properties. Figure 6.2 (d), (e) are in the logarithmic scale. The association of each LUCAS sample to the spaceborne

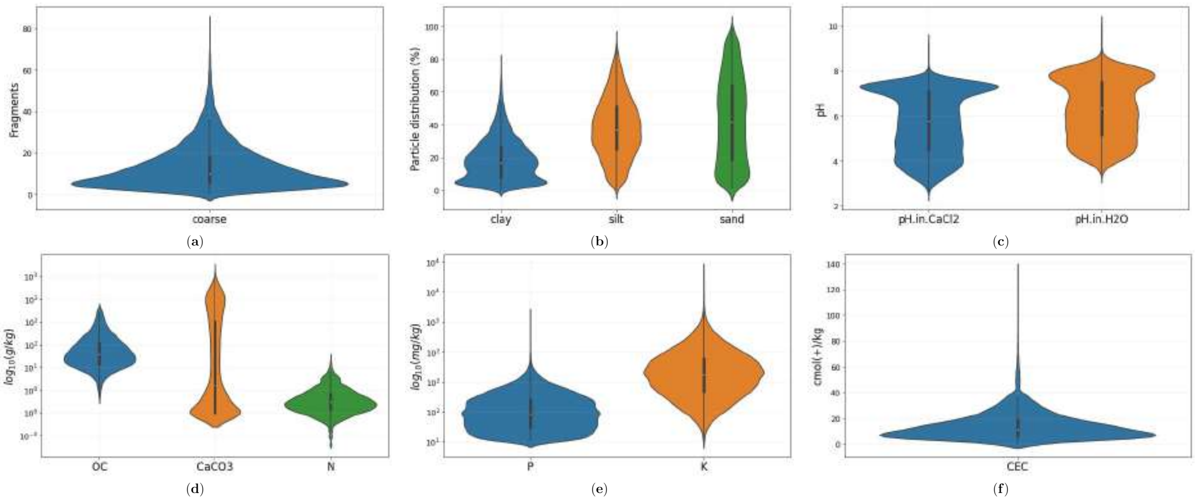


Figure 6.2: Violin plots of the soil properties considered in this work. (a) Coarse; (b) Clay, Silt and Sand; (c) pH in $CaCl_2$ and H_2O; (d) Organic Carbon (OC), Calcium Carbonate (CaCO3) and Nitrogen (N); (e) Phosphorous (P) and Potassium (K); (f) Cation exchange capacity (CEC). For the sake of visualization, the variables P, K and CEC are shown in a logarithmic scale.

multispectral signatures of Sentinel-3 has been achieved through the GPS coordinates associated with each sample.

6.2.3 Copernicus - Digital Elevation Model Data

To take into account the geometrical distortions of the land in the estimation of the soil properties, the DEM has been included, which is a representation of the terrain elevation [239]. The DEM was acquired from the Copernicus Land Monitoring Service, resampled to a resolution of 8 m, and then saved as a raster layer on QGIS. Moreover, the DEM was reprojected from the EPSG:4326-WGS 84 datum, in degrees, to another datum (EPSG:3035 ETRS89-extended/LAEA Europe), which is in meters. Figure 6.1b shows the maps obtained, which were later sampled with the points from the LUCAS dataset.

Afterward, the raster layers corresponding to the DEM information were processed using the SAGA GIS software (version 7.8.2) [240] to extract the following features: altitude, valley depth, slope, topographic wetness index (TWI), channel network base level (CNBL), vertical distance to channel network (VDCN), catchment slope and slope length.

The TWI is used to estimate where water is accumulated, the CNBL is the base level of groundwater and the VDCN is the vertical distance. The exact procedure for the feature extraction is described as follows. The *valley depth* function is called with the following parameters: *Tension Threshold* 1, *Maximum Iterations* 0, *Keep Ridge Level Above Surface* checked, *Ridge Detection Threshold* 4. The functions *slope* [241] (*Unit* radians) and *slope length* are. Then, the functions to calculate the *SAGA Wetness Index*, catchment area and catchment slope [242] are called with the following parameters: *Suction* 10, *Type of Area* square root of the catchment area, *Type of Slope* catchment slope, *Minimum slope* 0, *Offset*

Table 6.2: Variables considered by the state-of-the-art methods related to the estimation of soil parameters from multispectral signals. Most of these methods consider the SOC as the most important variable. This work goes further and provides all the necessary tools to estimate almost all soil characteristics simultaneously.

method	coarse	clay	silt	sand	pH_{CaCl2}	pH_{H2O}	SOC	CaCO3	N	P	K	CEC	Area
Meng et al. [150]							X						North east China (315 samples)
Forkuor et al. [144]		X	X	X			X		X			X	Rural watershed (580 km^2)
Safanelli et al. [145]		X		X		X	X	X				X	European Croplands (7142 samples)
Trontelj et al. [153]									X	X	X		Slovenia (350 samples)
Li et al. [154]		X					X		X				19 sampling sites (180 samples)
Zhou et al. [146]							X		X				Switzerland (150 samples)
Our Proposal	X	X	X	X	X	X	X	X	X	X	X	X	**Europe (20,000 samples)**

Slope 0.1, *Slope Weighting* 1. The channel network was created through the catchment area as an initiation grid, setting the initiation type as "Greater than" and the "threshold" as 10 million, using the *channel network* function with the following parameters: *Min. Segment Length* 10. Finally, the function *vertical distance to channel network* was used to extract the VDCN and CNBL with the following parameters: *Tension Threshold* 1, *Maximum Iterations* 0, *Keep Base Level below Surface* checked.

6.2.4 Comparison with Existing Datasets

Table 6.2 recapitulates the main characteristics of the state-of-the-art soil parameter estimation experiments in comparison with the ANN methods proposed. All of these estimate a single or small subset of soil characteristics. Of great importance is the constraint on the volume of data employed in the experiments, posing challenges for the utilization of data-hungry approaches like machine learning. Lastly, the geographical area under test is usually region- or state-wide, with none of the previous works evaluating larger areas, such as continent-wide areas. The proposal of this study is the only one that considers a large set of multiple variables at the same time. Most importantly, it includes the largest area, covering a continent-wide area corresponding to the European region.

6.2.5 Data Split

All the experiments were carried out on the proposed dataset, which was divided into training, validation and test sets, using the rule 80%- 10%-10%. The quantiles for each variable are endured to be the same in all the sets in order to have the same data distribution in all the sets.

Each soil property s_i^n was normalized using a robust scaler [243]. The use of this preprocessing allows for a mitigation of the scale effect and the effects of outliers. The robust scaler subtracts the mean and scale on the base of the interquartile range, leading to a more robust rule. Formally, it is defined as:

$$x_i^n = \frac{s_i^n - median(s_i)}{IQR_{0.25-0.75}} \tag{6.3}$$

where s_i^n is the i-th soil property at the original scale while x_i^n is the same property preprocessed with the robust scaler. $IQR_{0.25-0.75}$ is the interquartile range starting from the 25% quartile to the 75% quartile. Furthermore, the input features ($\mathbf{m}$ and $\mathbf{d}$) were normalized to the zero mean and unit standard deviation.

6.3 Results of the proposed Digital Soil Mapping algorithms

In this section, the performance of all the considered methods is assessed. An analysis of the feature importance is also presented in order to highlight the role of each single feature in the soil parameter prediction.

6.3.1 Experiments

Table 6.3(a-c) show the results achieved by all the methods considered and measured with R^2, RMSE and MAE, respectively. Two groups of experiments are shown, depending on the input features: multispectral input ($\mathbf{m}$) and multispectral input with DEM derivatives ($\mathbf{m}, \mathbf{d}$). For each group of experiments, the performance of the state-of-the-art and ANN methods is shown. For each input, the underlined values highlight the best methods for each soil variable, while the bold values highlight the best methods on average.

ANN Single and ANN Multi are the best-performing methods in terms of R^2, whatever the input features are. ANN Single and ANN Multi, with the multispectral input ($\mathbf{m}$), are the best-performing methods in terms of RMSE and MAE. In terms of RMSE, the best-performing methods when the inputs are the multispectral data and DEM derivatives ($\mathbf{m}$,$\mathbf{d}$) are ANN Single and RF. In terms of MAE, the best-performing method when the inputs are the multispectral data and DEM derivatives ($\mathbf{m}$,$\mathbf{d}$) is the SVR. To enable a visual comparison of the results, in Figure 6.3, the maps have been rendered through inverse distance weighting (IDW) interpolation of the ground truth, predictions and errors in terms of RMSE for each soil variable. As it is possible to observe, in most cases, the maps corresponding to the ground truth and the prediction of each soil variable are visually similar, indicating an accurate estimation. In agreement with the numerical results presented in Table 6.3, CaCO3, pH$_{H2O}$ and pH$_{CaCl2}$ are the ones that appear most similar. Interestingly, it is worth noting that the spatial distribution of the error is not homogeneous. In fact, an observation of the error maps of each triplet (last map) makes it

Table 6.3: Performance of the considered methods measured in terms of R^2, RMSE and MAE. For each table, the first group of rows refers to the use of multispectral input ($\mathbf{m}$)*, while the second group refers to the use of multispectral input and DEM derivatives* ($\mathbf{m}, \mathbf{d}$)*. For each input, the underlined values highlight the best methods for each soil variable, while the bold values highlight the best methods on average. For R^2 the higher is the value, the better is the method, while for RMSE and MAE the lower is the value, the better is the method.*

(a) R^2

Type	**Model**	silt	sand	pH_{H2O}	pH_{CaCl2}	coarse	clay	P	OC	N	K	CaCO3	CEC	avg
($\mathbf{m}$)	**GB**	0.13	0.18	0.42	0.41	0.05	0.18	0.08	0.16	0.11	0.08	0.25	0.08	0.18
	RF	0.19	0.25	0.48	0.47	0.06	0.26	0.07	0.13	0.12	0.10	0.33	0.11	0.21
	SVM	0.13	0.19	0.44	0.44	0.04	0.14	0.07	0.11	0.07	0.04	0.14	0.02	0.15
	ANN Single	0.23	0.28	0.54	0.53	0.11	0.26	0.13	0.20	0.17	0.13	0.37	0.12	0.26
	ANN Multi	0.25	0.30	0.51	0.50	0.13	0.26	0.12	0.21	0.15	0.18	0.43	0.14	**0.27**
($\mathbf{m}, \mathbf{d}$)	**GB**	0.22	0.27	0.47	0.47	0.14	0.28	0.13	0.19	0.13	0.10	0.33	0.14	0.24
	RF	0.36	0.41	0.54	0.54	0.16	0.39	0.12	0.20	0.17	0.12	0.41	0.22	0.30
	SVR	0.27	0.32	0.55	0.54	0.12	0.31	0.14	0.19	0.17	0.10	0.34	0.17	0.27
	ANN Single	0.33	0.39	0.57	0.57	0.17	0.35	0.12	0.25	0.20	0.12	0.46	0.20	**0.31**
	ANN Multi	0.34	0.39	0.56	0.56	0.14	0.38	0.12	0.25	0.21	0.15	0.44	0.23	**0.31**

(b) RMSE

Type	**Model**	silt	sand	pH_{H2O}	pH_{CaCl2}	coarse	clay	P	OC	N	K	CaCO3	CEC	avg
($\mathbf{m}$)	**GB**	0.96	0.92	0.76	0.77	0.96	0.89	0.87	0.90	0.95	1.04	0.88	0.99	0.91
	RF	0.92	0.88	0.72	0.73	0.95	0.84	0.87	0.91	0.95	1.03	0.83	0.97	0.88
	SVR	0.95	0.91	0.75	0.75	0.96	0.91	0.88	0.93	0.97	1.07	0.94	1.02	0.92
	ANN Single	0.90	0.86	0.67	0.69	0.91	0.85	0.85	0.88	0.92	1.02	0.80	0.97	**0.86**
	ANN Multi	0.88	0.85	0.70	0.71	0.91	0.84	0.86	0.87	0.93	1.00	0.75	0.97	**0.86**
($\mathbf{m}, \mathbf{d}$)	**GB**	0.90	0.86	0.73	0.73	0.91	0.83	0.85	0.88	0.94	1.03	0.83	0.96	0.87
	RF	0.82	0.78	0.68	0.68	0.90	0.77	0.85	0.88	0.92	1.02	0.78	0.91	**0.83**
	SVR	0.87	0.84	0.67	0.68	0.92	0.82	0.84	0.88	0.92	1.04	0.83	0.94	0.85
	ANN Single	0.84	0.79	0.66	0.66	0.90	0.79	0.86	0.86	0.91	1.02	0.75	0.92	**0.83**
	ANN Multi	0.85	0.80	0.67	0.67	0.90	0.79	0.86	0.85	0.90	1.03	0.79	0.92	0.84

(c) MAE

Type	**Model**	silt	sand	pH_{H2O}	pH_{CaCl2}	coarse	clay	P	OC	N	K	CaCO3	CEC	**avg**
($\mathbf{m}$)	**GB**	0.77	0.76	0.63	0.64	0.72	0.69	0.61	0.58	0.63	0.53	0.50	0.69	0.65
	RF	0.73	0.72	0.57	0.58	0.71	0.64	0.61	0.58	0.63	0.52	0.45	0.67	0.62
	SVR	0.77	0.75	0.59	0.60	0.67	0.66	0.57	0.52	0.59	0.48	0.42	0.66	0.61
	ANN Single	0.70	0.68	0.54	0.54	0.68	0.64	0.57	0.57	0.62	0.51	0.43	0.67	**0.60**
	ANN Multi	0.70	0.68	0.56	0.57	0.69	0.65	0.59	0.54	0.61	0.52	0.42	0.66	**0.60**
($\mathbf{m}, \mathbf{d}$)	**GB**	0.73	0.72	0.61	0.61	0.67	0.65	0.59	0.56	0.62	0.52	0.47	0.66	0.62
	RF	0.65	0.63	0.54	0.53	0.67	0.58	0.59	0.55	0.60	0.51	0.42	0.61	0.57
	SVR	0.69	0.66	0.53	0.53	0.64	0.58	0.53	0.48	0.54	0.46	0.38	0.60	**0.55**
	ANN Single	0.66	0.63	0.52	0.52	0.66	0.58	0.59	0.53	0.57	0.52	0.39	0.60	0.56
	ANN Multi	0.68	0.65	0.53	0.53	0.67	0.59	0.59	0.53	0.59	0.52	0.44	0.61	0.58

clear that for each variable, there are zones where the estimation is less accurate. This could be due to several factors. First, although the sample collection is standardized, it is subject to human factors and errors. Secondly, there may be environmental factors that mitigate the observability of the phenomenon.

