Tesi di Dottorato

Università degli Studi di Napoli "Federico II"

Dipartimento di Ingegneria Elettronica
e delle Telecomunicazioni

Dottorato di Ricerca in
Ingegneria Elettronica e delle Telecomunicazioni

A Deep Learning Framework in Selected Remote Sensing Applications

---

Massimiliano Gargiulo

Il Coordinatore del Corso di Dottorato
Ch.mo Prof. Daniele Riccio

Il Tutore
Ch.mo Prof. Giuseppe Ruello

A. A. 2020–2021

Si te veco: me veco.
Si mme vire: te vire.
Si tu parle, c'è l'eco
e chist'eco song'i.
Si te muove: me movo.
Si te sento: me sento.
Si me truove, te trovo...
Si me trovo, si tu!
(Eduardo De Filippo)

to all my family

Preface

This thesis presents the results of my research activities carried out within the PhD degree in Information Technology and Electrical Engineering at the University of Naples "Federico II" from January 2018 until April 2021. The main research topic is designing and implementing a deep learning framework applied to remote sensing. Remote sensing techniques and applications play a crucial role in observing the Earth evolution, especially nowadays, where the effects of climate change on our life is more and more evident. A considerable amount of data are daily acquired all over the Earth. Effective exploitation of this information requires the robustness, velocity and accuracy of deep learning. This emerging need inspired the choice of this topic.

The conducted studies mainly focus on two European Space Agency (ESA) missions: Sentinel 1 and Sentinel 2. Images provided by the ESA Sentinel-2 mission are rapidly becoming the main source of information for the entire remote sensing community, thanks to their unprecedented combination of spatial, spectral and temporal resolution, as well as their open access policy. The increasing interest gained by these satellites in the research laboratory and applicative scenarios pushed us to utilize them in the considered framework. The combined use of Sentinel 1 and Sentinel 2 is crucial and very prominent in different contexts and different kinds of monitoring when the growing (or changing) dynamics are very rapid.

Starting from this general framework, two specific research activities were identified and investigated, leading to the results presented in this dissertation. Both these studies can be placed in the context of data fusion.

The first activity deals with a super-resolution framework to improve Sentinel 2 bands supplied at 20 meters up to 10 meters. Increasing the spatial resolution of these bands is of great interest in many remote sensing applications, particularly in monitoring vegetation, rivers, forests, and so on.

The second topic of the deep learning framework has been applied to the multispectral Normalized Difference Vegetation Index (NDVI) extraction, and the semantic segmentation obtained fusing Sentinel 1 and S2 data. The S1 SAR data is of great importance for the quantity of information extracted in the context of monitoring wetlands, rivers and forests, and many other contexts.

In both cases, the problem was addressed with deep learning techniques, and in both cases, very lean architectures were used, demonstrating that even without the availability of computing power, it is possible to obtain high-level results. The core of this framework is a Convolutional Neural Network (CNN). CNNs have been successfully applied to many image processing problems, like super-resolution [1], pansharpening [2], classification [3], and others, because of several advantages such as (i) the capability to approximate complex non-linear functions, (ii) the ease of training that allows to avoid time-consuming handcraft filter design, (iii) the parallel computational architecture. Even if a large amount of "labelled" data is required for training, the CNN performances pushed me to this architectural choice. In our S1 and S2 integration task, we have faced and overcome the problem of manually labelled data with an approach based on integrating these two different sensors. Therefore, apart from the investigation in Sentinel-1 and Sentinel-2 integration, the main contribution in both cases of these works is, in particular, the possibility of

designing a CNN-based solution that can be distinguished by its lightness from a computational point of view and consequent substantial saving of time compared to more complex deep learning state-of-the-art solutions.

The reminder of the manuscript is organized as follows. In the first Chapter, the fundamental concepts of remote sensing are recalled. In the successive Chapter, key information about machine learning and deep learning design and performances is introduced. This study has the objective of processing and interpreting the images using deep learning-based methods. In the third Chapter, we describe a specific deep learning algorithm for super-resolution, started first as a simple CNN-based solution, then evolved into a refined version thanks to innovative improvements. The presented approach evidences the importance of the deep learning approaches in the super-resolution of Sentinel-2 bands. Further, in Chapter 4, the DL framework is applied to a data fusion problem between Sentinel-1 and Sentinel-2. Also, in this case, the proposed deep learning approach is described step-by-step, starting from the fundamental solution in which a shallow CNN is used in NDVI estimation and evolving toward deep learning approaches that can extract accurate information from Sentinel-1 data benefiting from the helpful information of Sentinel-2. In Chapter 5, we present the results of the proposed framework in selected applications. In particular, the super-resolution has been tested in the context of small fires (tens-hundreds of meters) and the segmentation in the monitoring of wetlands. Eventually, we further discuss the possibility and importance of deep learning strategies in different remote sensing contexts.

Massimiliano Gargiulo
University of Naples "Federico II"
April 2021

This page intentionally left blank.

Contents

List of Tables

This page intentionally left blank.

List of Figures

Chapter 1

Remote Sensing Principles

Remote Sensing can be considered the set of techniques, tools, and interpretative means that allow you to extend the human eye's perceptive capabilities by providing the observer with qualitative and quantitative information on objects placed at a distance. Its beginning is historically linked to the birth and development of photographic technique, which allowed both to permanently record observations and extend the possibilities of perception of an individual.

Specifically, Remote Sensing's beginning can be considered in 1840 when the balloons acquired the first images of an extended area with the newly invented photo camera. At the end of the last century, the newest platform was the renowned pigeon fleet operating as a novelty in Europe. Aerial photography became a recognized tool during World War I and filled during the Second. The official entry of sensors into space began with an automatic camera aboard German V-2 missiles launched from White Sands. Although Remote Sensing's beginning can link to the development of the traditional analog photographic technique, today, the sensors used in this context are mainly mounted on a satellite (generic satellites in orbit around the Earth in Fig. 1.1), or a set of satellites that compose a constel-

lation. In recent years, the RS also takes advantage of acquiring images possibility thanks to the use of drones or again aircraft. Sensors that acquired black and white images of Earth were mounted on meteorological satellites starting in the 1960s. Since the Remote Sensing techniques investigate the characteristics of surfaces placed at a distance from the observer by exploiting the outgoing electromagnetic energy (emitted + reflected) measured by special instruments as a vehicle for transporting information, then the RS reached a subsequent increasing interest, with operating systems for the acquisition of images on the Earth with a certain periodicity, in 1970 with instruments onboard the Skylab (and later the Space Shuttle) and on Landsat. Landsat was the first satellite specifically dedicated to monitoring lands and oceans to map cultural and natural resources. The first non-military radar system was the Shuttle Imaging Radar mounted by JPL aboard the Space Shuttle in 1982. Other nations then developed similar sensors or with different capabilities. Since 1980, Landsat has been privatized, and in several countries, including France, the United States, Russia, and Japan, a wider and more commercial use of RS has begun. To understand the importance of the RS imagery, I show the first Landsat image that sparked the strong interest in satellites for Earth monitoring

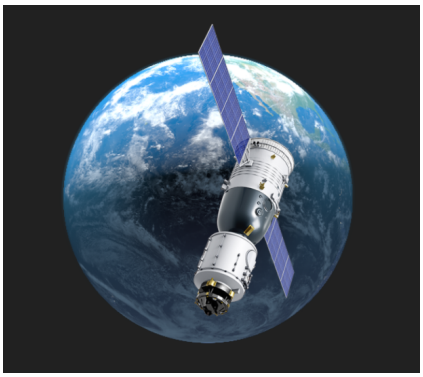
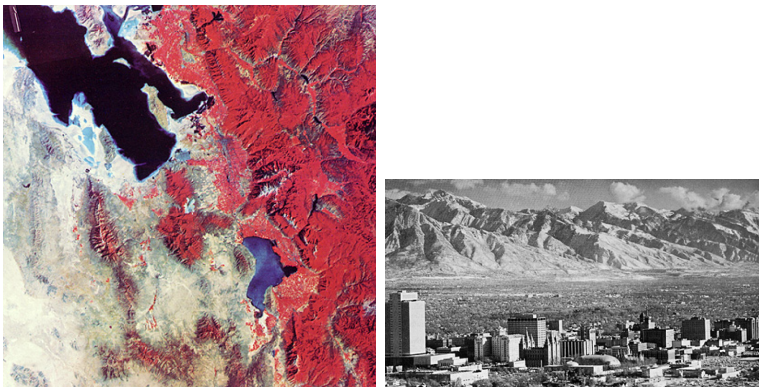


Figure 1.1. Satellites in orbit around the Earth.



(a) Landsat 1 Image **(b)** Picture of the same area

Figure 1.2. (a) The first image acquired by Landsat-1 15 days after its launch (August 7th, 1972), and (b) A picture of the same area (North of the Utah, United States of America).

1.2a. The scene below is from Landsat 1 and depicts North Central Utah taken 15 days after launch (August 7, 1972).

This kind of image is rich in information. In particular, in this case, it is used a near-infrared band instead of red, and in this false-color composition, it is possible to associate the bright red with healthy vegetation (dense forests or grassy lands) reflects a lot.To highlight the richness and importance of the satellite imagery, another point of view of the same scene acquired from the ground are reported in Fig. 1.2b. The picture in Fig. 1.2b does not allow us in-depth analysis. For example, a large-scale classification could not be done.

1.1 The Remote Sensing Process

Remote Sensing is a process that exploits the different ways in which electromagnetic energy from a source interacts with natural surfaces (1) in order to obtain information on their characteristics. The electromagnetic wave interacts with the atmosphere (2) before reaching the target.

The interaction between the incident energy and the natural surface (3) allows determining a particular surface with a unique "trace". This trace is called a spectral signature (4). The satellites or planes acquire these measurements the different spectral responses from the sensors mounted on board, and these measurements are collected (5). The processing of the data collected by the different types of possible sensors allows us to obtain information on the state of health or the surfaces' characteristics under investigation. This process's ultimate goal is generally the production of maps that provide helpful information for studying and managing the environment (6).