Instead, focusing on the R^2, (see Table 6.3(a)), it is possible to observe that some variables can be estimated more accurately than others. For instance, the methods for the estimation of pH_{H20} and pH_{CaCl2} achieved an R^2 value higher than 0.5. On the contrary, the methods for the estimation of the variables *coarse*, *P* and *N* achieved an R^2 value lower than 0.20. Overall, the use of DEM derivatives permits the achievement of an increment

Table 6.4: The best methods for the estimation of each soil parameter, considering both types (m) and (m,d). The methods in bold represent the best type with respect to (m) and (m,d) for each metric and each soil parameter.

Soil Parameter	**(m)** R^2	**RMSE**	**MAE**	**(m,d)** R^2	**RMSE**	**MAE**
silt	ANN Multi	ANN Single	ANN Multi	**RF**	**RF**	**RF**
sand	ANN Multi	ANN Multi	ANN Multi	**RF**	**RF**	**RF/ANN Single**
pH_{H2O}	ANN Single	ANN Single	ANN Single	**ANN Single**	**ANN Single**	**ANN Single**
pH_{CaCl2}	ANN Single	ANN Single	ANN Single	**ANN Single**	**ANN Single**	**ANN Single**
coarse	ANN Multi	ANN Single/ANN Multi	ANN Multi	**ANN Single**	**RF/ANN Single/ANN Multi**	**SVR**
clay	RF/ANN Single/ANN Multi	RF/ANN Multi	RF/ANN Single/ANN Multi	**RF**	**RF**	**RF/SVR/ANN Single**
P	ANN Single	ANN Single	ANN Single	**SVR**	**SVR**	**SVR**
OC	ANN Multi	ANN Multi	ANN Multi	**ANN Single/ANN Multi**	**ANN Multi**	**SVR**
N	ANN Single	ANN Single	ANN Single	**ANN Multi**	**ANN Multi**	**SVR**
K	**ANN Multi**	**ANN Multi**	ANN Multi	ANN Multi	RF/ANN Single	**SVR**
CaCO3	ANN Multi	**ANN Multi**	ANN Multi	**ANN Single**	**ANN Single**	**SVR**
CEC	ANN Multi	RF/ANN Single/ANN multi	ANN Multi	**ANN Multi**	**RF**	**SVR/ANN Single**
avg	ANN Multi	ANN Single/ANN Multi	ANN Multi	**ANN Single/ANN Multi**	**RF/ANN Single**	**SVR**

of 19% and 15% in the case of ANN Single and Multi, respectively. In particular, the use of DEM derivatives improved the performance in the estimation of soil textures (silt, sand, coarse, and clay), in terms of R^2, on average by 43% and 30% in the case of ANN Single and ANN Multi, respectively. In the case of ANN Multi, the improvement is mitigated by the fact that this method predicts multiple variables at the same time. This result was expected since the soil properties are closely related to geological formations and landscape positions and, in particular, the soil textures are highly correlated to the parameters derived from the DEM [244]. Furthermore, the use of DEM significantly improved the estimation of the parameter CEC by about 65%, whichever method is employed. This behavior is due to the fact that the CEC is often correlated with the DEM because exchangeable cations can be mobilized and leached to lower landscape positions [245].

Figure 6.4 shows, for each soil variable, the scatter plots gathered by the best overall model (ANN Multi). Each plot shows the prediction vs. the ground truth values, and ideally, each point should lie on the bisect of the graph. The trend line (in orange) is also shown and its coefficients are provided. Finally, to summarize the results achieved by each method and to show which combination of types is best, Table 6.4 reports, for each soil parameter and each metric (considering both (**m**) and (**m**,**d**)), the best technique for the estimation. For all the soil parameters, apart from K, the use of DEM in combination with multispectral imagery improves the estimation accuracy, whichever evaluation metric is used. It is worth noting that in some cases, ANNs struggle to exploit the multimodality (**m**,**d**) with respect to other machine learning methods, which in turn adapt better to a diverse input. Nevertheless, ANNs on average outperform all the other methods evaluated, as confirmed in Table 6.3(a) and (b).

6.3.2 Explainability Discussion

As described in Section 6.1.1, it was also investigated the importance of an input feature in predicting a given output using a feature permutation approach. This investigation is fundamental to an understanding of the actual advantages of using the combination of

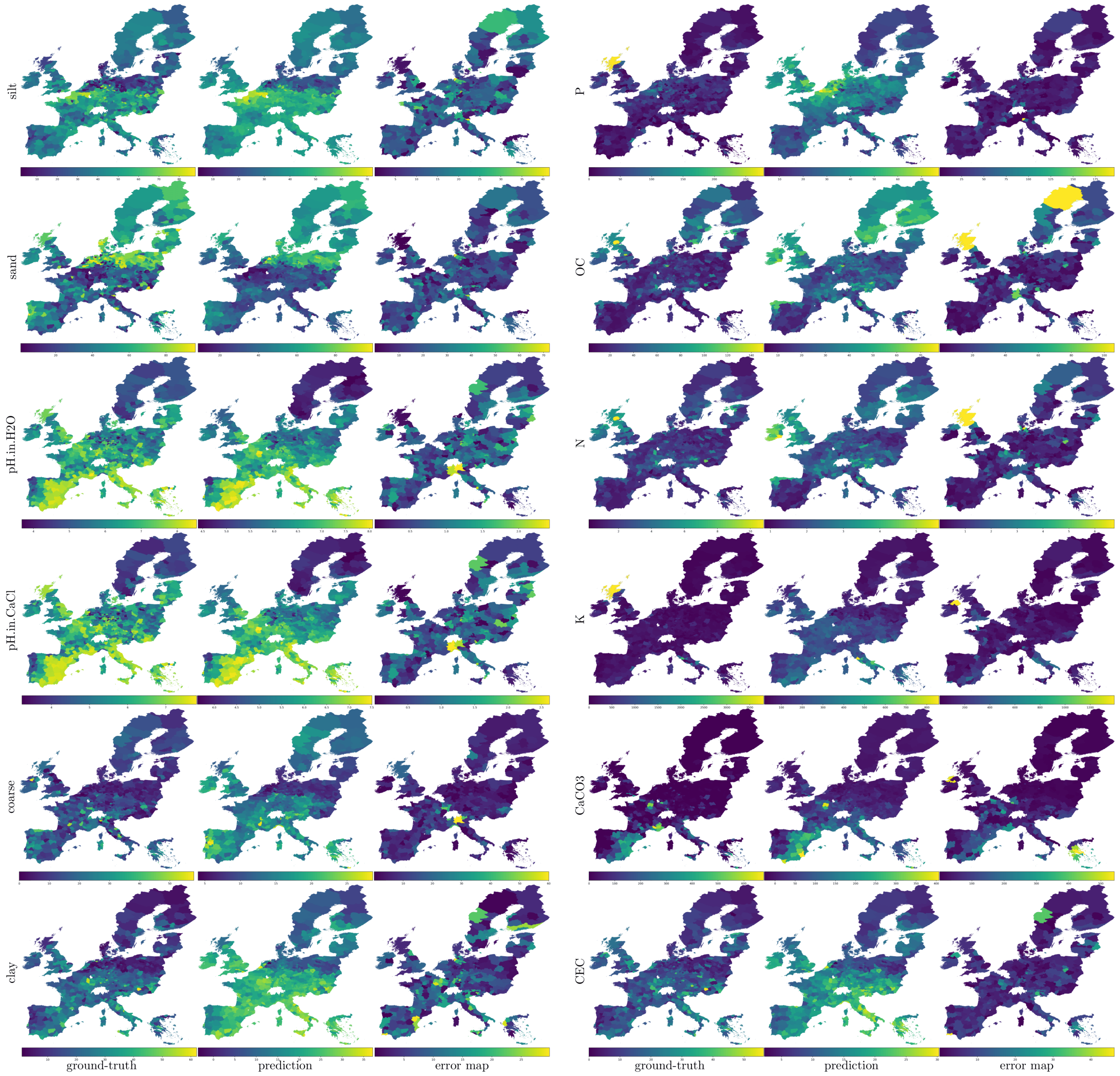
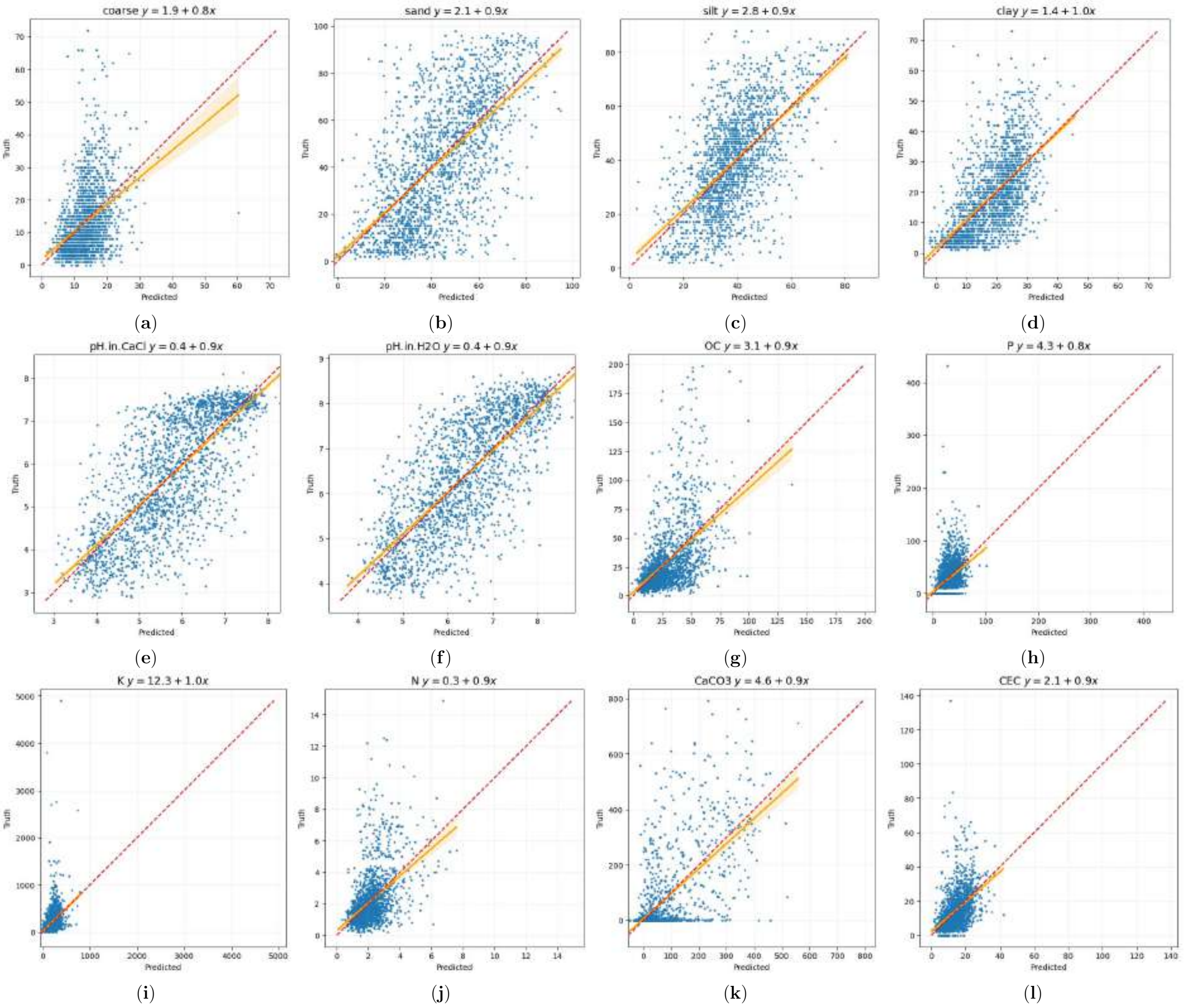


Figure 6.3: Visual comparisons of the results have been rendered through inverse distance weighting (IDW) interpolation of the ground points relative to the ground truth, the predictions, and the errors in terms of RMSE of each soil variable. The blue and yellow colors represent the minimum and maximum values of each soil property, respectively. Colors relative to intermediate values are obtained through quantile color coding. It is worth noting that in most cases the ground truth (the first image of each triplet) and the prediction (the second image) are visually similar, indicating an accurate estimation of the soil properties.

spectral and DEM information to describe the properties of the soil.

Figures 6.5 and 6.6 show the feature importance in the estimation of each soil variable. The blue bars represent the bands of the multispectral signal, while the orange bars represent the DEM derivatives. It is possible to observe that for some variables, such as *coarse* and *P*, all the bands have more or less the same significance, while with respect to other variables, there are spiking bands that heavily impact on the results, such as bands 8 and 9 in the estimation of pH in H_2O.

In the case of the geomorphological features, the band that most significantly impacted



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

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 6.4: Scatter plots corresponding to the predictions of ANN Multi in relation to the soil parameters. The x axis refers to the predictions, while the y axis refers to the ground truth values. For reference, each plot includes the perfect prediction line $y = x$ (dashed red) and the trend line relative to the estimations (solid orange). Each subfigure represents the scatter plot of a given soil variable.

on the predictions is the *valley depth.* This is because the *valley depth* is a vital indicator of a depositional (sedimentary) environment [246]. On the contrary, the *slope length* does not provide a significant contribution to the prediction of the soil properties.

Figure 6.7 shows the comparison between the two groups of considered features (multispectral and DEM) in terms of their importance in the estimation of the soil variables. For a given variable, blue and orange bars represent the percentage of multispectral bands and DEM, respectively. Overall, the use of multispectral features is more important than the use of DEM derivatives with a ratio on average of 60% vs. 40%. In some cases, such as the estimation of pH_{CaCl} and pH_{H2O}, the use of DEM derivatives counts for about 30%, thus confirming that geomorphological features are less important in the estimation of pH levels.

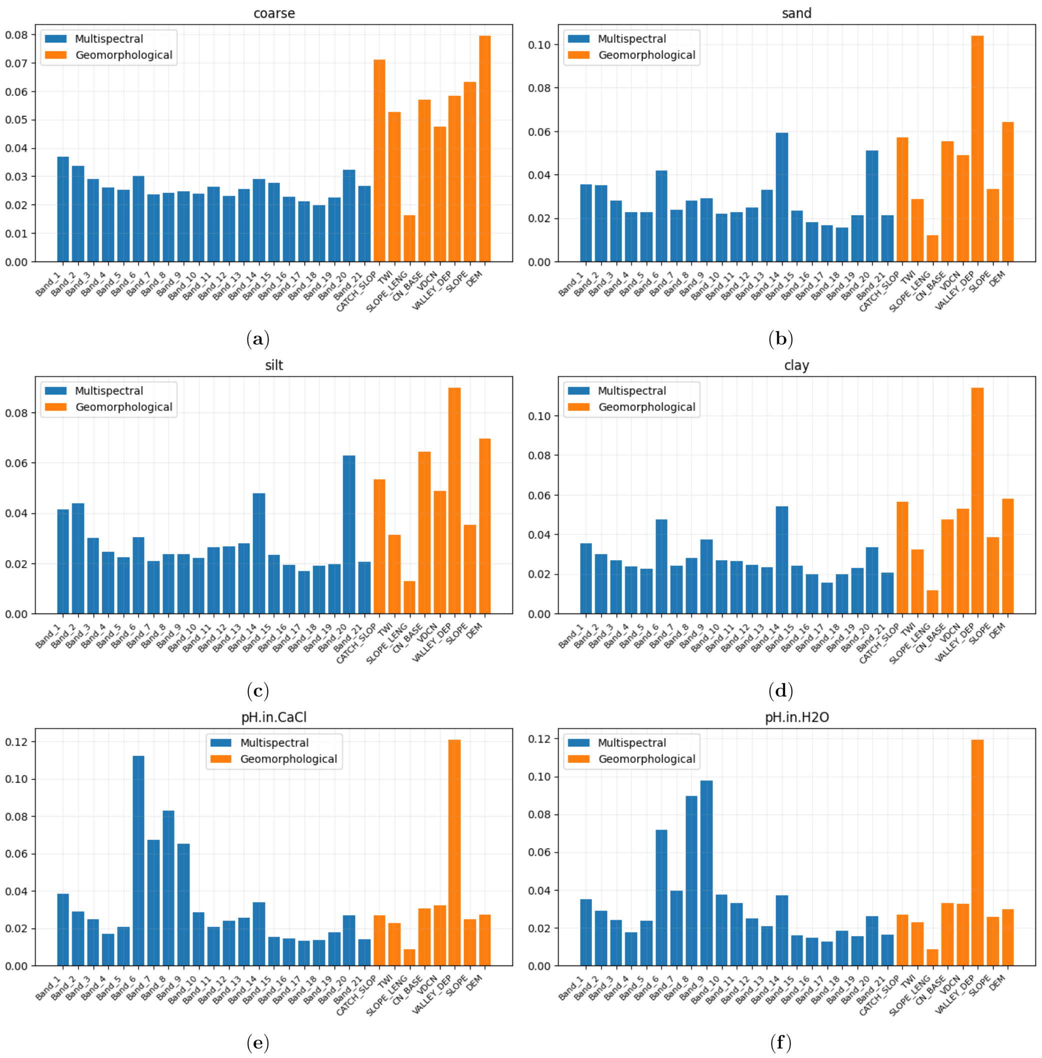
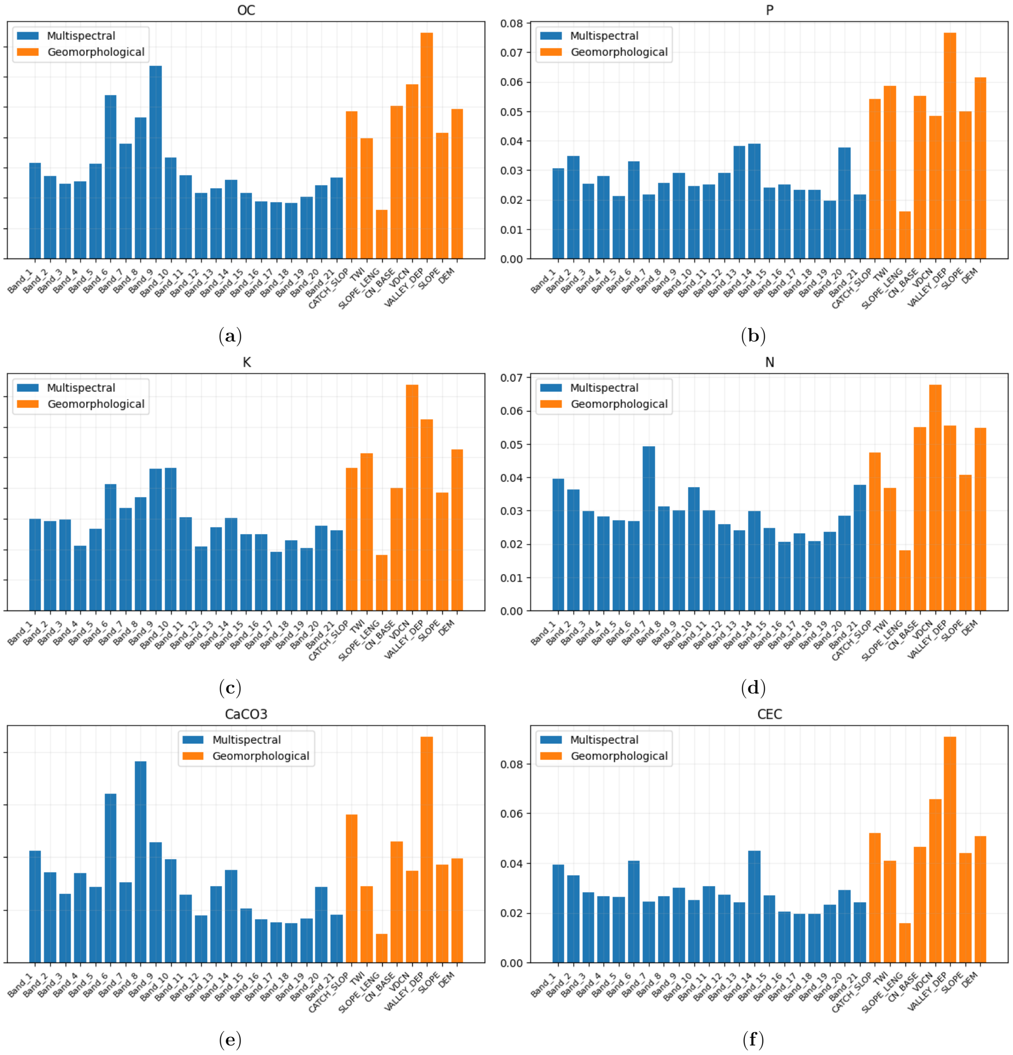


Figure 6.5: Feature importance of each soil property obtained using random forests. The blue bars represent the importance of the multispectral bands, while the orange bars are relative to the DEM-derivatives. (a) Coarse; (b) Sand; (c) Silt; (d) Clay; (e) pH in $CaCl_2$; (f) pH in H_2O.

6.4 Usefulness of spectral and Digital Terrain Model information for Digital Soil Mapping

This work has evaluated different machine learning-based methods for the estimation of the multiple soil characteristics of a continent-wide area corresponding to the European region from multispectral signal and DEM derivatives. The multispectral signals, DEM derivatives, and soil characteristics have been gathered respectively from the Sentinel-3 satellite, the European Copernicus mission and the LUCAS library, respectively. All the data collected were then geographically matched to create a uniform and multisource benchmark of 20,000 samples. On this dataset, several machine learning methods, representing the state of the art for the estimation of soil characteristics, have been benchmarked. These methods were adapted to use multispectral signals, DEM derivatives and a combination of each of



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

Figure 6.6: Feature importance of each soil property obtained using random forests. Blue bars represent the importance of the multispectral bands, while orange bars are relative to the DEM-derivatives. (a) Organic Carbon (OC); (b) Phosphorous (P); (c) Potassium (K); (d) Nitrogen (N); (e) Calcium Carbonate (CaCO3); (f) Cation exchange capacity (CEC).

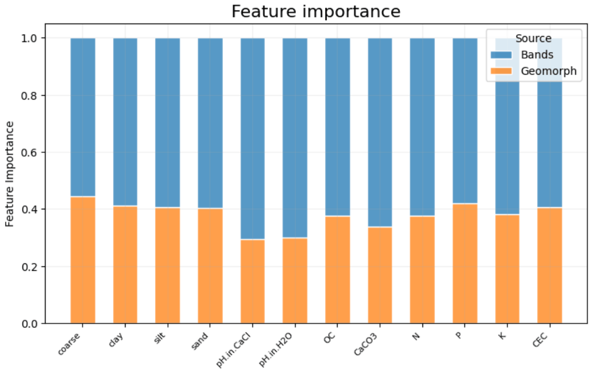


Figure 6.7: Aggregated feature importance of each soil variable. The contributions of the multispectral and geomorphological DEM derivatives are shown in blue and orange, respectively. It is worth noting that the importance of the multispectral variables is greater than that of the DEM derivatives.

them. An ANN-based method capable of predicting all the soil properties at the same time has also been proposed and included in the investigation. Three metrics were used for the performance assessment: MAE, RMSE and R^2. Overall, neural networks showed the best performance in describing the data. The experiments of this work also demonstrated how the use of DEM derivatives improves the quality of the predictions in contrast with the use of the multispectral signal alone. The improvement in terms of R^2 increment is 19% on average, this being greatly appreciated in the prediction of soil textures where it reaches the 43%. Moreover, the correlation between DEM and CEC has allowed for a significant improvement of about 65%. Further analysis of the feature importance revealed a high impact of the multispectral bands 8 and 9 in the estimation of the pH in H_2O. In the case of the geomorphological features, the DEM derivative *valley depth* is the variable that most significantly impacted on all the predictions. Overall, the use of multispectral bands is more important than the use of DEM derivatives by a ratio of 60–40%. The study area includes the entire European region, comprising an extensive collection of soil samples with a remarkable diversity and heterogeneity. The analysis presented in this thesis provides insights into the potential of machine learning techniques to generalize over a vast geographical area. Nevertheless, given the substantial variations in soil properties across different regions, the validity of these findings for areas beyond Europe should be empirically verified.