1. *Energy source* is the first requirement of Remote Sensing in order to illuminate the portion of the territory ("scene") that you want to study. The types of energy source can therefore be natural (Sun and Earth) or artificial, as in the case of radar;

2. the electromagnetic wave *interacts with the atmosphere*, composed of water vapor, gas, and suspended dust, therefore this interaction can modify the radiation, specifically generating refraction and/or absorption phenomena. This interaction is crucial in order to interpret the final data;

3. once it reaches the natural surfaces, the electromagnetic energy is *absorbed, re-emitted or reflected* depending on the physical characteristics of the surface, its conditions, and other factors;

4. the *spectral signature of surfaces* is a sort of imprint (obtained from the interaction between the incident energy and the surface) which is called the spectral signature, as mentioned, and allows us to distinguish between bodies or surfaces that show themselves different along the frequencies observed;

5. specific sensors measure and record the energy reflected or emitted

from the surface. The sensors are located on platforms always placed at a distance from the studied object. In the case of radar, the sensor is also a source of energy by emitting microwaves;

6. The *data* recorded by the sensors display as an image. *Thematic maps* are the final product of this process that various users can use for different purposes: from the environmental analysis necessary for proper territorial planning to the valuable cartographic reproduction for the preparation of atlases and tourist maps.

These aspects will be investigated in more detail in the following sections.

1.2 Electromagnetic Radiation

The observable measured in Remote Sensing (the vehicle of information) is the electromagnetic energy, which represents the link between the remote sensor and the phenomenon under investigation; the variations in the characteristics of the electromagnetic radiation become the source of a large amount of data, which allow to interpret and obtain important information on the different aspects of the investigated phenomenon. According to the wave theory, radiant energy can be described as a harmonic wave propagating in space and consisting of two orthogonal fields: the electric field **E**, whose amplitude varies continuously according to the direction of propagation of the electromagnetic wave and the magnetic field **M**.

The propagation speed of the wave in the air is constant and equal to the product of wavelength and frequency and is approximate: $c = 300,000 km/s^2$; therefore, both fields **E** and **M** also travel at the speed c, known as the speed of light. Furthermore, according to Planck's quantum hypothesis, electromagnetic energy is not uniformly distributed along the spectrum but propagates by finite quantities, called photons or energy

quanta, that, according to the corpuscular theory, are massless energy particles traveling at the speed of light. The energy carried by a photon is inversely proportional to its wavelength: the longer the wavelength, the lower its energy content.

1.2.1 Electromagnetic Spectrum

The electromagnetic spectrum covers many decades in frequency (or wavelength) from near 0 Hz to 300 GHz. The radio spectrum is used in many different applications, such as navigation, broadcasting, radar, and passive Remote Sensing. The electromagnetic spectrum represents the continuous distribution of electromagnetic energy ordered by increasing wavelengths λ, originating at tiny wavelengths. For practical and operational reasons, it is divided into different intervals called spectral bands. The spectral bands (the wavelengths of electromagnetic radiation) are included in the atmospheric windows, corresponding to the portions of the spectrum in which the atmosphere is transparent for the particular wavelength considered, a reason for discontinuity in the spectral bands. Sensors are generally designed to avoid the portions of the spectrum in which the atmosphere is opaque. Considering the distribution of energy as a function of wavelength, the different mechanisms of interaction with matter, and the characteristics of atmospheric transparency, the spectrum useful for Remote Sensing systems can be conceptually divided into two main windows: the optical *(visible + infra red)* window and the radar window.

The optical window is between $100nm$ and $20\mu m$ wavelength, and the energy studied is reflected or naturally emitted by surfaces. On a practical - operational level, the optical window consists of three main bands, each subdivided into several sub-bands, as shown in Fig. 1.1. In the radar window, the energy is transmitted and received through a specific antenna. It is divided into the bands, shown in Tab. 1.2, comprising the microwave band extending between 0.3 GHz and 300 GHz (or, equivalently, between

Table 1.1. Main bands and their sub-bands of the optical window

Band	Band	Wavelength λ	Range $\Delta\lambda$	Frequency ν
Ultraviolet UV	Ultraviolet	100 - 380 nm	280 nm	3000 - 789 THz
Visible VIS	Violet	380 - 430 nm	50 nm	789 - 697 THz
	Blue	430 - 475 nm	45 nm	697 - 631 THz
	Blue-Green	475 - 490 nm	15 nm	631 - 612 THz
	Green	490 - 550 nm	60 nm	612 - 545 THz
	Yellow	550 - 580 nm	30 nm	545 - 517 THz
	Orange	580 - 620 nm	40 nm	517 - 484 THz
	Red	620 - 750 nm	130 nm	484 - 400 THz
Infrared IR	Near Infrared	0.75 - 0.9 μm	0.15 μm	400 - 333 THz
		0.9 - 1.3 μm	0.4 μm	333 - 231 THz
	Short Wave Infrared	1.5 - 2.5 μm	1.0 μm	200 - 120 THz
	Mid Wave Infrared	3.5 - 5.2 μm	1.7 μm	86 - 58 THz
	Thermal Infrared	7.0 - 20.0 μm	13 μm	43 - 15 THz

1 m and 1 mm in wavelength).

Regarding the energy involved, the optical window can be distinguished into two main parts:

- from 0.1 μm to 3 μm: this is the area of the spectrum most suitable for investigating the spectral behaviour of surfaces using reflected solar energy; the most important and critical coefficients in this part of the spectrum are spectral reflectivity and spectral transmissivity;

- From 3 μm to 20 μm: the effects of the energy spontaneously emitted by the surfaces at room temperature, i.e. around 300 K, are predominant; it is the domain of the spectral emissivity coefficient that strongly conditions the interpretation of the data in terms of surface temperature since it allows to pass from the measured data to the actual temperature of the different surfaces.

The emissivity coefficient expresses how a given surface emits electromagnetic energy concerning the ideal blackbody model. Any external surface of a body, if at a temperature above absolute zero (zero on the Kelvin scale equal to -273.14C), emits its electromagnetic radiations,

which depend on the temperature of the body and the physical-chemical-geometric characteristics of its surface and reflects, absorbs or lets itself crossed by electromagnetic radiations coming from the outside according to the laws enunciated by Planck and Stefan-Boltzmann. The law formulated by Planck showed (see Fig. 1.3) that the spectral radiance of a body for a specific frequency f at absolute temperature T is given by:

$$R(\lambda) = \frac{2 \cdot \pi \cdot h \cdot f^3}{c^2} \frac{1}{e^{\frac{h \cdot f}{\lambda \cdot k \cdot T}} - 1} \tag{1.1}$$

where k is the Boltzmann constant, h the Planck constant and c the speed of light in the vacuum. Further, Stefan and Boltzmann experimentally verified that the maximum value of spectral radiance (emitted by the ideal blackbody) is proportional to the fourth power of the absolute temperature:

$$R = \sigma \cdot T^4 \tag{1.2}$$

where σ is the Boltzmann-Stefan constant. The spectral emissivity coefficient regulates this phenomenon. Therefore, the value of emissivity de-

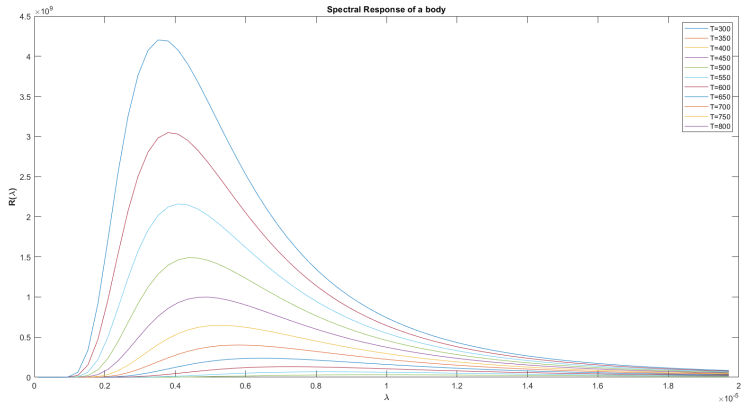


Figure 1.3. Planck's curve for given Temperature

pends on the particular wavelength considered and the higher the body's temperature, the lower the wavelength at which electromagnetic emission is maximum (Wien's law). The mathematical formulation of Wien's law, able to determine the wavelength corresponding to the maximum emission, is the following:

$$\lambda_{MAX} = \frac{2898}{T} \tag{1.3}$$

T is the absolute temperature expressed in Kelvin, and the corresponding λ_{MAX} is expressed in μm.

For example, for the Sun, which has an apparent black body temperature of about 6000 K, the maximum emission is around 0.5 micron (yellow-green light), while bodies at room temperature (around 300 K) peak at around 10 microns. What has been said follows that the criteria for selecting sensors and bands of the electromagnetic spectrum, each time more appropriate in Remote Sensing applications, depend on the temperature of the surfaces under consideration. As to the interaction between a surface and the electromagnetic energy incident on it, it should be remembered that the incident energy is partly reflected, partly absorbed and partly transmitted as a function of the reflection (r), absorption (a) and transmission (t) parameters, respectively, which depend on the physical nature of the surfaces, on their degree of roughness and the wavelength considered. Considering the ratios of the three radiant fluxes, for the incident radiant flux, the following three coefficients are obtained:

- reflection coefficient or reflectivity, = r / i;
- absorption coefficient or absorptivity, a / i;
- transmission coefficient or transmissivity, t / i.