The outcomes achieved for some soil variables display more promising results with respect to others; for instance, pH and soil textures (clay, sand, and silt) exhibit superior predictability compared to potassium and nitrogen. Nevertheless, the numerical outcomes, measured in terms of R^2, closely align with the findings from analogous scientific papers [146], even if on a significantly larger dataset. This study validates the efficacy of remote sensing methodologies for soil parameter estimation. Despite the presence of estimation errors, these methods offer numerous advantages over conventional approaches. Remote sensing offers significant advantages in terms of spatial coverage, real-time monitoring, non-invasiveness and the ability to capture multispectral information. These advantages make it a powerful tool for soil parameter estimation that complements or even surpasses traditional methodologies in terms of efficiency, accuracy and practicality.

In future works, it would be interesting to include hyperspectral signals in the assessment of the machine learning methods and to compare the quality of the predictions with that of the multispectral signals.

Chapter 7

Multimodality in a Real Scenario - Estimation of Soil Parameters for Agricultural Areas Management

This chapter refers to the paper *Multimodal Earth Observation Modeling using AI* [24] presented at the *MESAS 2023* conference (International Conference on Modelling and Simulation for Autonomous Systems) organized by NATO Modelling and Simulation Centre of Excellence [1].

In this chapter, a real use case strategy for the management of agricultural resources which is of uttermost importance among the many tasks related to Earth Observation (EO). The methods analyzed in the previous chapter are combined to propose a single pipeline that takes advantage of multimodal approaches to investigate the needs of agricultural areas, by estimating their properties.

In particular, the pipeline, shown in Figure 7.1, aimed at automatically identifying agricultural soils and estimating their chemical and physical properties. Thanks to this automatic estimation, it is possible to implement agricultural policies aimed, on the one hand at finding the most compatible soils for a given crop, and on the other hand, at implementing timely interventions on existing crops. The designed pipeline consists of 1) segmenting the soil to understand if a terrain is used for agriculture or not, and 2) estimating the properties of the soil in agriculture areas. As shown in chapter 6 and 5, the use of AI techniques allows for powerful, data-hungry tools that in many tasks have proven to be fast and accurate in both segmentation and parameter estimation, being able to generalize the problem to different contexts when modeled with sufficient data. Moreover, as previously demonstrated, multimodal approaches also combine different types of sources (or modalities) to exploit their inherent advantages and improve performance.

Regarding the multimodal semantic segmentation, the Ticino dataset [21] proposed

[1] MESAS conference: `https://www.mscoe.org/event/mesas-2023/`

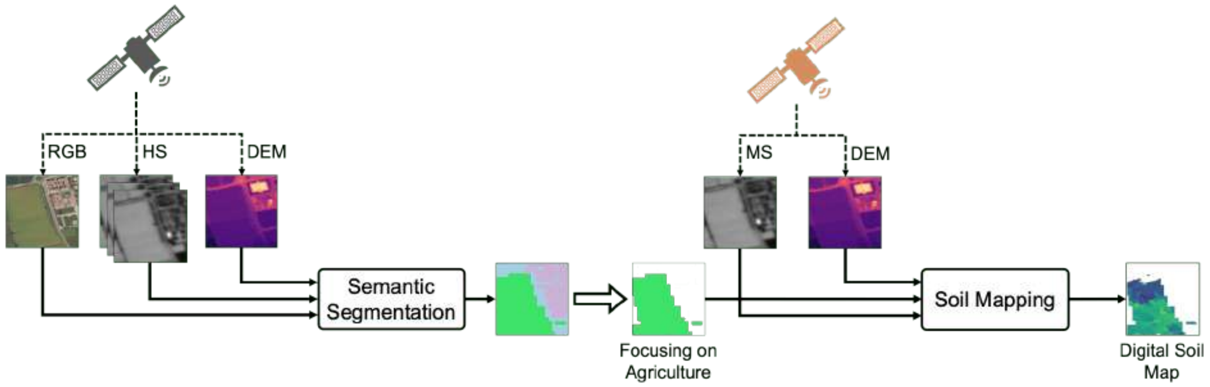


Figure 7.1: The analysis of the chemical and physical properties of agricultural soils is carried out in two stages: in the first, such soils are automatically located by semantic segmentation, while in the second, soil properties are predicted from the combination of different signals. Both stages require the use of multimodal sources that include spectral information.

and described in chapter 5 has been used. For the estimation, the multimodal dataset used in chapter 6 has been used to combine spectral and elevation information. The datasets differ from each other because there is currently no HS dataset that covers an area with a truly large set of parameters estimated by combining spectral information and labeling; therefore, in order to have a reasonable dataset that can take advantage of the real benefits of AI, the choice was to rely on multispectral information. To demonstrate the advantages of multimodal approaches in a possible real scenario, neural networks have been used, experimenting with different combinations of inputs and including single vs multimodal comparisons to show how much the second approaches can improve on a more practical task.

This work focuses on the prediction of textural features (which include silt, sand and clay), pH$_{H_2O}$ and pH$_{CaCl_2}$. The soil textures play a determining role when it comes to behaviors such as water-holding capacity, drainage characteristics, nutrient retention, and susceptibility to erosion, influencing plant growth and agricultural productivity and, thus, being fundamental to agricultural decisions. In fact, different plants have specific pH requirements for optimal growth. The pHs affect nutrient availability and microbial activity in the soil, giving important information on the soil health.

7.1 Segmentation and Estimation

This section will provide a comprehensive description of the pipeline. Firstly, this discussion will focus on Semantic Segmentation as a tool used to identify agricultural areas and the methods used to demonstrate the improvement achieved by a multimodal approach. Secondly, it will describe the process of estimating soil parameters, while consistently examining the differences between single- and multimodal approaches.

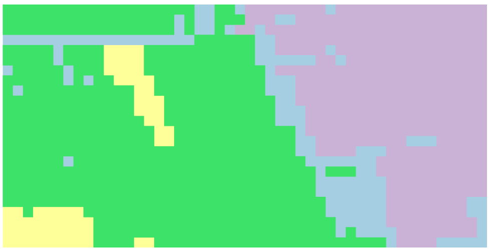


Figure 7.2: Regrouped SAU labeling.

7.1.1 Semantic Segmentation module

As mentioned above, the Ticino dataset [21] has been used. In particular, the pansharpened version has been considered where PAN and the two spectral cubes have been fused together to obtain a final all-comprehensive HS image of 182 bands and 5 m/px (HS↑). For the experiments, the SAU labeling was considered to identify the agricultural areas. It is originally characterized by 10 classes that, for the purpose of this work, were regrouped into 5 specific classes: *Background*, *Agriculture*, *Man-made areas*, *Water bodies*, *Natural Vegetation*, where *Agriculture* is the union of *Other agricultural crops*, *Forage crops*, *Corn*, *Industrial plants*, *Rice*, and *Seeds*. A sample result of the regrouping can be seen in Figure 7.2. The goal of this regrouping is to facilitate the fast identification of agricultural areas and non-agricultural areas. All the dataset was used with the same splitting in training, validation and test. In particular, the 1502 samples with all available modalities were respectively split into 1051, 225, and 226 images.

To achieve the segmentation and the comparisons, the same architecture and setup as the best CNN method tested in 5.3.1 have been used. Single modalities with RGB and HS↑ and the two combinations (RGB + HS↑) and (RGB + HS↑ + DTM) were tested with the Middle fusion technique and the ResNet18 encoder that demonstrated to be the best method in the analysis of the previous chapter.

7.1.2 Digital Soil Mapping Module

Symmetrically, the dataset used for Digital Soil Mapping experiments is the same dataset used in chapter 6. It includes Sentinel-3 multispectral, DEM information from the Copernicus project and parts of the LUCAS [155] dataset [23]. The estimation focused on the prediction of textural features (which include silt, sand and clay), pH_{H_2O} and pH_{CaCl_2}.

To estimate the soil properties, the ANN multi-model described in chapter 6 for estimating the parameters at the same time has been used with the same setup. Even in this case, a comparison between single- and multimodal approaches with (MS) and (MS + DEM).

Table 7.1: *Segmentation results of Soil Agricultural Use divided by modalities combination. Bold values represent the best performance obtained on the rows. For all the metrics, the higher is the better.*

		Single modality		Multimodality	
Class	Metric	RGB	HS↑	(RGB + HS↑)	(RGB + HS↑ + DTM)
Agriculture	Acc	0.31	0.38	**0.46**	**0.46**
	IoU	0.21	0.28	0.33	**0.34**
	Precision	0.34	0.47	0.51	**0.54**
Man-made areas	Acc	0.89	0.89	**0.90**	**0.90**
	IoU	**0.77**	0.76	**0.77**	**0.77**
	Precision	**0.85**	0.83	0.84	0.84
Water bodies	Acc	0.56	**0.72**	0.66	0.69
	IoU	0.46	0.55	0.55	**0.56**
	Precision	0.72	0.69	**0.77**	0.75
Natural Vegetation	Acc	0.82	0.83	0.84	**0.85**
	IoU	0.64	**0.67**	**0.67**	**0.67**
	Precision	0.75	**0.78**	0.77	0.76
Overall	Acc	0.65	0.71	0.72	**0.73**
	IoU	0.52	0.57	0.58	**0.59**
	Precision	0.67	0.69	**0.72**	**0.72**

7.2 Results and Discussion

In this section, the results achieved on both tasks are discussed, focusing on the advantages of multimodal approaches and showing the importance of pursuing these kinds of strategies.

7.2.1 Soil Agricultural Use Semantic Segmentation

Table 7.1 shows the results obtained for each class and the overall mean on the four classes, comparing the different combinations of input used in the experiments. The performance is again evaluated in terms of Accuracy (Acc), Intersection over Union (IoU) and Precision and the bold value represents the best performance on each row of the table.

As demonstrated before, it is immediately observable that, *Overall*, the multimodal approaches take the lead in terms of performance for every metric. In particular, using three modalities (RGB + HS↑ + DTM) always achieved the best average performance. By comparing these approaches with the RGB-only technique, an important increment is visible on each metric. In detail, Acc improves by about 8%, IoU by 7% and Precision by 5%. If multimodal approaches are compared with the HS↑-only method, the increment is inferior, showing the importance of HS data, but respectively, each metric improves by 2%, 2% and 3%.

The primary subject of this research is the Agriculture class. In this case, it is worth mentioning that the (RGB + HS↑ + DTM) combination reaches the best performance with a significant increment with respect to RGB-only and HS↑-only. This result remarks the

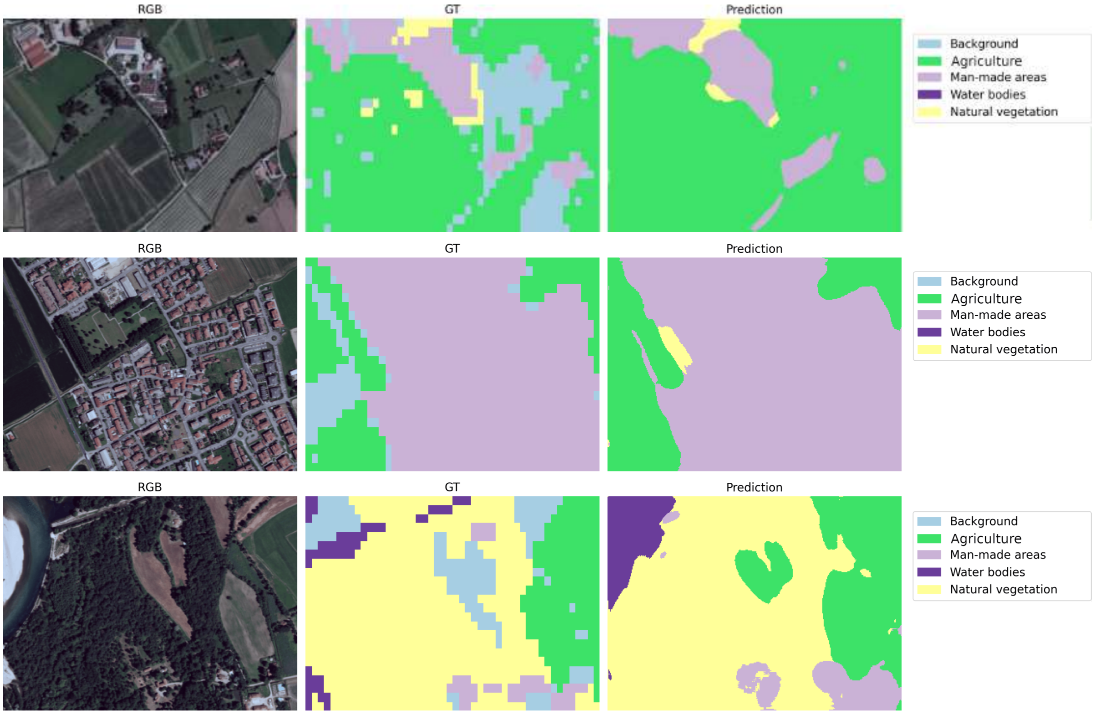


Figure 7.3: Visual prediction of Soil Agricultural Use segmentations with middle fusion approach using all the modalities (RGB+HS↑+DTM).

importance of combining more modalities because each of them brings different advantages and being able to exploit all of them together improves the power of an AI model. The RGB-only modality exhibited the worst performance across all metrics, indicating that it is not the optimal modality for this kind of class.

The difference between using three modalities and RGB-only is 15% for Acc, 13% for IoU and 20% for Precision. HS↑-only achieved better performances in identifying Agriculture pixels than RGB-only with a difference from multimodal of 8%, 8% and 7% for Acc, IoU and Precision, respectively.

Figure 7.3 shows the visual results of the segmentation model with (RGB + HS↑ + DTM) with the RGB image in the left column, the ground truth in the center and the prediction in the right column. It is possible to note that the segmentation recognizes the different elements that compose the scene, both in their location and semantics.

From the Overall and more specific results, it is possible to conclude that generally, a multimodal approach is preferable because the complementarity of the modalities allows for a better description and understanding of the soil. HS↑ alone is already a great improvement compared with the standard RGB, but using it in combination with better spatial resolution and morphological information, utterly improves overall results and especially improves the identification capabilities of *Agriculture* areas.

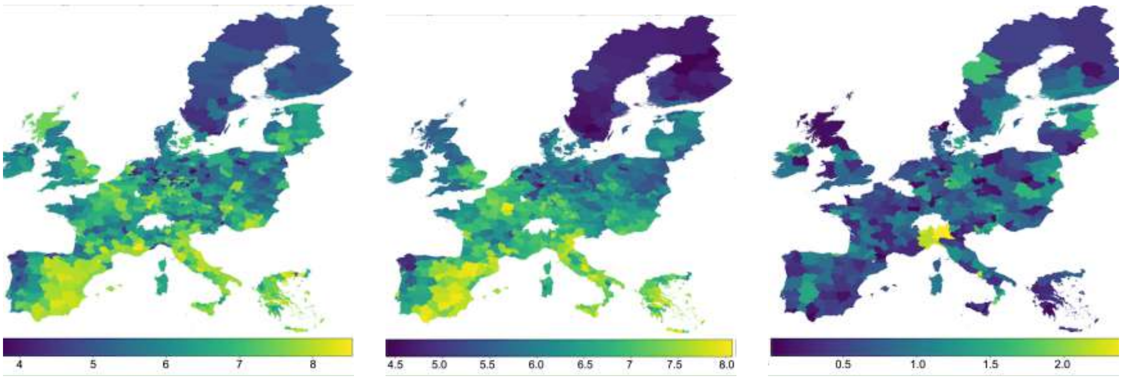
Table 7.2: Performance of the considered methods measured in terms of R^2, RMSE and MAE. For each table, the first group of rows refers to the use of multispectral input (MS), while the second group refers to the use of multispectral input and DEM derivatives (MS + DEM). For each input, the bold values highlight the best methods on average. For R^2 the higher is the better, while for RMSE and MAE the lower is the better.

Parameter	Metric	Single modality (MS)	Multimodality (MS+DEM)
Textures	R^2 (↑)	0.27	**0.37**
	RMSE (↓)	0.86	**0.81**
	MAE (↓)	0.68	**0.64**
pH$_{H_2O}$	R^2 (↑)	0.51	**0.56**
	RMSE (↓)	0.70	**0.67**
	MAE (↓)	0.56	**0.53**
pH$_{CaCl_2}$	R^2 (↑)	0.50	**0.56**
	RMSE (↓)	0.71	**0.67**
	MAE (↓)	0.57	**0.53**
	R^2 (↑)	0.43	**0.50**
Overall	RMSE (↓)	0.76	**0.72**
	MAE (↓)	0.60	**0.57**

7.2.2 Digital Soil Mapping

Table 7.2 reports the results achieved estimating the Texture and pHs of the soil, comparing an ANN single modality approach based on multispectral input with an ANN multimodality approach using multispectral and DEM derivatives as input. The evaluation metrics used to measure the performance are still R^2 (coefficient of determination), RMSE (Root Mean Square Error) and MAE (Mean Absolute Error). R^2 is the only one where the higher values indicate better performance, while the others work in the opposite way.

Considering these three parameters and looking at Table 7.2, it is unequivocal that every time the DEM is involved the performance is better. Each of the metric, for *Textures*, pH$_{CaCl_2}$ and pH$_{H_2O}$ is increased with the multimodal approach. Overall the improvements are of 0.07, 0.04 and 0.03 for R^2, RMSE and MAE, respectively. The pHs, in particular, have a R^2 of 0.56 with multimodal methods, proving to be estimable with this kind of approach. On the other hand, *Textures* achieved a generally lower performance, but the use of DEM has guaranteed an improvement of 0.10 on R^2, showing room for improvements in this line of study. Nonetheless, the usefulness of RS technology in their ability to perform accurate acquisition in real-time is strong enough to pursue this strategy and continue to investigate multimodal RS approaches. The results clearly show that introducing multimodality for the estimation of parameters, combining multispectral and DEM information, improves our ability to estimate the parameters of the soil. Figure 7.4 shows the visual results of the estimation of pH$_{H_2O}$. These results have been generated through inverse distance weighting (IDW) interpolation: left is the ground truth, center is



(a) Ground Truth (b) Prediction (C) Error Map

Figure 7.4: Digital soil mapping of pH$_{H_2O}$ rendered through inverse distance weighting (IDW) interpolation: (a) is the ground truth, (b) is the predictions, and (c) is the error map in terms of RMSE.

the predictions and right is the error map in terms of RMSE.

7.3 Final remarks on the Estimation of soil parameters for Agricultural areas management

The presented pipeline is able to automatically find agricultural areas and consequently estimate their soil properties. This methodology can support the management of the resources, making it easier to maintain high standards in soil health. The pipeline makes use of multimodal RS images with different characteristics, and it creates maps of parameters for each segmented area of interest. It consists of two modules: 1) Semantic Segmentation; and 2) Digital Soil Mapping. The first uses RGB, HS and DTM information to provide an accurate segmentation of the areas and in particular of the terrain dedicated to agriculture. The second makes use of multispectral and DEM images to extract *Textures*, pH$_{H_2O}$ and pH$_{CaCl_2}$, creating new maps that convey information for the handling of the territory. The scope of this work is to demonstrate how multimodal approaches can impact performance in a real use case, thus improving their efficiency and accuracy.

The results, both on Semantic Segmentation and Digital Soil Mapping, remark again that combining complementary information from different modalities improves the overall performance. In the case of semantic segmentation, exploiting the advantages of HS already proves to be a better choice than typical RGB information. Combining HS with RGB and DTM utterly improves this performance. In the same way, the use of MS-only information in Digital Soil Mapping was overcome by combining MS and DEM. To further note is the advantages of using RS images that nowadays are easy to acquire, giving us a real-time technology and thus having a great advantage over collecting data on the field.

In conclusion, this procedure demonstrates that multimodal approaches in tasks such as semantic segmentation and regression of parameters with RS technology are viable and more efficient than the standard single modality approach for EO. In the future, with the increment of HS data, this modality could also be integrated into the Digital Soil Mapping module, further improving our ability to describe the soil. The same procedure could be divided, and the estimation could be used to find feasible terrains for agricultural areas, finding a balance between our necessities and the use of Earth's resources.

Chapter 8

Conclusions

In this thesis, the usefulness of multimodal approaches in the remote sensing (RS) field of study has been comprehensively investigated. Earth Observation (EO) is a field of study of great importance, involved in many tasks. Therefore, being able to use the most efficient and effective techniques is necessary. Among the most powerful techniques that allow the extraction of the best information from images, AI and multimodality have represented a really big advancement in different fields of study. These approaches create a great opportunity to better understand and describe the terrain, bringing huge advantages in many applications. This work investigated these approaches through the different steps that characterize the use of RS for EO and real applications. The analysis focuses on different tasks that are proven to be of great importance in this scope. Starting from the acquisition of data to a real use case and possible applications of AI technology, in this thesis, the use of information diverse from the typical RGB images has been investigated, creating a compendium on multimodal AI-based RS methods for EO. In particular, this work focused on the following tasks: Hyperspectral Pansharpening [20], Unsupervised Segmentation of hyperspectral images [10], Multimodal Supervised Semantic Segmentation [21, 22], Digital Soil Mapping [23], and Multimodality in a Real Scenario through the Estimation of Soil Parameters for Agricultural Areas Management [24].

As defined above, when it comes to RS and multimodal approaches, the first fundamental step to address is the acquisition and in particular the analysis of data with their complementarity information and characteristics. The data used in this work comprehend spatial, spectral and morphological information, spanning RGB, multispectral (MS), hyperspectral (HS), panchromatic (PAN), and digital elevation model (DEM) that have been combined to improve performance in many areas of the EO. These areas include HS pansharpening, HS unsupervised segmentation, supervised semantic multimodal segmentation of RS images, and digital soil mapping. For all of them a dataset including different sources has been built, investigating the modalities and their advantages.

One of the first and most important elements of an RS pipeline is the enhancement of the original data. In this matter, hyperspectral pansharpening is a fundamental strategy that combines high-spatial-resolution panchromatic images and hyperspectral data to achieve high spatial and spectral resolution new data that share the best characteristics of both information. The analysis elaborated in this work investigated the current state of the art, illustrating its strengths and weaknesses. A novel dataset has been built, with the highest number of images compared with any other datasets used in this field. The most popular and effective hyperspectral pansharpening techniques have been adapted to this new dataset in the first comprehensive analysis of a statistically relevant dataset.

The unsupervised segmentation analyzed the intrinsic capacity of HS information in identifying different materials and thus its usefulness for RS images. The advantages of an unsupervised segmentation concern in particular the creation of ground truth, making the process faster and less subjected to errors. The proposed method based on superpixels outperformed other techniques being at the same time more flexible, robust to different kinds of noises and not needing training, thus being helpful in the creation of new multimodal datasets.

The analysis and investigation of different modalities and techniques helped to create the Ticino dataset. This dataset presents the highest number of modalities than any other multimodal dataset and a high cardinality suitable for deep learning methods. In the supervised semantic multimodal segmentation investigation, the usefulness of multimodal approaches has been proved with two sets of experiments and by comparing different deep learning techniques and combinations of modalities. The first set of experiments consisted of comparing different combinations of modalities with early and middle fusion techniques based on CNNs. The model used for the experiments was a U-shaped network with a Resnet18 backbone. The results demonstrate that overall single modality approaches with only RGB or HS data were outperformed by multimodality approaches. In particular, the middle fusion technique achieved the best results, demonstrating that, even if, multimodality alone improves our ability to describe the terrain, choosing the right method of fusion is of utter importance for the final results. To further analyze this conclusion, the second set of experiments investigates different techniques of fusion based on Transformer models and in particular the Swin-Upernet model. This second set of experiments demonstrates that fusion methods that tend to extract high-level features from each modality are more capable of exploiting the advantages of heterogeneous and complementary modalities. In particular, a late concatenation method performed better than any other approach. Moreover, the test demonstrated that better performance can be achieved without increasing the complexity and resources used. In fact, the two methods Late concatenation and Token Fusion at Attention Level, used a comparable number of parameters with RGB only technique.

Finally, the Digital Soil Mapping investigation demonstrates that combining spectral and morphological information, respectively with MS images and DEM, improves the possibility of estimating the textures and chemical parameters of the soil. The experiments reported compared MS single modality with MS and DEM multimodality using various machine learning techniques. The results, based on 3 evaluation metrics, demonstrated the advantages of combining different information. Moreover, an analysis of the importance of the variable in this task, showed how much DEM can be relevant for the estimation. The results demonstrate a ratio of 60-40% between MS and DEM.