The whole process of interaction with matter obeys the principle of conservation of energy expressed by Kirchhoff's law (r + a + t = 1), and the three coefficients are linked to each other by this equation. The Remote

Sensing principle is based on the ability to differentiate the largest possible number of elements on the territory (soil, vegetation, water, urbanized etc.) thanks to their different radiometric behaviour, that is, trying to analyze their characteristics in the electromagnetic spectrum at the different wavelengths, as shown in Fig. 1.4. The spectral signature describes

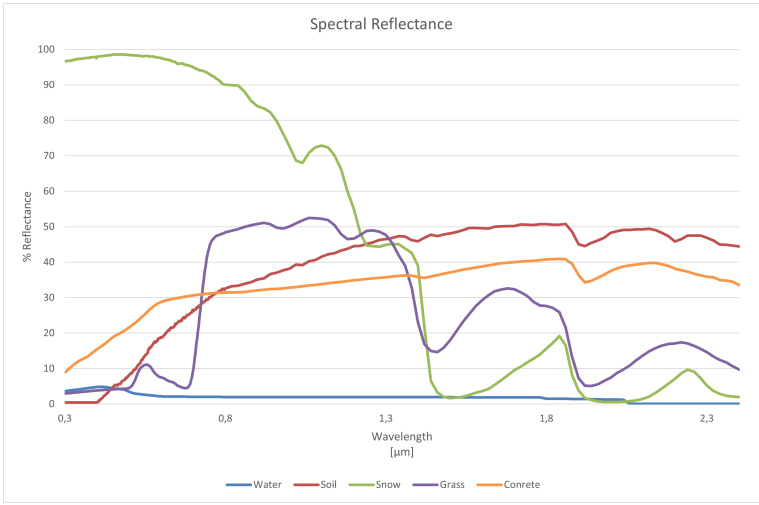


Figure 1.4. Spectral signatures of grass vegetation, water, conrete, snow and soil

the different radiometric behaviour of surfaces in the electromagnetic spectrum. The vegetation reflects energy, in the green and infrared bands, and absorbs energy in the blue and red bands. The behaviour then differs depending on the type of vegetation, the densities, the phenological state, the phytosanitary status, moisture content etc. The reflectance of bare soils depends on their physical (texture, structure, moisture content, etc.) and chemical (organic matter content, etc.) composition. The spectral response of water varies with the wavelength variation depending on the nature of the water and the conditions in which it is found. Its simple identification is easily possible in the near-infrared (0.7-1.2 mm) because it absorbs all the incident energy, while the study of its conditions is car-

ried out in the visible bands. The values assumed by these coefficients (r, a, and t) depend fundamentally on the wavelength taken into consideration, also varying according to the chemical-physical nature and roughness characteristics of the surface under examination. In fact, the proportions of reflected, absorbed and transmitted energy vary according to the type of material constituting the surface: with the same surface, then, the quantities of reflected, absorbed and transmitted energy vary according to the wavelength. This implies that two surfaces may be spectrally indistinguishable in a spectral window and perfectly separable at another. The majority of surfaces are opaque for large regions of the electromagnetic spectrum or are characterised by almost negligible transmissivity values. In conditions of an opaque body, it is possible to reduce the relation expressed by Kirchhoff's law to the two components of reflection and absorption only, thus obtaining: $r + a = 1$. Each region of the electromagnetic spectrum plays an important role in characterising surface properties, and the used methodology depends on the wavelength domain of the radiation. In particular:

- ultraviolet, visible and near-infrared use spectral reflectance and the sensor onboard a satellite, or an aircraft in general, measures the energy spectrum reflected from the scene under examination;
- thermal infrared uses the energy emitted by the Earth and the sensor directly detects the natural emission of objects located on the Earth's surface;
- Microwaves rely both on energy naturally emitted by the Earth's surface and on energy reflected and originally produced by an artificial instrument; a radar emits electromagnetic radiation, and then a sensor detects the fraction of this that is reflected by objects on the Earth's surface.

Regardless of the objects on the Earth's surface and the nature of the transmission, the atmosphere can absorb, reflect, or transmit electromagnetic radiation in different ways and is considered a disturbance when it is not the object of the detection. From the target to the sensor, the electromagnetic energy is affected by the presence in the atmosphere of solid and liquid particles, gases and aerosols, which influence the radiance value.

The interaction between electromagnetic radiation and the atmosphere consists of two main mechanisms that act on solar radiation: absorption, which reduces the amount of energy reaching the Earth, and scattering, which redistributes the radiant energy in space, changing the direction of propagation. These two phenomena's overall effect is to reduce the incident radiant flux, producing an overall decrease in atmospheric transparency. The absorption phenomenon makes the atmosphere opaque, i.e. presenting shallow transmissivity values in certain electromagnetic spectrum intervals.

From these physical observations, the choice of the region of the electromagnetic spectrum depends on the percentage of energy that is transmitted through the atmosphere, specifically the atmospheric transmissivity, which is the inverse of atmospheric opacity. The atmospheric transmissivity varies according to the electromagnetic frequency and the entire microwave band (300 MHz-300 GHz). Furthermore, for all frequencies below about 15 MHz, the ionospheric conditions determine the opacity related to the transmission of electromagnetic waves, so it is clear that the opacity and the atmospheric transmissivity are fundamental for the determination of the most suitable frequencies for active and passive Remote Sensing. Consequently, the choice of usable wavelengths is limited to particular domains such as: the visible ($\lambda = [0.4, 0.7]\mu m$), the near-infrared ($\lambda = [0.7, 1.3]\mu m; [1.5, 1.8]\mu m; [2, 2.5]\mu m$), the mid-infrared ($\lambda = [3, 5]\mu m$), the thermal infrared ($\lambda = [8, 14]\mu m$) and the microwave domain ($\lambda = [1mm, 1m]$). These domains represent the windows of transparency of the Earth's atmosphere to electromagnetic radiation. In partic-

ular, among these domains, the microwave spectrum can assure successful transmissions through the atmosphere in cloud coverage conditions.

1.2.2 The reflection

Most Remote Sensing systems operate in the regions of the electromagnetic spectrum where wavelengths are mainly reflected by the surfaces on the Earth. The roughness of the surface and the wavelength of the incident radiation are key factors to evaluate the reflection properties of a specific surface.

One of the most important and frequently measured parameters in Remote Sensing is, therefore, reflectivity, which allows the spectral signature of a surface to be defined and identified. The spectral signature, or spectral reflectance curve, provides a measure of the ability of a given surface to reflect incident energy at various wavelengths and can be mathematically expressed through the following relationship:

$$\rho_\lambda = \frac{r}{i} \tag{1.4}$$

where r is the reflected energy, and i is the incident energy. Reflectivity measurements typically involve the region of the spectrum between ultraviolet and infrared; the reflectance curve, therefore, describes the reflectivity trend for a given surface in the spectral range $\lambda = 0.4 - 2.5m$. Depending on the environmental conditions (time of year, physical and chemical condition of the surface) and on the shooting conditions (sun - surface - sensor geometry), the reflectance surface curve varies. Drawing average reflectance curves, as shown in Fig. 1.4 it is particularly useful to provide important information on the behaviour of the surfaces under examination and to have an indicative value of how the variations described above influence the reflection.

1.3 Passive and Active Sensors

The term sensor refers to an electronic device capable of detecting electromagnetic energy from a specific scene and converting it into information. Then, the on-board recorders recorded and stored it in the form of an electrical signal. On this electrical signal, it is possible to extract useful information in a plethora of Remote Sensing application. A first and fundamental classification is made based on the sensor's functionality for the measurement of electromagnetic radiation. Consequently, the following two types of Remote Sensing are distinguished:

- Passive Remote Sensing: The sensor only measures the electromagnetic radiation emitted or reflected by the object analyzed;
- Active Remote Sensing: the sensor emits electromagnetic radiation and detects the fraction reflected by the objects placed on the Earth's surface.

Further, an important difference between active and passive instruments is related to the energy source. In the active case, it is inherent in the instrument itself. Conversely, in the passive case (photographic cameras, scanners, thermal imaging cameras), the energy source is external to the instrument, and the sensor measures the energy coming from the surface by spontaneous emission and by reflection.

In other words, the passive sensors detect electromagnetic radiation reflected, or naturally emitted, by the objects under examination located on the Earth's surface using natural sources, such as the Sun. The systems for passive Remote Sensing are of two categories:

- sensors operating in the visible (VIS) and near and medium Infrared, which collect the electromagnetic radiation emitted by the Sun and reflected from the objects on Earth's surface;

- sensors that operate mainly in the thermal infrared, which collect the radiations emitted directly from the objects on Earth's surface.

The reflected energy measurements can only occur when the Sun illuminates the objects under observation and therefore not at night; the detection of the energy emitted (in the thermal infrared case) can instead be carried out both during the sunshine hours and night.

On the other hand, the active sensors detect the electromagnetic radiation generated by themselves. The emitted radiation reaches the object under observation, and its reflected fraction is detected and measured by the sensor, following the interaction with the surface. Active Remote Sensing systems are divided into scattering systems, such as LIDAR(Light Detection and Ranging or Laser Imaging Detection and Ranging), which operate in the visible and infrared, and radar systems that operate in the microwave range. Among the main advantages offered by active sensors is the possibility of carrying out measurements at any time of day and night and, in the case of radar, also in all weather conditions.

1.3.1 Data Resolution Types

Each sensor is characterised by four properties:

- the spatial resolution;
- the radiometric resolution;
- spectral resolution;
- temporal resolution.

The spatial resolution is the minimum area on the ground seen by the instrument from a given height at a given time. It is represented by the size of the surface element recognisable in an image recorded by a Remote

Sensing system or, again, by the minimum distance within which two objects appear distinct in the image. The detail in an image depends on the spatial resolution of the sensor. The spatial resolution of passive sensors depends first and foremost on their Instantaneous Field of View (IFOV). The IFOV is the angular cone of the sensor's visibility and determines the area of the Earth's surface that is "seen" at a given height at a particular time. The observed area's size is determined by multiplying the IFOV by the distance from the ground to the sensor. This area on the ground is called the resolution cell and determines the sensor's maximum spatial resolution. If the object is smaller than this, it cannot be identified. Commercial satellites provide images with resolutions ranging from a few metres to several kilometres. Radiometric resolution represents the smallest difference in intensity that a sensor can detect between radiant energy values. Radiometric characteristics describe the information contained in an image. It is also defined as the number of discrete levels into which a signal can be divided. The data in an image is represented by positive digital numbers ranging from 0 to a power of 2. The image's data is generally displayed in a range of grey tones, with black representing the digital number 0 and white representing the maximum value (for example, 255 in 8-bit data). Comparing a 2-bit image with an 8-bit image, we can see a big difference in the level of detail detectable as a function of radiometric resolution. The higher the number of grey levels, the better the radiometric resolution. The spectral resolution is the width of the sensor's spectral bands, i.e. the minimum interval between the average wavelengths of two spectral bands that a sensor can separate. The better the spectral resolution, the narrower the wavelength range for a particular band. Many Remote Sensing systems record the energy of separate wavelength ranges at different spectral resolutions. These are called multispectral sensors. Advanced multispectral sensors, called hyperspectral, record hundreds of narrow spectral bands in the visible, near-infrared, and mid-infrared por-

Table 1.2. Main bands and their sub-bands of the microwave window, where HF, UHF and VHF are respectively high frequency, ultrahigh frequency and very high frequency

Bands	Frequency Range	Wavelengths Range [m]
HF	3-30 MHz	100.0- 10.00
VHF	30-300 MHz	10.00 - 1.000
UHF	0.3-3 GHz	1.000 - 0.100
P	0.225-0.39 GHz	1.333 - 0.769
L	0.39-1.55 GHz	0.769 - 0.193
S	1.55-4.2 GHz	0.193 - 0.071
C	4.2-5.75 GHz	0.071 - 0.052
X	5.75-10.9 GHz	0.052 - 0.027
Ku	10.9-22 GHz	0.027 - 0.014
Ka	22-36 GHz	0.014 - 0.008
Q	36-46 GHz	0.008 - 0.006
V	46-56 GHz	0.006 - 0.005
W	56-100 GHz	0.005 - 0.003

tions of the electromagnetic spectrum. Their very high spectral resolution makes it possible to discriminate between different objects based on their spectral response in each of the bands. Temporal resolution is the time between successive acquisitions of the same area. The revisiting period of a satellite sensor is normally several days. The temporal resolution of a sensor depends on various factors, including satellite and sensor characteristics, overlapping of the imaging amplitude and latitude. **

1.4 Microwave Remote Sensing

As already mentioned before, the microwave portion of the spectrum covers the range from approximately 1 cm to 1 m in wavelength (see Table 1.2), and both active and passive forms of Remote Sensing are possible in microwave sensing. For instance, this longer wavelength radiation allows the detection of microwave energy under almost all weather (cloud, haze, or dust)and environmental conditions so that data can be collected at any time. However, the heaviest rainfall as the longer wavelengths are not

susceptible to atmospheric scattering affects shorter optical wavelengths. Passive microwave sensing is similar in concept to thermal Remote Sensing. The microwave energy recorded by a passive sensor can be emitted by the atmosphere, emitted or reflected from the surface or transmitted from the subsurface. Microwave energy of some magnitude is emitted by all objects or surfaces, depending on its temperature and moisture properties, though the amounts are generally tiny compared to optical wavelengths. To detect enough energy to record a signal, most passive microwave sensors are characterized by low spatial resolution. Passive microwave sensors are typically radiometers or scanners and operate using a single antenna used to detect and record the microwave energy.

Active microwave sensors illuminate the target with their own source of microwave radiation. Active microwave sensors are generally divided into two distinct categories: imaging and non-imaging. The sensor transmits a microwave (radio) signal and detects the backscattered portion of the signal. The most used imaging active microwave sensors are RADAR, an acronym for RAdio Detection And Ranging, i.e., target detection and target distance determination by exploiting the EM radiation-matter interaction. In this context, imaging radar is only exclusively focused. As with passive microwave sensing, a major advantage of RADAR is that it can be used to image the surface at any time, day or night, and all-weather conditions. Because of the fundamentally different operational way between the active RADAR and the passive sensors (in visible and infrared of the spectrum), radar and optical data can be complementary to one another as they offer different perspectives of the Earth's surface, providing different information content. However, the applications are the same, and therefore many works tend to integrate this different information. In particular, radar systems evolved tremendously over the years. Their purposes initially limited to target detection and distance, allow for target tracking, imaging, identification, and classification. As a result, the range of these

modern systems applications is wide, from traditional military and civil aircraft and vehicle tracking to 2-D and three-dimensional imaging, Earth monitoring, and many others.

1.4.1 SAR: Synthetic Aperture Radar

During its forty-year history, SAR has become a fundamental tool for the Earth and other celestial bodies' observation. Because it is an active microwave sensor, it allows continuous monitoring of the sensed surface in all weather conditions. The electrical and structural properties of the surfaces can gather via a proper modelization from its information. Furthermore, the SAR technology's global monitoring capability has led federal space agencies, intergovernmental organizations, and private companies to use a series of SAR systems in the last decades. However, because of acquisition geometry and speckle noise, SAR imagery interpretation is challenging and is still an exemption of SAR expert users.

A radar emits radiofrequency EM waves towards a region of interest and holds the EM energy reflected from objects perhaps present in that ward. A Radar system is composed of a transmitter, a receiver, an antenna, and an electronics system to process and record the data. The transmitter generates a signal, successive short bursts (or pulses of microwave) at regular intervals, which is focused by the antenna into a beam. The RADAR beam illuminates the surface obliquely at a right angle to the platform direction. The EM wave incident on the target provokes electrical currents and reradiates EM energy into the environment. Then, the reflected (or backscattered) energy is detected by the antenna. By measuring the time delay, the location of the target can be determined. The backscattered signals make a two-dimensional image of the surface thanks to the sensor movement. Likewise, other facades on the ground contribute to the received signal, and this is called clutter. A portion of the returned signal propagates back to the radar is caught by the receiver antenna. The re-

ceived echo is processed by the radar receiver. The range, R, of a detected spot can be computed by the beaten time, ΔT. Since the EM wave travels to the target and back to the radar, the distance is obtained by 2R:

$$R = \frac{c \cdot \Delta T}{2} \tag{1.5}$$

The microwave portion of the spectrum is relatively larger than the visible and infrared part, and there are several wavelength ranges or bands commonly used which given code letters during World War II and remain to this day, as reported in Tab. 1.2. Images acquired at different radar bands have significant differences due to the way the radar energy interacts with the objects under investigation.

1.4.2 Operating Modes

A SAR system can work in three different modes: stripmap, spotlight, and scan-SAR, as shown in Fig. 1.5. In the stripmap configuration, the antenna beam is maintained at a fixed angle concerning the flight direction, and the antenna footprint covers a strip on the sensed surface following the system movement. A stripmap image is then limited in the range direction and unlimited in the azimuth one. In the spotlight configuration, the antenna beam is steered along the system path to cover a confined area on the ground, and it determines an improvement in the azimuth resolution. The better azimuth gives a traded off by a limited image in the along-track direction. Finally, the scan-SAR configuration mode allows for an expansion of the acquired image in the across-track direction. In this configuration, the synthetic aperture is split into orthogonal sub-apertures; each one looked at a different angle. The system cyclically switches the beam among the different angles, thus covering a larger area than the stripmap configuration. However, the shorter synthetic aperture determines a worse azimuth resolution. To avoid range-doppler ambiguities, the side-looking

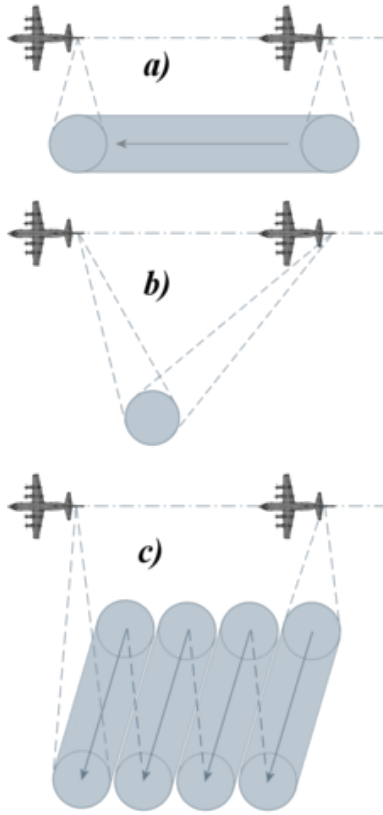


Figure 1.5. Different operating modes: a)stripmap-SAR; b) spotlight-SAR; c) Scan-SAR

viewing geometry is used, in which the radar antenna is directed to the left or right of the flight track and typically perpendicular to the flight direction.

1.4.3 Scattering and Polarizations

In Earth observation, the object observed by radar is generally a portion of the surface of which a map is generated by associating the amplitude, phase and polarization data of the electromagnetic field diffused by

each cell of spatial resolution. The cell can be seen as the intersection of the "main lobe" of the radar antenna (real or synthetic) with the surface itself. Therefore, the dimensions of the cell are determined by both the satellite parameters and those of the radar system. The most common product that a satellite radar provides an image of the surface consisting of the two-dimensional distribution of the scattering intensity that comes from each image element called a pixel (Fig. 14.3). This map is obtained by associating the value of the corresponding backscattering coefficient to the coordinates of the pixel center, defined as the backscattering cross-section per unit of surface

$$\sigma_0 = \frac{\sigma(\theta, \phi, \theta, \phi + \pi)}{\Delta A} \tag{1.6}$$

where ΔA is the area of the spatial resolution cell. Since the backscattering coefficient depends on the local characteristics of the surface, the radar image is the starting point for the production of maps of the parameters of interest to users. When the other system parameters are equal (frequency, local angle of incidence, spatial resolution, satellite altitude, etc.), the field's polarisation greatly influences the radar response. Apart from the SRTM missions, the first systems continuously in orbit operated by transmitting and receiving the same linear polarization, producing maps of the horizontal co-polar backscattering coefficients, σ^0_{hh}, or vertical co-polar, σ^0_{vv}, are JERS, RADARSAT missions, and ERS-1 and ERS-2 missions. To define the vertical and horizontal polarization directions, reference is made to the plane locally tangent to the geoid. The ENVISAT mission, started in 2002, was the first to provide images also of the backscattering coefficient on cross-polarization, σ^0_{hv}, where the first subscript (h) indicates the received polarization and the second (v) the transmitted one. Given the importance of polarisation in the generation of products from radar measurements, the most recent satellite systems operate on multiple polarizations or are completely polarimetric, which means they measure both

the amplitude and the phase of the horizontal and vertical components of the field. The essential product of a polarimetric radar is the scattering matrix [**S**] which binds the incident field

$$\underline{E}_{0i} = E_{0v}^{(i)} \underline{v_0} + E_{0h}^{(i)} \underline{h_0} \tag{1.7}$$

and backscattered field

$$\underline{E}_{0s} = E_{0v}^{(s)} \underline{v_0} + E_{0h}^{(s)} \underline{h_0} \tag{1.8}$$

through the

$$\begin{bmatrix} \underline{E}_{0v}^{(s)} \\ \underline{E}_{0h}^{(s)} \end{bmatrix} = \frac{e^{-jk_0R}}{R} [\mathbf{S}] \begin{bmatrix} \underline{E}_{0v}^{i} \\ \underline{E}_{0h}^{(i)} \end{bmatrix} = \frac{e^{-jk_0R}}{R} \begin{bmatrix} f_{vv} & f_{vh} \\ f_{hv} & f_{hh} \end{bmatrix} \begin{bmatrix} \underline{E}_{0v}^{i} \\ \underline{E}_{0h}^{(i)} \end{bmatrix} \tag{1.9}$$