All the presented works demonstrated how multimodality can impact RS applications and all the steps that characterize them. Starting from the acquisition of new data, to their enhancement and finally to their use for EO, in this thesis, new datasets, analysis, and methods are presented to make full use of multimodality in this field. One of the most important results is represented by the possibility of using each of these steps in a real use case that can actually show the potential of multimodality and why it is necessary to perpetrate the research. To corroborate this possibility, this thesis also presented a real use case where the estimation of the parameters of agricultural areas for the evaluation of soil health takes advantage of multimodality to achieve better results than single modality. The strategy combined different steps of the multimodal pipeline considered in this thesis, using the Ticino dataset, the enhancement of HS images, the multimodal supervised semantic segmentation and the Digital Soil Mapping to improve the estimation of textures and pH parameters. Even in this case, the usefulness of multimodality is demonstrated.

To summarize, a pipeline for EO analysis and application has been presented in this thesis. This pipeline exploited the potential of multimodal approaches, covering all the steps that go from the acquisition of data to real applications and descriptions of the soil. In particular, the thesis addressed a specific task typical of EO, involving the use of data that diverge from the typical RGB images. The scope was to demonstrate that multimodality is one of the most promising routes to investigate in the RS field of study. To this purpose, different kinds of data, including HS images, were studied and analyzed, building new datasets, investigating the state of the art and creating new techniques for EO. In each of these tasks, this thesis shows how multimodality can improve and outperform single modality approaches, building, at the same time, techniques and tools that can further help future investigations. In fact, many tasks can be addressed starting from this work. Future works can involve the tools introduced and provided in this work and also different tasks. The refinement of the labeling for the Ticino dataset with the help of unsupervised techniques, a completely new HS pansharpening method, and the use of HS data for the estimation of parameters are all possible improvements and advancements of the works proposed in this thesis. Instead, other tasks that can benefit from the use of multimodality, even using the proposed investigations as a base, include spatio-temporal fusion heterogeneous data, new techniques for improving pansharpening, weather prediction, climate change management, refugees, and emergency identification and handling. All these tasks are of crucial importance for our life on Earth and how we build our future, and all these tasks would benefit from more data and different information that can be combined to improve our understanding of EO.

References

[1] Hualou Long, Yingnan Zhang, Li Ma, and Shuangshuang Tu. Land use transitions: Progress, challenges and prospects. *Land*, 10(9):903, 2021.

[2] Thomas Blaschke. Object based image analysis for remote sensing. *ISPRS journal of photogrammetry and remote sensing*, 65(1):2–16, 2010.

[3] Hua-Dong Guo, Li Zhang, and Lan-Wei Zhu. Earth observation big data for climate change research. *Advances in Climate Change Research*, 6(2):108–117, 2015.

[4] Sentinel-3. QGIS Software.

[5] Sentinel-3. sentinel-3 mission.

[6] Prisma satellites data. ASI.

[7] Sancho Salcedo-Sanz, Pedram Ghamisi, María Piles, Martin Werner, Lucas Cuadra, A Moreno-Martínez, Emma Izquierdo-Verdiguier, Jordi Muñoz-Marí, Amirhosein Mosavi, and Gustau Camps-Valls. Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion*, 63:256–272, 2020.

[8] Rajendra P Sishodia, Ram L Ray, and Sudhir K Singh. Applications of remote sensing in precision agriculture: A review. *Remote Sensing*, 12(19):3136, 2020.

[9] Annett Frick and Steffen Tervooren. A framework for the long-term monitoring of urban green volume based on multi-temporal and multi-sensoral remote sensing data. *Journal of geovisualization and spatial analysis*, 3(1):6, 2019.

[10] Mirko Paolo Barbato, Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Unsupervised segmentation of hyperspectral remote sensing images with superpixels. *Remote Sensing Applications: Society and Environment*, 28:100823, 2022.

[11] Thilo Wellmann, Angela Lausch, Erik Andersson, Sonja Knapp, Chiara Cortinovis, Jessica Jache, Sebastian Scheuer, Peleg Kremer, André Mascarenhas, Roland Kraemer, et al. Remote sensing in urban planning: Contributions towards ecologically sound policies? *Landscape and urban planning*, 204:103921, 2020.

[12] CJ Van Westen. Remote sensing for natural disaster management. *International archives of photogrammetry and remote sensing*, 33(B7/4; PART 7):1609–1617, 2000.

[13] Wei Han, Xiaohan Zhang, Yi Wang, Lizhe Wang, Xiaohui Huang, Jun Li, Sheng Wang, Weitao Chen, Xianju Li, Ruyi Feng, et al. A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:87–113, 2023.

[14] V.L. Mulder, Sytze de Bruin, Michael Schaepman, and T.R. Mayr. The use of remote sensing in soil and terrain mapping - a review. *Geoderma*, 162:1–19, 04 2011.

[15] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[16] Qibin He, Xian Sun, Wenhui Diao, Zhiyuan Yan, Fanglong Yao, and Kun Fu. Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling. *IEEE Transactions on Image Processing*, 32:1474–1487, 2023.

[17] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354, 2020.

[18] Victor Klemas. Airborne remote sensing of coastal features and processes: An overview. *Journal of Coastal Research*, 29(2):239–255, 2013.

[19] James M Murphy and Mauro Maggioni. Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1829–1845, 2018.

[20] Simone Zini, Mirko Paolo Barbato, Flavio Piccoli, and Paolo Napoletano. Deep learning hyperspectral pansharpening on large scale prisma dataset, 2023.

[21] Mirko Paolo Barbato, Flavio Piccoli, and Paolo Napoletano. Ticino: A multi-modal remote sensing dataset for semantic segmentation. *Available at SSRN 4535928*, 2023.

[22] Veronica Grazia Morelli, Mirko Paolo Barbato, Flavio Piccoli, and Paolo Napoletano. Multimodal fusion methods with vision transformers for remote sensing semantic segmentation. In *2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2023 (in press).

[23] Flavio Piccoli, Mirko Paolo Barbato, Marco Peracchi, and Paolo Napoletano. Estimation of soil characteristics from multispectral sentinel-3 imagery and dem derivatives using machine learning. *Sensors*, 23(18):7876, 2023.

[24] Mirko Paolo Barbato, Flavio Piccoli, and Paolo Napoletano. Multimodal earth observation modeling using ai. In *International Conference on Modelling and Simulation for Autonomous Systems*. Springer, 2023 (in press).

[25] B Ayerdi M Graña, MA Veganzons. Hyperspectral Remote Sensing Scenes. `http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes`, 2020.

[26] Lin He, Dahan Xi, Jun Li, Honghao Lai, Antonio Plaza, and Jocelyn Chanussot. Dynamic hyperspectral pansharpening cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2023.

[27] Demetrio Labate, Kazem Safari, Nikolaos Karantzas, Saurabh Prasad, and Farideh Foroozandeh Shahraki. Structured receptive field networks and applications to hyperspectral image classification. In *Wavelets and Sparsity XVIII*, volume 11138, pages 218–226. SPIE, 2019.

[28] Naoto Yokoya and Akira Iwasaki. Airborne hyperspectral data over chikusei. *Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27*, 5, 2016.

[29] Laetitia Loncan, Luis B De Almeida, José M Bioucas-Dias, Xavier Briottet, Jocelyn Chanussot, Nicolas Dobigeon, Sophie Fabre, Wenzhi Liao, Giorgio A Licciardi, Miguel Simoes, et al. Hyperspectral pansharpening: A review. *IEEE Geoscience and remote sensing magazine*, 3(3):27–46, 2015.

[30] Marion F Baumgardner, Larry L Biehl, and David A Landgrebe. 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3. *Purdue University Research Repository*, 10(7):991, 2015.

[31] Lin He, Jiawei Zhu, Jun Li, Antonio Plaza, Jocelyn Chanussot, and Bo Li. Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(8):3092–3100, 2019.

[32] Lin He, Jiawei Zhu, Jun Li, Deyu Meng, Jocelyn Chanussot, and Antonio Plaza. Spectral-fidelity convolutional neural networks for hyperspectral pansharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:5898–5914, 2020.

[33] Pats Chavez, Stuart C Sides, Jeffrey A Anderson, et al. Comparison of three different methods to merge multiresolution and multispectral data- landsat tm and spot panchromatic. *Photogrammetric Engineering and remote sensing*, 57(3):295–303, 1991.

[34] Te-Ming Tu, Ping Sheng Huang, Chung-Ling Hung, and Chien-Ping Chang. A fast intensity-hue-saturation fusion technique with spectral adjustment for ikonos imagery. *IEEE Geoscience and Remote sensing letters*, 1(4):309–312, 2004.

[35] Craig A Laben and Bernard V Brower. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening, January 4 2000. US Patent 6,011,875.

[36] Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of ms +pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007.

[37] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[38] Guy P Nason and Bernard W Silverman. The stationary wavelet transform and some statistical applications. *Wavelets and statistics*, pages 281–299, 1995.

[39] Mark J Shensa et al. The discrete wavelet transform: wedding the a trous and mallat algorithms. *IEEE Transactions on signal processing*, 40(10):2464–2482, 1992.

[40] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987.

[41] Wenzhi Liao, Xin Huang, Frieke Van Coillie, Sidharta Gautama, Aleksandra Pižurica, Wilfried Philips, Hui Liu, Tingting Zhu, Michal Shimoni, Gabriele Moser, et al. Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2984–2996, 2015.

[42] Kai Zhang, Feng Zhang, Wenbo Wan, Hui Yu, Jiande Sun, Javier Del Ser, Eyad Elyan, and Amir Hussain. Panchromatic and multispectral image fusion for remote sensing and earth observation: Concepts, taxonomy, literature review, evaluation methodologies and challenges ahead. *Information Fusion*, 2023.

[43] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016.

[44] Giuseppe Scarpa, Sergio Vitale, and Davide Cozzolino. Target-adaptive cnn-based pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9):5443–5457, 2018.

[45] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017.

[46] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989, 2018.

[47] Xiangyu Liu, Qingjie Liu, and Yunhong Wang. Remote sensing image fusion based on two-stream fusion network. *Information Fusion*, 55:1–15, 2020.

[48] Jiajun Cai and Bo Huang. Super-resolution-guided progressive pansharpening based on a deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5206–5220, 2020.

[49] Yuchen Xie, Wei Wu, Haiping Yang, Ning Wu, and Ying Shen. Detail information prior net for remote sensing image pansharpening. *Remote Sensing*, 13(14):2800, 2021.

[50] Yuxuan Zheng, Jiaojiao Li, Yunsong Li, Kailang Cao, and Keyan Wang. Deep residual learning for boosting the accuracy of hyperspectral pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 17(8):1435–1439, 2019.

[51] Weiying Xie, Jie Lei, Yuhang Cui, Yunsong Li, and Qian Du. Hyperspectral pansharpening with deep priors. *IEEE transactions on neural networks and learning systems*, 31(5):1529–1543, 2019.

[52] Jianjun Liu, Zhiyong Xiao, Yufeng Chen, and Jinlong Yang. Spatial-spectral graph regularized kernel sparse representation for hyperspectral image classification. *ISPRS International Journal of Geo-Information*, 6(8):258, 2017.

[53] Shuzhen Zhang, Shutao Li, Wei Fu, and Leiyuan Fang. Multiscale superpixel-based sparse representation for hyperspectral image classification. *Remote Sensing*, 9(2):139, 2017.

[54] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Deep learning for classification of hyperspectral data: A comparative review. *IEEE geoscience and remote sensing magazine*, 7(2):159–173, 2019.

[55] Junfeng Wu, Zhiguo Jiang, Haopeng Zhang, Bowen Cai, and Quanmao Wei. Semi-supervised conditional random field for hyperspectral remote sensing image classification. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 2614–2617. IEEE, 2016.

[56] John E Vargas, Alexandre X Falcão, Jefersson A dos Santos, Júlio César Dalla Mora Esquerdo, Alexandre Camargo Coutinho, and JFG Antunes. Contextual superpixel description for remote sensing image classification. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1132–1135. IEEE, 2015.

[57] Nicolas Gillis, Da Kuang, and Haesun Park. Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4):2066–2078, 2014.

[58] Yuxiang Zhang, Kang Liu, Yanni Dong, Ke Wu, and Xiangyun Hu. Semisupervised classification based on slic segmentation for hyperspectral image. *IEEE Geoscience and Remote Sensing Letters*, 17(8):1440–1444, 2019.

[59] Juan Mario Haut, Mercedes E Paoletti, Javier Plaza, Antonio Plaza, and Jun Li. Visual attention-driven hyperspectral image classification. *IEEE transactions on geoscience and remote sensing*, 57(10):8065–8080, 2019.

[60] Haotian Zhang, Jing Yao, Li Ni, Lianru Gao, and Min Huang. Multimodal attention-aware convolutional neural networks for classification of hyperspectral and lidar data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022.

[61] Yongshan Zhang, Xinwei Jiang, Xinxin Wang, and Zhihua Cai. Spectral-spatial hyperspectral image classification with superpixel pattern and extreme learning machine. *Remote Sensing*, 11(17):1983, 2019.