In 1.9 the $\underline{E}_{0i}$ field that affects the observed object at a distance R and the $\underline{E}_{0s}$ field that returns to the radar are expressed through the two complex vertical and horizontal components. Since the absolute phase of the field is arbitrary, the vertical component is often assumed to be real, and the phase difference $\delta_{vi} - \delta_{hi}$ between the two components is associated with the horizontal one. The elements of [**S**], co-polar and cross-polarization complex scattering functions,

$$f_{pq} = f(f, \theta_i, \phi_i, A_m) \qquad p, q \equiv v, h \tag{1.10}$$

are functions of the frequency f, the local direction of incidence θ_i, ϕ_i, and the biological, physical and morphological parameters A_m, (m = 1, 2, ...), the surface element that spreads the incident electromagnetic wave. In radar applications where the phase reference is not maintained, the amplitude and polarization state of the electric field are often described through the modified Stokes vector, whose elements are more directly connectable to the surface density of power carried by the wave

$$\underline{Y}_m = \begin{bmatrix} |E_v|^2 \\ |E_h|^2 \\ 2\mathscr{R}[E_v E_h^*] \\ 2\mathscr{I}[E_v E_h^*] \end{bmatrix} \tag{1.11}$$

When using this representation, the Stokes vector of the backscattered wave $\underline{Y_{sm}}$ is linked to that of the incident wave $\underline{Y_{im}}$ through the Muller Matrix $[\mathscr{M}]$:

$$\underline{Y}_{sm} = \frac{1}{R^2}[\mathscr{M}]\underline{Y}_{im} \tag{1.12}$$

The 4×4 $[\mathscr{M}]$ matrix, whose elements are related to the previously defined scattering functions, provides the scattering coefficients for any combination of the incident (transmitted) and diffuse (received) field polarizations. These elements are in turn functions of the biological, physical and morphological parameters of the pixel to which they refer. A fundamental aspect, decisive for satellite monitoring and remote sensing in general, is to extract the information of application interest from the scattering functions or the elements of the Muller Matrix measured by radar systems. Given the complexity of the problem, this often requires sufficiently complete sets of measures (even at multiple frequencies and/or different angles of incidence) and the development of sophisticated inversion and estimation techniques, such as those based on computational intelligence (for example, neural network algorithms).

1.4.4 Polarizations

As mentioned before, another important feature of microwave radiation is the polarization, which is the orientation of the electric field. The microwave radiation is transmitted horizontally (H) or vertically polarization (V). Similarly, the backscattered energy is received by the antenna in both horizontally or vertically polarization. Thus, considering transmitted

e received acquisition mode, four combinations can be considered:

- HH: horizontal-horizontal in both transmission and reception,
- VV: vertical-vertical in both transmission and reception,
- HV: horizontal in transmission and vertical in reception, and
- VH: vertical in transmission and horizontal in reception.

The HH and VV combinations are defined as co-polarized, and the other two as cross-polarized. As for the wavelength variations, the radiation from a specific surface is differently backscattered. Because of the fundamentally different operational way between the different polarization modes, the resulting combinations can be complementary to each other and provide different pieces of information.

This page intentionally left blank.

Chapter 2

Convolutional Neural Networks and Deep Learning

The terms artificial intelligence, machine learning, and deep learning are sometimes misused as synonyms of the former. In 1956, John McCarthy coined the term "Artificial Intelligence" (AI) at the Dartmouth Summer Research Project on Artificial Intelligence conference. AI involves all those operations that are characteristic of the human intellect and performed by computers. These include planning, language understanding, object and sound recognition, learning and problem-solving. Fascinating is the relationship between AI and Internet-of-Things (IoT) similar to that between the brain and the human body. Our brain makes decisions based on sensory inputs, just as, thanks to artificial intelligence, with the IoT, we can intelligently process all the data acquired from sensors scattered around the planet and make the necessary decisions. Machine learning is essentially a set of algorithms for the implementation of AI; a kind of AI subgroup that focuses on the ability of machines to receive a series of data and learn models on their own, modifying algorithms as they receive more information about what they are processing. In 1958, Rosenblatt

developed the first machine learning algorithms [4]. The term "machine learning" is understood as "the ability of a machine to learn without being explicitly programmed". Artificial vision systems or a computational system's ability to recognize objects acquired digitally by image sensors represented a classic example of machine learning. Finally, deep learning, on the other hand, is one of the many approaches related to machine learning and has taken its cue from the structure of the brain or the interconnection of various neurons. Deep learning is often referred to simply as a "deep neural network", referring to the many layers involved. Deep learning uses huge neural networks models with various processing units; take advantage of computational advances and training techniques to learn complex models through a considerable amount of data. One of the most common is precisely the multiple-layered models of inputs, commonly known as deep neural networks, described in the literature [5]. Deep Neural Networks include multiple non-linear operations, mainly of the convolutional or subsampling type (pooling layer). In addition to the two mentioned operations type, the Rectified Linear Unit (ReLu) layers and the fully connected layers are commonly used, strongly connected levels placed at the end of the network in the first segmentation solutions structure. These layers all together constitute the basic block usually present in each CNN and are cascaded between them. Before 2006, searching for the optimal solution in these deep cascade-based architectures' parameter space was a formidable research challenge due to the extremely high number of variables. Recently, the deep learning algorithms have significantly decreased the complexity of this problem [6]. To understand the complexity of this kind of problems is possible to visualize a deep architecture's parameter space like a desert with many dunes, and searching for the best solution is like searching for a ball among that dunes.

In the following sections, the main principles of machine learning and deep learning is described. Then, the description of some of the most used

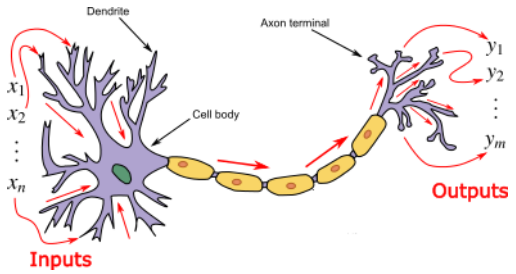


Figure 2.1. Biological Neural Network

neural networks is schematized, and some notions on how to carry out adequate training. Finally, in the context of deep learning, the supervised training considered in this work is described in more details.

2.1 Neural Network: from Perceptron to Multi-Layer Perceptron

Neural networks emulate the human brain's behaviour, enabling computer programs to recognize models and explain common problems in AI, machine learning, and deep learning. In the last decades, motivated by their strengths, many works have used this branch of techniques in many research fields. Neural Network is bio-inspired and tries to emulate human brains[7], [8].The human brains consist of a large number (approximately 10^{11}) of highly connected elements (approximately 10^4 connections per element) called neurons. These neurons have three principal components, as shown in Fig. 2.1: *dendrites*, receptive networks of nerve fibre, *cell body*, and *axon*, a single output fibre. The dendrites can carry out the electrical signals to the cell body, which elaborates the incoming signals, and finally, the axon carries out the elaborated signal to other neurons. The synapse is the contact point between an axon of one cell and a dendrite of another cell.

An artificial neuron network (ANN) is a computational model based

on biological neural networks' structure and functionalities. The ANNs are non-linear statistical model and can found and model the complex relationships between inputs and outputs. The ANNs are sometimes used as a black-box function approximation tool and are considered the most cost-effective and ideal methods to estimate complex functions or statistical distributions. Further, with relatively simple ANNs, people without specific expertise in a specific context can get quick and cheap solutions to complex problems. As biological neural structures can change throughout life, so ANNs can change throughout a specific training phase. These changes mainly consist of strengthening or weakening synaptic junctions. The complexity of the brain is difficult to achieve by Artificial Neural Networks(ANN). However, ANNs are designed to emulate two main characteristics of biological neural networks: the high interconnections between the neurons and the ability to determine specific functions with these interconnections.

2.1.1 Perceptron

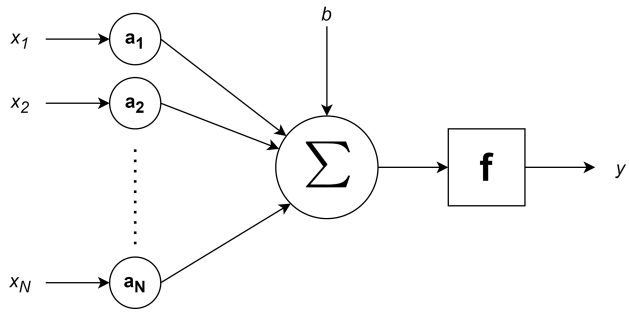


Figure 2.2. A perceptron.

The main and easiest artificial neuron inspired by the biological counterpart is the *perceptron*, which performs a weighted sum of the inputs, as shown in Fig. 2.2. Using a vectorial notation, the output y of the

perceptron can be written as follows:

$$y = f(\mathbf{a}^{\mathbf{T}}\mathbf{x} + b) \tag{2.1}$$

where, $\mathbf{x}$ is the input, b is a *bias*, $\mathbf{a}$ is the weight vector and f is a function called *activation function*.

This simple model can be related to a single biological neuron. The weight $\mathbf{a}$ corresponds to the strength of a synapse, the cell body to the summation and the activation function (f), and the neuron output y as the output signal on the axon.

The values $\mathbf{a}$ and b are learnable parameters of the neuron, with the observation of examples of input/output. Typically the activation function is chosen by the designer according to some specific goal. The network designer determines the better activation functions for different purposes and the learning strategy. Fig. 2.6 shows the most common linear and non-linear activation functions, that can be used in different problems to solve.

Despite processing many inputs, a single neuron is not sufficient for a more complex problem, so the use of several neurons operating in parallel is necessary. This set of neurons is called a layer.

A single-layer network of M neurons is shown in Fig. 2.3. Note that each of the N inputs is connected to each of the neurons and that the weight vector became a matrix that has M rows. As can be seen in Fig. 2.3, each neuron can have a different activation function.

2.1.2 Multi Layer Perceptron

Human brains organize their concepts in a hierarchical structure, so the artificial brains do. First, human brains learn simple concepts and then combine them to represent more sophisticated ideas. Motivated by this learning technique, researchers have dedicated many efforts to using

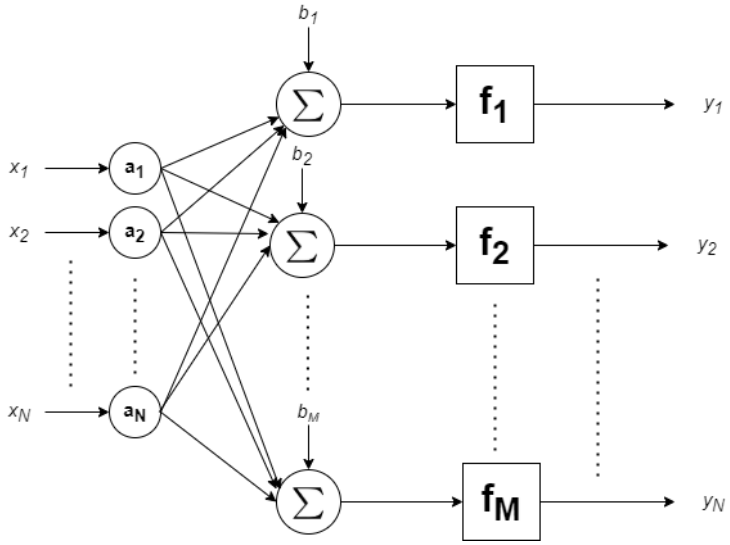


Figure 2.3. A layer composed by M neurons .

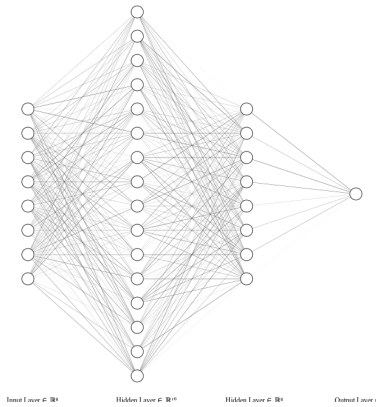


Figure 2.4. An example of a Multi Layer Perceptron.

many abstraction and processing levels to solve computational problems. The network output can be written as:

$$\mathbf{y} = f(\mathbf{A}\mathbf{x} + \mathbf{b}) \tag{2.2}$$

To simulate this combination of simple concepts, in ANN, it is very common to consider more than one layer. In this case, the output of one layer represents the next layer's input and so on. The last layer is called output layer. The other layers are named hidden layers.

The network obtained by these layer-based structures are commonly called multilayer perceptron or multilayer networks, as shown in Fig. 2.4, and can solve more complex problems than the single perceptron or the single-layer networks.

2.2 From Multi-Layer Perceptron to Deep Learning

In recent years, deep learning research has gained remarkable momentum in both academia and industry. In particular, deep learning techniques had a terrific impact on several computer sciences and engineering fields, including object recognition, speech recognition, natural language processing, robotics, driverless cars, AI gaming and remote sensing.

As already described, artificial neural networks (ANN) try to emulate remarkable human abilities. In particular, the human visual system can easily solve complex pattern recognition problems that is a complex function for the most powerful computers. This happens thanks to the interconnected neurons. Based on this observation, the artificial neuron is designed to perform a straightforward task. Given a set of inputs, (x_1, ..., x_K), it outputs a nonlinear function of its weighted average:

$$y = f(b + \sum_{k=1}^{K} w_k \cdot x_k) \tag{2.3}$$

Doing this matches simple shapes, such as edges, lines or blobs, in the case of images. The outputs of a large layer of neurons operating in parallel become the input of a further layer of neurons, and so the next layer

combines basic features to extract features of higher semantic value. This proceeds through several layers, permitting a high level of abstraction. By modifying neuron weights in response to suitable input stimuli, the network learns how to perform all sorts of desired tasks. In a traditional fully-connected network, each neuron takes as input the outputs of all neurons of the previous layer and feeds its output to all neurons of the next layer. Consequently, a deep ANN includes a very large number of weights, which must be learned on a proportionally large training set, calling for an exceeding computational complexity. Convolutional neural networks (CNN) have been designed to overcome this problem by renouncing full connectivity. In CNN, each neuron has a limited receptive field, processing features observed only in the neuron's local neighbourhood. This makes full sense for many sources, notably for images, where spatial features are intrinsically local (especially in lower layers) and spatially invariant. Due to this latter property, in particular, one can use the very same set of weights for all neurons of the same layer, by which the output at neuron (i, j) can express as the convolution with the previous layer input

$$y_{i,j} = f(b + \sum_{m=1}^{M} \sum_{n=1}^{N} w_{n,m} \cdot x_{i+n,j+m}) \tag{2.4}$$

or, in compact matrix notation:

$$y = f(\mathbf{b} + \mathbf{w} * \mathbf{x}) \tag{2.5}$$

where * denotes convolution. CNNs reduce the number of connections among neurons drastically, hence the number of free parameters to learn, enabling deep learning in practical applications. CNNs have become very popular in recent years, thanks to software available online, relatively fast training achievable with cheap and powerful GPUs, and also thanks to the huge mass of labelled visual data available on the web [], essential

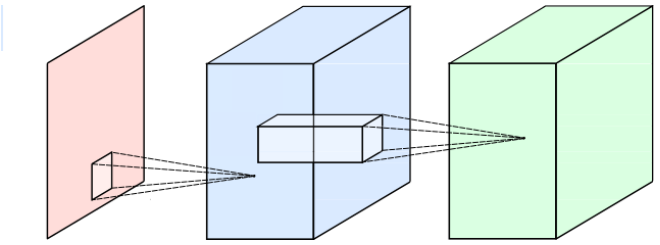


Figure 2.5. An Example of Convolutional Neural Network.

for training this kind of networks. When dealing with images, input variables are two-dimensional, spatially related entities. Spatial dependencies are propagated throughout the network, which justifies why the features output by intermediate (hidden) layers are represented as images. Eventually, depending on the net's specific task, the output of the network may be itself an image (think of denoising, segmentation, super-resolution), or else a set of spatially unrelated decision variables (detection, classification, recognition).

2.3 Convolutional Neural Networks (CNNs)

The Convolutional Neural Network, also known as CNN or ConvNet, is a discriminative deep architecture [9] and this is chosen in this work because this kind of ANN is the best for the images. Firstly, CNN is very similar to an ordinary ANN. Indeed both architectures consist of neurons having learnable (i.e., tunable) weights and biases. Furthermore, it receives an input (i.e., a single vector) and transforms the input through several hidden layers. In each hidden layer, there is a set of learning units called neurons. The neurons of a hidden layer are fully connected with neighbourhood layers. In the typical classification problems, the fully connected layer is referred to as the output layer. Each neuron receiving several inputs takes a weighted sum of them, passes it through an activation function, and responds with an output. The deep ANN structure is a

conventional training method in many areas, but there is a problem related to the full connectivity between nodes when the input is a high-resolution image, and so the conventional ANN cannot process well. Therefore, the convolutional layers were proposed instead of full connectivity in the neural network layers. The CNN consists of several layers of convolutions with non-linear activation functions to compute the output. An idea of the interconnection between different layer in CNN is shown in Fig. 2.5. Each layer applies different filters (in the order of hundreds to thousands) and combines their results. During the training phase, CNN automatically learns its filters' values based on the given task. Specifically, CNN sees localized shapes in each region of the input and follows that the output gives information about this area. In its first layer, CNN may learn to detect the edges from the analyzed area's pixels. Then, in the second layer, the CNN combines the previous output to detect simple shapes. Using these shapes, the higher sequent layers of CNN may learn higher-level features like facial shapes and so forth. A classifier is used to exploit these high-level features in the final layer (sometimes a fully connected layer). Deep CNNs are very successful in learning task-specific features, which have provided much-improved results. Generally, it is not easy to obtain a substantially large training set (composed of input and labelled output), which is a key challenge for solving a new task. Besides, CNNs have reported being extensively used in unsupervised methods [10], [11], [12].

2.3.1 CNN: the basic components

Of course, the first and most important layer is the Convolutional one. The convolutional layer aims to extract the significant features of the images and is the main layer of the network. It is designed to identify patterns, such as curves, angles, circumferences or squares depicted in an image. The convolutional levels can be multiple, but this depends on the

network architecture: the greater their number, the greater the complexity of the characteristics they can identify. The filter (whose size is usually equal to 3×3), also called Kernel, is the matrix involved with the input and allows you to obtain an Activation Map, called Feature Map.

In summary, CNN learns the values of these filters during training, and it is clear that the more filters are used, the more features are extracted. The obtained feature map depends on depth, stride and zero-padding. The first corresponds to the number of filters used, the stride to the number of pixels by which the filter moves when observing the input image, while the zero-padding is the number of zeros to complete the input matrix the edges (wide convolution).

Pooling Layer

On deeper levels, the number of features generally increases, so it becomes essential to find a way to reduce these feature sets' size. The pooling layer decreases the spatial dimensions while preserving the essential information, thus reducing the number of trainable parameters and the computation time required for training. This layer also makes the network invariant to small transformations such as distortion or translation concerning the initial image. Spatial Pooling can be of different types: Global Average Pooling (GA Pooling), Region of Interest Pooling (RoI pooling), and the most used Max-Pooling. The GA Pooling can allow an extreme reduction in dimensionality by reducing a tensor from the dimension $H \times W \times D$ to the dimension $1 \times 1 \times D$ and has been used a lot to reduce overfitting due to the reduction of the number of parameters within the network. The GAP layer reduces each feature map into a single number that summarizes the essential information in the feature map. RoI Pooling is a type of pooling layer that solves the problem of having areas of interest of different sizes in input and produces a small feature map of a fixed size, i.e. 7×7. This dimension is a hyperparameter of the network. This has allowed for a speeding up of the training phase, making the architectures

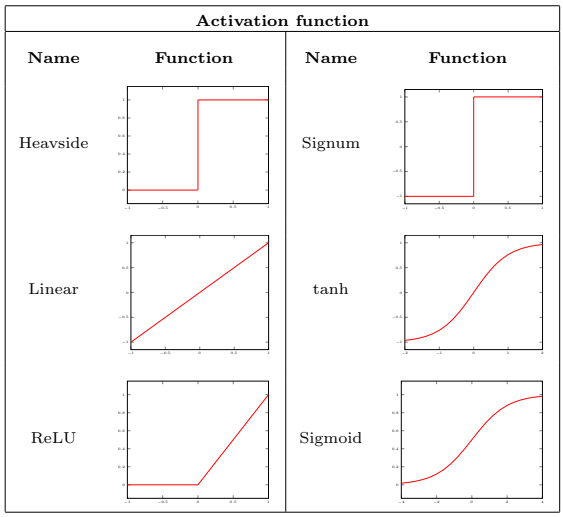


Figure 2.6. The most used activation functions.

that use them among the fastest in the field of Object Detection. It is used to produce in a single pass a single feature map from all those proposed by the Region Proposal Networks (RPN). Max-Pooling has often used in the construction of neural networks; in this level, the maximum or at least the largest value for each feature map is calculated. It allows us to identify if the required feature is present at the level it observes.