[62] Baokai Zu, Kewen Xia, Tiejun Li, Ziping He, Yafang Li, Jingzhong Hou, and Wei Du. Slic superpixel-based l2, 1-norm robust principal component analysis for hyperspectral image classification. *Sensors*, 19(3):479, 2019.

[63] Sen Jia, Shuguo Jiang, Zhijie Lin, Nanying Li, Meng Xu, and Shiqi Yu. A survey: Deep learning for hyperspectral image classification with few labeled samples. *Neurocomputing*, 448:179–204, 2021.

[64] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international*

conference on artificial intelligence and statistics, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[65] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[66] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected topics in applied earth observations and remote sensing*, 7(6):2094–2107, 2014.

[67] Ghasem Abdi, Farhad Samadzadegan, and Peter Reinartz. Spectral–spatial feature learning for hyperspectral imagery classification using deep stacked sparse autoencoder. *Journal of Applied Remote Sensing*, 11(4):042604, 2017.

[68] Chen Xing, Li Ma, and Xiaoquan Yang. Stacked denoise autoencoder based feature extraction and classification for hyperspectral images. *Journal of Sensors*, 2016, 2016.

[69] Jakub Nalepa, Michal Myller, Yasuteru Imai, Ken-ichi Honda, Tomomi Takeda, and Marek Antoniak. Unsupervised segmentation of hyperspectral images using 3-d convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters*, 17(11):1948–1952, 2020.

[70] Mengmeng Zhang, Wei Li, Ran Tao, Hengchao Li, and Qian Du. Information fusion for classification of hyperspectral and lidar data using ip-cnn. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021.

[71] Arati Paul and Sanghamita Bhoumik. Classification of hyperspectral imagery using spectrally partitioned hyperunet. *Neural Computing and Applications*, 34(3):2073–2082, 2022.

[72] Lukasz Tulczyjew, Michal Kawulok, and Jakub Nalepa. Unsupervised feature learning using recurrent neural nets for segmenting hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 18(12):2142–2146, 2020.

[73] Jialong Chen, Yuebin Wang, Liqiang Zhang, Meiling Liu, and Antonio Plaza. Drflvat: Deep representative feature learning with virtual adversarial training for semi-supervised classification of hyperspectral image. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[74] Naftaly Wambugu, Yiping Chen, Zhenlong Xiao, Kun Tan, Mingqiang Wei, Xiaoxue Liu, and Jonathan Li. Hyperspectral image classification on insufficient-sample and

feature learning using deep neural networks: A review. *International Journal of Applied Earth Observation and Geoinformation*, 105:102603, 2021.

[75] Giacomo Aletti, Alessandro Benfenati, and Giovanni Naldi. A semi-supervised reduced-space method for hyperspectral imaging segmentation. *Journal of Imaging*, 7(12):267, 2021.

[76] Yao Ding, Xiaofeng Zhao, Zhili Zhang, Wei Cai, Nengjun Yang, and Ying Zhan. Semi-supervised locality preserving dense graph neural network with arma filters and context-aware learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021.

[77] Fulin Luo, Zehua Zou, Jiamin Liu, and Zhiping Lin. Dimensionality reduction and classification of hyperspectral image via multistructure unified discriminative embedding. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.

[78] Guangyun Zhang, Xiuping Jia, and Jiankun Hu. Superpixel-based graphical model for remote sensing image mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 53(11):5861–5871, 2015.

[79] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[80] Ahmad Obeid, Ibrahim M Elfadel, and Naoufel Werghi. Unsupervised land-cover segmentation using accelerated balanced deep embedded clustering. *IEEE Geoscience and Remote Sensing Letters*, 2021.

[81] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.

[82] J González Santiago, Fabian Schenkel, Wolfgang Gross, and Wolfgang Middelmann. An unsupervised labeling approach for hyperspectral image classification. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:407–415, 2020.

[83] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018.

[84] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. Torontocity: Seeing the world with a million eyes. *arXiv preprint arXiv:1612.00423*, 2016.

[85] Sharada Prasanna Mohanty, Jakub Czakon, Kamil A Kaczmarek, Andrzej Pyskir, Piotr Tarasiewicz, Saket Kunwar, Janick Rohrbach, Dave Luo, Manjunath Prasad, Sascha Fleer, et al. Deep learning for understanding satellite imagery: An experimental survey. *Frontiers in Artificial Intelligence*, 3, 2020.

[86] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE, 2017.

[87] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[88] Jefersson A dos Santos, Otávio AB Penatti, Philippe-Henri Gosselin, Alexandre X Falcaão, Sylvie Philipp-Foliguet, and Ricardo da S Torres. Efficient and effective hierarchical feature propagation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(12):4632–4643, 2014.

[89] Michele Volpi and Vittorio Ferrari. Semantic segmentation of urban scenes by learning local class interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2015.

[90] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417, 2021.

[91] Sumit Kumar Arora. Spacenet information. Medium, 2018.

[92] ISPRS 2D Semantic Labeling. Isprs, 2018.

[93] Dstl Satellite Imagery Feature Detection. Kaggle.

[94] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112:102926, 2022.

[95] 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*, pages 3431–3440, 2015.

[96] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[97] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[98] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[99] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[100] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[101] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[102] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[103] Bei Fang, Ying Li, Haokui Zhang, and Jonathan Cheung-Wai Chan. Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism. *Remote Sensing*, 11(2):159, 2019.

[104] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[105] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[106] Yongcheng Liu, Bin Fan, Lingfeng Wang, Jun Bai, Shiming Xiang, and Chunhong Pan. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS journal of photogrammetry and remote sensing*, 145:78–95, 2018.

[107] Dongcai Cheng, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Fusionnet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12):5769–5783, 2017.

[108] Dimitrios Marmanis, Jan D Wegner, Silvano Galliani, Konrad Schindler, Mihai Datcu, and Uwe Stilla. Semantic segmentation of aerial images with an ensemble of cnss. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, 3:473–480, 2016.

[109] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS journal of photogrammetry and remote sensing*, 140:20–32, 2018.

[110] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[111] Corentin Henry, Seyed Majid Azimi, and Nina Merkle. Road segmentation in sar satellite images with deep fully convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(12):1867–1871, 2018.

[112] Yongyang Xu, Liang Wu, Zhong Xie, and Zhanlong Chen. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sensing*, 10(1):144, 2018.

[113] Hongzhen Wang, Ying Wang, Qian Zhang, Shiming Xiang, and Chunhong Pan. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sensing*, 9(5):446, 2017.

[114] Guangming Wu, Xiaowei Shao, Zhiling Guo, Qi Chen, Wei Yuan, Xiaodan Shi, Yongwei Xu, and Ryosuke Shibasaki. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing*, 10(3):407, 2018.

[115] Ruirui Li, Wenjie Liu, Lei Yang, Shihao Sun, Wei Hu, Fan Zhang, and Wei Li. Deepunet: A deep fully convolutional network for pixel-level sea-land segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(11):3954–3962, 2018.

[116] Kaiqiang Chen, Kun Fu, Menglong Yan, Xin Gao, Xian Sun, and Xin Wei. Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(2):173–177, 2018.

[117] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[118] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[119] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[120] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[121] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[122] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[123] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[124] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

[125] Zhiyong Xu, Weicun Zhang, Tianxiang Zhang, Zhifang Yang, and Jiangyun Li. Efficient transformer for remote sensing image segmentation. *Remote Sensing*, 13(18):3585, 2021.

[126] Hong Wang, Xianzhong Chen, Tianxiang Zhang, Zhiyong Xu, and Jiangyun Li. Cctnet: Coupled cnn and transformer network for crop segmentation of remote sensing images. *Remote Sensing*, 14(9):1956, 2022.

[127] Liang Gao, Hui Liu, Minhang Yang, Long Chen, Yaling Wan, Zhengqing Xiao, and Yurong Qian. Stransfuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10990–11003, 2021.

[128] Cheng Zhang, Wanshou Jiang, Yuan Zhang, Wei Wang, Qing Zhao, and Chenjie Wang. Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022.

[129] Teerapong Panboonyuen, Kulsawasd Jitkajornwanich, Siam Lawawirojwong, Panu Srestasathiern, and Peerapon Vateekul. Transformer-based decoder designs for semantic segmentation on remotely sensed images. *Remote Sensing*, 13(24):5100, 2021.

[130] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.

[131] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021.

[132] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.

[133] Gemine Vivone, Andrea Garzelli, Yang Xu, Wenzhi Liao, and Jocelyn Chanussot. Panchromatic and hyperspectral image fusion: Outcome of the 2022 whispers hyperspectral pansharpening challenge. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:166–179, 2022.

[134] Yabei Li, Junge Zhang, Yanhua Cheng, Kaiqi Huang, and Tieniu Tan. Semantics-guided multi-level rgb-d feature fusion for indoor semantic segmentation. In *2017 IEEE international conference on image processing (ICIP)*, pages 1262–1266. IEEE, 2017.

[135] Abdulaziz Amer Aleissaee, Amandeep Kumar, Rao Muhammad Anwer, Salman Khan, Hisham Cholakkal, Gui-Song Xia, and Fahad Shahbaz Khan. Transformers in remote sensing: A survey. *Remote Sensing*, 15(7):1860, 2023.

[136] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[137] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 839–848, 2020.

[138] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3347–3356, 2022.

[139] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198*, 2020.

[140] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[141] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11782–11791, 2021.

[142] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12972–12980, 2021.

[143] Moslem Ladoni, H. Bahrami, Seyed Kazem Alavi Panah, and Ali Norouzi. Estimating soil organic carbon from soil reflectance: A review. *Precision Agriculture*, 11:82–99, 02 2009.

[144] Gerald Forkuor, Ozias K. L. Hounkpatin, Gerhard Welp, and Michael Thiel. High resolution mapping of soil properties using remote sensing variables in south-western burkina faso: A comparison of machine learning and multiple linear regression models. *PLOS ONE*, 12:1–21, 01 2017.

[145] José Safanelli, Sabine Chabrillat, Eyal Ben-Dor, and José Demattê. Multispectral models from bare soil composites for mapping topsoil properties over europe. *Remote Sensing*, 12:1369, 04 2020.

[146] Tao Zhou, Yajun Geng, Cheng Ji, Xiangrui Xu, Hong Wang, Jianjun Pan, Jan Bumberger, Dagmar Haase, and Angela Lausch. Prediction of soil organic carbon and the c:n ratio on a national scale using machine learning and satellite data: A comparison between sentinel-2, sentinel-3 and landsat-8 images. *Science of The Total Environment*, page 142661, 01 2021.

[147] Panos Panagos, Marc Van Liedekerke, Arwyn Jones, and Luca Montanarella. European soil data centre: Response to european policy support and public data requirements. *Land use policy*, 29(2):329–338, 2012.

[148] Jie Hu, Jie Peng, Yin Zhou, Dongyun Xu, Ruiying Zhao, Qingsong Jiang, Tingting Fu, Fei Wang, and Zhou Shi. Quantitative estimation of soil salinity using uav-borne hyperspectral and satellite multispectral images. *Remote Sensing*, 11(7):736, 2019.

[149] Long Guo, Haitao Zhang, Tiezhu Shi, Yiyun Chen, Qinghu Jiang, and M Linderman. Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images. *Geoderma*, 337:32–41, 2019.

[150] Xiangtian Meng, Yilin Bao, Jiangui Liu, Huanjun Liu, Xinle Zhang, Yu Zhang, Peng Wang, Haitao Tang, and Fanchang Kong. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. *International Journal of Applied Earth Observation and Geoinformation*, 89:102111, 07 2020.

[151] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[152] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[153] Olga Chambers et al. Machine learning strategy for soil nutrients prediction using spectroscopic method. *Sensors*, 21(12):4208, 2021.

[154] Ruixue Li, Bo Yin, Yanping Cong, and Zehua Du. Simultaneous prediction of soil properties using multi_cnn model. *Sensors*, 20(21):6271, 2020.

[155] Gergely Tóth, Arwyn Jones, and Luca Montanarella. The lucas topsoil database and derived information on the regional variability of cropland topsoil properties in the european union. *Environmental monitoring and assessment*, 185, 02 2013.

[156] Sreeram V Menon and Chandra Sekhar Seelamantula. Robust savitzky-golay filters. In *2014 19th International Conference on Digital Signal Processing*, pages 688–693. IEEE, 2014.

[157] Jozsef Dombi and Adrienn Dineva. Adaptive savitzky-golay filtering and its applications. *International Journal of Advanced Intelligence Paradigms*, 16(2):145–156, 2020.

[158] Kurt Heil and Urs Schmidhalter. An evaluation of different nir-spectral pretreatments to derive the soil parameters c and n of a humus-clay-rich soil. *Sensors*, 21(4):1423, 2021.