Activation Layer

The choice of the activation layer is crucial to obtain a valuable result. With the wrong activation layer, the results can totally be without significance, so some activation function (or layer) is briefly described in the following. The Activation Layers (some examples are reported in Fig. 2.6) are immediately used after Convolution (or Pooling) layers and is performed pixel by pixel.

The *Heavside* activation function is mainly used in binary classification, where the neuron has to decide true or false.

The *Sigmoid* function is widely diffused in the machine learning ap-

proached, especially for the logistic regression and some basic neural network implementations. However, the Sigmoid function has a main drawback: the derivative value is minimal, and it can be a problem in the back-propagation algorithm. The Sigmoid is defined as follows:

$$f(x) = \frac{1}{1 + e^{-\beta \cdot x}} \tag{2.6}$$

In the multi-class segmentation context, the Softmax function is preferred to the Sigmoid. In principle, both limit the output values for each class between 0 and 1, but the Sigmoid does it independently for each pixel, while the Softmax introduces the constraint of the unitary sum of the pixels to obtain a probability distribution. Mathematically, the Softmax activation function is the generalization of the Sigmoid function, and is defined as follows:

$$f(x) = \frac{e^{-\beta_i \cdot x}}{\sum_{k=1}^{K} e^{-\beta_k \cdot x}} \tag{2.7}$$

To realize faster training for large networks, most recent deep learning architectures use Rectified Linear Unit (ReLU) [13]. The $ReLU$ layer introduce non-linearity into the network since this layer replaces each negative number of the previous layer with the value zero allowing CNN to remain mathematically stable. It is often used for training because its derivative calculation is fast, and the value it assumes is zero only if the input is less than zero.

2.3.2 Recurrent Neural Networks (RNNs) or Long Short Term Memory Networks (LSTMs)

The Recurrent Neural Network (RNN) can be considered to be a deep generative architecture [14] as shown in Fig. 2.7.

The depth of an RNN may be as large as the length of the input data sequence. Therefore, the RNN is particularly useful for modelling the

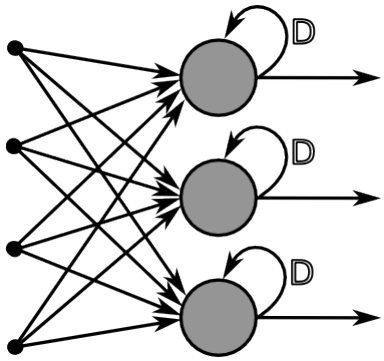


Figure 2.7. The Recurrent Neural Network's concept

sequence data in text and speech. Despite the potential strength, their use was restricted until recently due to the so-called "vanishing gradient" problem [15]. New optimization methods to train generative RNNs that modify stochastic gradient descent have been presented in the literature recently, [16], [17].

2.3.3 Stacked Auto-Encoder

Stacked auto-encoders can be a good example of how the earlier-mentioned Greedy Layer-Wise unsupervised pre-training is exploited. An auto-encoder refers to an ANN to learn efficient coding by encoding a set of data, as depicted in Fig. 2.8. The encoded data conveys a compressed representation of the data set. In other words, the auto-encoder can be exploited to perform data compression or dimensionality reduction. An input layer composes a typical auto-encoder architecture, and in series, there are several significantly smaller hidden layers, and finally, an output layer. The hidden layers encode the input data set while the output layer attempts to reconstruct the input layer. According to Bourlard [18] , if the auto-encoder architecture consists of just linear neurons or just a single sigmoid hidden layer, the autoencoder's optimal solution is correlated to the Prin-

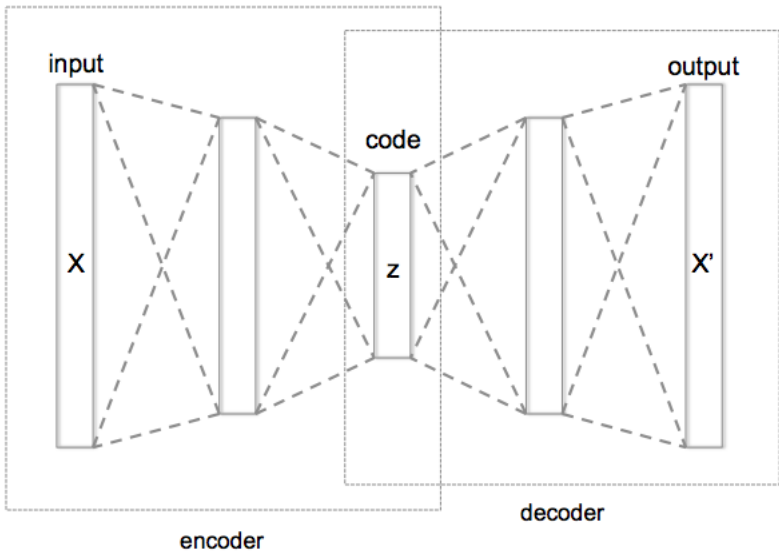


Figure 2.8. The Stacked auto-encoder's concept

cipal Component Analysis (PCA).

2.3.4 Design of a Convolutional Neural Network

In the design phase of a CNN, it is required to choose the model, the training dataset, and the training algorithm.

The *model* is the type of neural network chosen for the planned activity. The choice is constrained by the available computing power and the type of analysis or task to be performed. However, a too simple model could lead to underfitting, while an excessively complex model leads to overfitting. Therefore, it could be possible to build a model from scratch, layer by layer, or to exploit already known architectures/models that constitute state-of-the-art, such as VGG-NET or ResNet.

The dataset is composed, in our case, of a set of images, used to test and train the network architecture. Usually, we can choose a pre-existing dataset, used before in similar tasks, or create a novel dataset specific for our task. Once selected (or created), images can randomly be divided into: *Training-Set, Validation-Set and Test-Set.* The Training-Set is used to allow the network model, chosen in the first step, to learn from the data; the Validation Set to make evaluations on the model during the training phase (on the fly), and finally a Test-Set to try to verify that the model already trained can meet the initial expectations. In some cases, part of the validation set is also used as a Test-Set.

Modifications can be made to the chosen model, which may be structural, or the hyperparameters can be varied. The variation of the hyperparameters is a mathematical operation that affects the network's final performance: the hyperparameters characterize the Convolution and the train operation of the algorithm. In addition to the aforementioned, there are other elements of the arbitrary choice of great impact: the size of the input, the number of convolutional layers, the positioning of the pooling, since you can decide to apply it after each layer or alternately, but also the very choice of the pooling to be adopted. Finally, another consistent choice is related to the size of the filters since each convolutional layer can use filters with different types of "receptive fields". Other possible choice variables entrusted to the user can be the numerical choice of Fully-Connected Layers, the Loss Function's choice, or the neural activation function (which may be different for each layer). All these choices, which can modify the base block of a network, can bring about improvements and be properly tested.

The learning phase is very expensive from a computational perspective, so the different usage scenarios are fundamental. You can choose between training from scratch, Transfer Learning of a pre-trained network with fine-tuning or without further training.

Since deep networks require 2-3 weeks for training using multiple GPUs, it is common practice to take advantage of an already-trained network and use it for a different task. In particular, the weights from the training of a previous training are taken, even on a different task, and transferred to test the network or retrain a similar model: the network can start with already pre-trained weights. A pre-trained network can be used as a Feature extractor in which the last layer is removed (fully-connected). There is no further training. In this case, the information extracted is exploited to process the chosen data. Finally, it is possible to perform Fine-tuning, during which the weights of the pre-trained network are specialized to the current task. In the thesis's practical applications in question, we will evaluate comparisons between very deep pre-trained networks and lighter networks from scratch. The weights (or parameters) update was performed using an appropriate learning rule, specifically unsupervised and supervised learning.

Unsupervised machine learning allows to infer a function from an hidden structure of the "unlabeled" data, while supervised learning to infer a function from labeled training data. Unsupervised learning approaches [19], [20], [21] include, for example, clustering (K-Means) [22]. In supervised machine learning, the training data consists of a set of training examples, organized in pairs of an input object (typically a vector) and a desired output value:

$$\{\mathbf{p_1}, \mathbf{t_1}\}, \{\mathbf{p_2}, \mathbf{t_2}\}, \ldots, \{\mathbf{p_N}, \mathbf{t_N}\}, \tag{2.8}$$

where $\mathbf{p_i}$ is the input vector, and $\mathbf{t_i}$ is the target output vector, and N represent the number of input-target couple available.

A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. Supervised learning algorithms work

in two phases: training and testing. The training phase with mathematical optimization techniques (gradient descent) using backpropagation is based on two cyclical phases: propagation and updating of weights. In the propagation phase (forward step), the input elements cross the entire network and then recover, after propagation, the previously obtained outputs. The prediction error is calculated through the loss function, with which to calculate the gradient that will then be propagated backwards in the network. The second phase of updating the network provides that the gradient values are passed to the gradient descent algorithm (or any optimization algorithm), which will use them to update each neuron's weights. The weights are updated, weighing the error with a value called the learning rate. Training is carried out in two possible ways, namely online, full batch and mini-batch. In the first, each forward propagation step is immediately followed by an update step, while in the second, the propagation is carried out for all the examples of the training set, and after the update is made, in the last, the training dataset is randomly ordered and then it is broken down into smaller batches. The second approach leads directly to the final result but turns out to be infeasible and computationally exaggerated. In the first case, the online one gives efficiency problems since there are too many updates and has robustness problems, since there will be "incorrect" updates due to outliers' presence within the training set. The trade-off is the use of mini-batches.

2.3.5 Performance Optimization

To train a NN it is necessary to modify the weight matrix and the biases to have the desired behaviour to a given input. For this purpose it is necessary to introduce a performance function that should be optimized.