[159] Manuela Mancini, Giuseppe Toscano, and Åsmund Rinnan. Study of the scattering effects on nir data for the prediction of ash content using emsc correction factors. *Journal of Chemometrics*, 33(4):e3111, 2019.

[160] Space Transportation Costs. Trends in price per pound to orbit 1990-2000. *Futron Corporation, Bethesda, Maryland*, 2002.

[161] Harry Jones. The recent large reduction in space launch cost. 48th International Conference on Environmental Systems, 2018.

[162] Adam Okninski, Wioleta Kopacz, Damian Kaniewski, and Kamil Sobczak. Hybrid rocket propulsion technology for space transportation revisited-propellant solutions and challenges. *FirePhysChem*, 1(4):260–271, 2021.

[163] Michael A Wulder, Thomas R Loveland, David P Roy, Christopher J Crawford, Jeffrey G Masek, Curtis E Woodcock, Richard G Allen, Martha C Anderson, Alan S Belward, Warren B Cohen, et al. Current status of landsat program, science, and applications. *Remote sensing of environment*, 225:127–147, 2019.

[164] Michele Chevrel, MICHEL Courtois, and G Weill. The spot satellite remote sensing mission. *Photogrammetric Engineering and Remote Sensing*, 47:1163–1171, 1981.

[165] Darius Phiri, Matamyo Simwanda, Serajis Salekin, Vincent R Nyirenda, Yuji Murayama, and Manjula Ranagalage. Sentinel-2 data for land cover/use mapping: A review. *Remote Sensing*, 12(14):2291, 2020.

[166] Dionysios N Apostolopoulos and Konstantinos G Nikolakopoulos. Spot vs landsat satellite images for the evolution of the north peloponnese coastline, greece. *Regional Studies in Marine Science*, 56:102691, 2022.

[167] Jared Keith Krueger. *CLOSeSat: Perigee-lowering techniques and preliminary design for a small optical imaging satellite operating in very low earth orbit.* PhD thesis, Massachusetts Institute of Technology, 2010.

[168] Shutao Li, Xudong Kang, Leyuan Fang, Jianwen Hu, and Haitao Yin. Pixel-level image fusion: A survey of the state of the art. *information Fusion*, 33:100–112, 2017.

[169] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2014.

[170] Agenzia Spaziale Italiana ASI. Prisma algorithm theoretical basis document (atbd), 2021. Accessed on 3 April, 2023.

[171] Daniel Scheffler, André Hollstein, Hannes Diedrich, Karl Segl, and Patrick Hostert. Arosics: An automated and robust open-source image co-registration software for multi-sensor satellite data. *Remote sensing*, 9(7):676, 2017.

[172] Lucien Wald. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions.* Presses des MINES, 2002.

[173] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992.

[174] Jie Zhou, Daniel L Civco, and John A Silander. A wavelet transform method to merge landsat tm and spot panchromatic data. *International journal of remote sensing*, 19(4):743–757, 1998.

[175] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002.

[176] Alberto Arienzo, Gemine Vivone, Andrea Garzelli, Luciano Alparone, and Jocelyn Chanussot. Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):168–201, 2022.

[177] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6):691–699, 1997.

[178] Luciano Alparone, Andrea Garzelli, and Gemine Vivone. Spatial consistency for full-scale assessment of pansharpening. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 5132–5134. IEEE, 2018.

[179] Miguel Simoes, José Bioucas-Dias, Luis B Almeida, and Jocelyn Chanussot. A convex formulation for hyperspectral image superresolution via subspace-based regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3373–3388, 2014.

[180] D Fitton, E Laurens, N Hongkarnjanakul, C Schwob, and L Mezeix. Land cover classification through convolutional neur-al network model assembly: A case study of a local rural area in thailand. *Remote Sensing Applications: Society and Environment*, 26:100740, 2022.

[181] Tahereh Bahraini, Peyman Azimpour, and Hadi Sadoghi Yazdi. Modified-mean-shift-based noisy label detection for hyperspectral image classification. *Computers & Geosciences*, page 104843, 2021.

[182] Hugo Costa, Giles M Foody, and Doreen S Boyd. Supervised methods of image segmentation accuracy assessment in land cover mapping. *Remote Sensing of Environment*, 205:338–351, 2018.

[183] Ioannis Kotaridis and Maria Lazaridou. Remote sensing image segmentation advances: A meta-analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:309–322, 2021.

[184] Fabricio Breve. Interactive image segmentation using label propagation through complex networks. *Expert Systems With Applications*, 123:18–33, 2019.

[185] Yilang Shen, Tinghua Ai, Wende Li, Min Yang, and Yu Feng. A polygon aggregation method with global feature preservation using superpixel segmentation. *Computers, Environment and Urban Systems*, 75:117–131, 2019.

[186] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

[187] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[188] Yongji Wang, Qingwen Qi, Ying Liu, Lili Jiang, and Jun Wang. Unsupervised segmentation parameter selection using the local spatial statistics for remote sensing image segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 81:98–109, 2019.

[189] Bart Finkston. Mean shift clustering. `https://www.mathworks.com/matlabcentral/fileexchange/10161-mean-shift-clustering`, 2022. [Online; accessed February 16, 2022].

[190] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[191] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[192] Taoran Sheng and Manfred Huber. Unsupervised embedding learning for human activity recognition using wearable sensor data. In *The Thirty-Third International Flairs Conference*, 2020.

[193] Vijini Mallawaarachchi, Anuradha Wickramarachchi, and Yu Lin. GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics*, 36(11):3307–3313, 03 2020.

[194] Jakub Nalepa, Michal Myller, Marcin Cwiek, Lukasz Zak, Tomasz Lakota, Lukasz Tulczyjew, and Michal Kawulok. Towards on-board hyperspectral satellite image segmentation: Understanding robustness of deep learning through simulating acquisition conditions. *Remote Sensing*, 13(8):1532, 2021.

[195] Colin Clark. *Population growth and land use.* Springer, 1977.

[196] Pamela Tendaupenyu, Christopher Hilary Dennis Magadza, and Amon Murwira. Changes in landuse/landcover patterns and human population growth in the lake chivero catchment, zimbabwe. *Geocarto International*, 32(7):797–811, 2017.

[197] Antonie Veldkamp and Eric F Lambin. Predicting land-use change, 2001.

[198] Flavio Piccoli, Simone Giuseppe Locatelli, Raimondo Schettini, and Paolo Napoletano. An open-source platform for gis data management and analytics. *Sensors*, 23(8):3788, 2023.

[199] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019.

[200] Clément Dechesne, Clément Mallet, Arnaud Le Bris, and Valérie Gouet-Brunet. Semantic segmentation of forest stands of pure species as a global optimization problem. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:141, 2017.

[201] Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782, 2017.

[202] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, 1(1):293–298, 2012.

[203] Jagannath K Jadhav and RP Singh. Automatic semantic segmentation and classification of remote sensing data for agriculture. *Mathematical Models in Engineering*, 4(2):112–137, 2018.

[204] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[205] Qibin He, Xian Sun, Wenhui Diao, Zhiyuan Yan, Fanglong Yao, and Kun Fu. Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling. *IEEE Transactions on Image Processing*, 32:1474–1487, 2023.

[206] Microsoft Bing Maps. MBM. Microsoft Bing Maps.

[207] Geoportal of Lombardy Region. regione.lombardia.it.

[208] OpenStreetMap. OSM.

[209] Consultazione cartografia catastale. Agenzia delle Entrate, 2015.

[210] Carta dell'uso agricolo dati. SIARL.

[211] Bruno Aiazzi, Stefano Baronti, and Massimo Selva. Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3230–3239, 2007.

[212] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[213] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Int. conference on machine learning*, pages 10347–10357. PMLR, 2021.

[214] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proc. of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[215] David Hoffmann, Kai Norman Clasen, and Begüm Demir. Transformer-based multimodal learning for multi label remote sensing image classification. *arXiv preprint arXiv:2306.01523*, 2023.

[216] Philipp Dufter, Martin Schmitt, and Hinrich Schütze. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763, 2022.

[217] C. Ritter, D. Dicke, M. Weis, Horst Oebel, Hans-Peter Piepho, A. Büchse, and Roland Gerhards. An on-farm approach to quantify yield variation and to derive decision rules for site-specific weed management. *Precision Agriculture*, 9:133–146, 06 2008.

[218] Richard P. Dick. A review: long-term effects of agricultural systems on soil biochemical and microbial parameters. *Agriculture, Ecosystems & Environment*, 40(1):25–36, 1992. Biotic Diversity in Agroecosystems.

[219] Theodora Angelopoulou, Nikolaos Tziolas, Athanasios Balafoutis, G. Zalidis, and Dionysis Bochtis. Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sensing*, 11:676, 03 2019.

[220] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014.

[221] EM Barnes and MG Baker. Multispectral data for mapping soil texture: possibilities and limitations. *Applied Engineering in Agriculture*, 16(6):731–741, 2000.

[222] Philippe Lagacherie, Frédéric Baret, Jean-Baptiste Feret, José Madeira Netto, and Jean Marc Robbez-Masson. Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. *Remote Sensing of Environment*, 112(3):825–835, 2008.

[223] Gilson Augusto Helfer, Jorge Luis Victória Barbosa, Douglas Alves, Adilson Ben da Costa, Marko Beko, and Valderi Reis Quietinho Leithardt. Multispectral cameras and machine learning integrated into portable devices as clay prediction technology. *Journal of Sensor and Actuator Networks*, 10(3):40, 2021.

[224] Flavio Piccoli, Micol Rossini, Roberto Colombo, Raimondo Schettini, and Paolo Napoletano. A deep scalable neural architecture for soil properties estimation from spectral information. *Computers & Geosciences*, page 105433, 2023.

[225] Omosalewa Odebiri, Onisimo Mutanga, and John Odindi. Deep learning-based national scale soil organic carbon mapping with sentinel-3 data. *Geoderma*, 411:115695, 2022.

[226] Nyle C Brady, Ray R Weil, and Ray R Weil. *The nature and properties of soils*, volume 13. Prentice Hall Upper Saddle River, NJ, 2008.

[227] James B Campbell and Randolph H Wynne. *Introduction to remote sensing*. Guilford press, 2011.

[228] Copernicus - DEM. dem description.

[229] Hamza Keskin, Sabine Grunwald, and Willie G Harris. Digital mapping of soil carbon fractions with machine learning. *Geoderma*, 339:40–58, 2019.

[230] Tae-Hwy Lee, Aman Ullah, and Ran Wang. Bootstrap aggregating and random forest. *Macroeconomic forecasting in the era of big data: Theory and practice*, pages 389–429, 2020.

[231] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[232] Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424, 2021.

[233] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems*, 26:431–439, 2013.

[234] Copernicus. scihub-copernicus.

[235] NetCDF Format. format-description.

[236] Semi-Automatic Classification. qgis-plugin.

[237] Pat S Chavez et al. Image-based atmospheric corrections-revisited and improved. *Photogrammetric engineering and remote sensing*, 62(9):1025–1035, 1996.

[238] Cristiano Ballabio, Panos Panagos, and Luca Monatanarella. Mapping topsoil physical properties at european scale using the lucas database. *Geoderma*, 261:110–123, 2016.

[239] James A Thompson, Jay C Bell, and Charles A Butler. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma*, 100(1-2):67–89, 2001.

[240] SAGA-GIS. SAGA-GIS-Software.

[241] M.R. Travis, G.H. Elsner, W.D. Iverson, and C.G. Johnson. Viewit: computation of seen areas, slope, and aspect for land-use planning. *USDA F.S. Gen. Tech. Rep. PSW-11/1975*, 1975.

[242] Olaf Conrad, Benjamin Bechtel, Michael Bock, Helge Dietrich, Elke Fischer, Lars Gerlitz, Jan Wehberg, Volker Wichmann, and Jürgen Böhner. System for automated geoscientific analyses (saga) v. 2.1. 4. *Geoscientific Model Development*, 8(7):1991–2007, 2015.

[243] Insun Song, PooGyeon Park, and Robert W Newcomb. A normalized least mean squares algorithm with a step-size scaler against impulsive measurement noise. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 60(7):442–445, 2013.

[244] Zhengyong Zhao, Thien Lien Chow, Herb W Rees, Qi Yang, Zisheng Xing, and Fan-Rui Meng. Predict soil texture distributions using an artificial neural network model. *Computers and electronics in agriculture*, 65(1):36–48, 2009.

[245] Paul Horswill, Odhran O'Sullivan, Gareth K Phoenix, John A Lee, and Jonathan R Leake. Base cation depletion, eutrophication and acidification of species-rich grasslands in response to long-term simulated nitrogen deposition. *Environmental Pollution*, 155(2):336–349, 2008.

[246] DR Allen. Identification of sediments-their depositional environment and degree of compaction—from well logs. 1975.