There are two steps involved in this process:

- **finding a performance index**: a quantitative measure of network

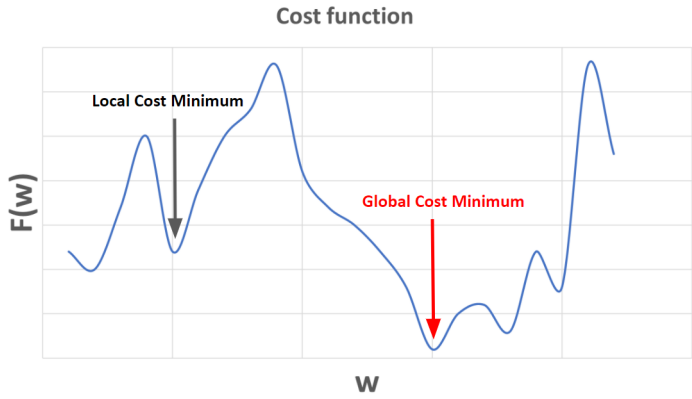


Figure 2.9. Non-convex optimization of a cost function

performance which is small when the network performs well and large when the network performs poorly;

- **optimization**: the research in the parameter space (adjust the network weights and biases) in order to reduce the performance index and reach an optimum point.

Usually the reached optimum is a local point since the optimization problem is not convex (Fig. 2.9).

Starting from there, the performance index to minimize is represented by $F(\mathbf{x})$, where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ is the parameter that is going to be adjusted. It will be assumed that the performance index is an analytic function, so that all of its derivatives exist. Then, it can be represented by its Taylor series expansion about some nominal point $\mathbf{x}^*$:

$$\begin{aligned} F(\mathbf{x}) = F(\mathbf{x}^*) + \nabla^T F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*} (\mathbf{x} - \mathbf{x}^*) + \\ + \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*} (\mathbf{x} - \mathbf{x}^*) + \ldots \end{aligned} \tag{2.9}$$

where $\nabla F(\mathbf{x})$ is the gradient and $\nabla^2 F(\mathbf{x})$ is the Hessian of the performance index $F(\mathbf{x})$.

Since the purpose of the training is the minimization of the performance index, it should be useful find the direction $\mathbf{p}$ that has the greatest slope. The maximum slope will occur when the inner product of the direction vector and the gradient is a maximum. This happens when the direction vector is the same as the gradient.

The Eq. 2.9 can be written as:

$$\begin{aligned} F(\mathbf{x}) = F(\mathbf{x}^* + \mathbf{\Delta x}) = F(\mathbf{x}^*) + \nabla^T F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*} (\mathbf{\Delta x}) + \\ + \frac{1}{2}(\mathbf{\Delta x})^T \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*} (\mathbf{\Delta x}) + \dots \end{aligned} \tag{2.10}$$

where $\Delta x = \mathbf{x} - \mathbf{x}^*$. Necessary condition for the optimality are:

- First Order: the gradient in $\mathbf{x}^*$ must be 0;
- Second Order: the Hessian matrix must be positive semi-definite.

The optimization will result in a direction research using an iterative process. At the *k-th* step, starting from $\mathbf{x}^k$, the result of the optimization step will be $\mathbf{x}^{k+1}$ that can be obtained as follows:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{p_k} \iff \mathbf{\Delta x} = \mathbf{x}^{k+1} - \mathbf{x}^k = \alpha_k \mathbf{p_k} \tag{2.11}$$

where $\mathbf{p_k}$ is the moving direction in the parameter space and α_k is the magnitude of this direction, that is called *learning rate*. Setting this parameter is one of the challenge for a good training. If it is too small, the search of the local optimal point is more accurate, but the convergence is slower; on the other hand, if it is too high the search of the optimal point is not much accurate and some local optimal point can't be found, but the

convergence it can be faster. There are a lot of techniques to find the local optimal point. The most popular is the Steepest Descent.

2.4 Motivation: Deep Learning in Remote Sensing

The deep architectures are categorized into three types, namely, generative, discriminative, and hybrid deep architectures, and these names are used depending to several usage. A generative deep architecture characterizes the high-order correlation properties of the input data for synthesis purposes. On the other hand, a discriminative deep architecture is used for pattern classification or recognition purposes. By combining the generative and discriminative deep architectures, a hybrid model may be constructed, particularly to carry out discrimination tasks, which are aided by the optimized outputs obtained from the generative architecture [5]. It is worth noting that the hybrid deep architecture is not the same as feeding the outputs of a traditional neural network to a Hidden Markov Model (HMM) [23] . However, regardless of their purpose, with the exception of Convolution Neural Networks (CNNs), deep architectures could not be successfully trained before 2006. The state-of-the-art deep learning algorithms, on the other hand, still found its operation upon multi-layer architectures according to the work conducted by Bengio et al. [24].

A key reason to use the afore-mentioned deep architectures is that they can more efficiently represent a complex non-linear function [6] compared to the MLP architectures. In other words, for a non-linear function to be compactly represented, a significantly large (i.e., deep) architecture is required. Another reason of using deep architectures is that they can support transfer learning, i.e., the ability of a learning algorithm to share knowledge across various tasks. Because deep learning algorithms learn features which capture underlying factors, they may be useful for carrying

out other tasks. This idea of knowledge sharing was demonstrated in [24]. In addition, in the two transfer learning challenges held in 2011, learning algorithms leveraging deep architectures exhibited superior performance compared to existing learning algorithms [25]. Furthermore, some works demonstrated the successful application of transfer learning [26]. However, with limited availability of data or computing resources (few GPUs available), it is sometimes advisable to move towards less deep solutions, equally able to approximate complex non-linear functions.

In the last decades, widely breakthroughs were obtained by deep learning techniques in diverse domains, such image analysis, speech recognition, autonomous cars, or the arts, encouraging to use that kind of solutions also in remote sensing context, like in [27]. Besides, the amount of remote sensing data is constantly growing due to the rise of very-high-resolution sensors and short satellites' revisit time. Deep learning is a hot topic in CV community and has had many successes in various field, specially with the speed up of the GPU. It gives very accurate results both in generative [28] and in discriminative [29, 30] problems. In particular, in the last years, there was a lot of breakthroughs in deep learning solutions that improved the results and the accuracy obtained by generative and discriminative neural networks. In the first field, there was a very prominent work, that is Generative Adversarial Network (GAN) [31], that inspired a lot of following related works, such as conditional GAN and cycle GAN [28, 32]. Instead, in discriminative issues, there was a series of interesting works that utilize the feature maps (or output of a particular layer) for semantic segmentation, that means to detect, localize and recognize object in computer vision images [29, 33–35]. Definitely, various further remote sensing problems have been favourably addressed with deep learning: object detection, image restoration, image enhancement, and so on. Deep learning has demonstrated to be relevant in many species of remote sensing imagery: synthetic aperture radar, multispectral and hyperspectral imagery,

and many others.

In [36], it is shown that a residual strategy is successful in deeper neural networks, so it contrast the higher training error when the neural network depth significantly increases. Inspired by this idea, in many remote sensing works, a residual approach is used, such as in [37] and in [38].

This page intentionally left blank.

Chapter 3

Deep Learning for Sentinel-2 Super Resolution

In this Chapter, we propose to leverage Convolutional Neural Networks (CNNs) to provide a fast, upscalable method for single-sensor fusion of Sentinel-2 (S2) data, to provide a 10 m super-resolution of the original 20 m bands. The proposed super-resolution approaches benefit from the details provided by 10-m Sentinel-2 bands to recover the missing spatial details of the lower resolution 20-m bands, following a pansharpening-like problem [39]. Using Convolutional Neural Networks (CNNs) successfully in generative and super-resolution frameworks, like in [40] and [41], encouraged us to develop new solutions in this application field. Without affecting the generality, the proposed methods have been tested on the Sentinel-2 bands. This section's main objective is to describe the Deep Learning (DL) based super-resolution approaches proposed in three years of my PhD. Other very recent approaches,both model-based [42, 43] and data-driven CNN-based [44], aimed to super-resolve the whole Sentinel-2 dataset testify the relevance of the topic to the community. A preview of the proposed solution compared with a simple bicubic interpolation is

ρ_8 (PAN-like) ρ_{11} (↑, bicubic) ρ_{11} (↑, proposed)

Figure 3.1. Example of super-resolution of ρ_{11} (SWIR band).

given in Fig. 3.1. On the left is one of the "guide" bands, ρ_8, in the middle is the bicubic interpolation of ρ_{11}, and on the right is reported the proposed upsampling.

In this section, we present the evolution of the super-resolution approach from a simple DL approach to an efficient and refined algorithm able to obtain valuable results with limited complexity. First, in [45], we focused on a general machine learning and deep learning approach. Specifically, the method, called M5, benefits from the details provided by 10-m Sentinel-2 bands to recover the missing spatial details of a single lower resolution 20-m band, the Short Wave Infrared one (SWIR) using a three-layer CNN.

Then, in a subsequent work [46], substantial and specific changes were considered:

a. in addition to 10-m PAN-like bands, we included all the 20-m Sentinel-2 images in the input stack. In the previous work, we only considered one MS band in input and one in output;

b. we added a high pass filter of the whole input stack as in [47];

c. we applied a batch normalization of the input stack to speed up the training phase [48];

d. we developed the residual solution that improves the performances in terms of some considered metrics, as reported in [36];

Thus, the Fast Upscaling of SEntinel-2 images (FUSE) approach is an evolution of the *proof-of-concept* works presented in [45, 46]. In particular, the major improvements with respect to the method in [49] reside in the following changes:

a. Architectural improvements with the introduction of an additional convolutional layer.

b. The conception of a new loss function which accounts for both spectral and structural consistency.

c. An extensive experimental evaluation using diverse datasets for testing that confirms the generalization capabilities of the proposed approach.

The rest of the Chapter is organized as follows. First, more details about the Sentinel-2 dataset's characteristic are given. Then, the proposed approaches, from the "simplest" to the most sophisticated solution, are described.

3.1 Sentinel 2

The Sentinel-2 mission consists of two multispectral platforms (see in Fig.3.2): Sentinel-2A, launched on June 23, 2015; and Sentinel-2B, launched on March 7, 2017.

The Sentinel-2 mission provides a multi-resolution stack composed of 13 spectral bands between the visible and short-wave infrared (SWIR), distributed over three resolution levels, as shown in Tab.3.1. Four bands lying between visible and near-infrared (NIR) are given at the finer resolution of 10 m, while the remaining ones are provided at 20 (six bands) and

Figure 3.2. Sentinel-2 satellite.

60 (three bands) m due to a trade-off between storage and transmission bandwidth limitations.

The Sentinel-2 satellites render data complementary to missions such as the SPOT and LANDSAT satellites. Also, in combination with radar data from the Sentinel-1 mission, such as climate change (monitoring deforestation or desertification), land cover/use mapping, water management, soil protection, emergency management, monitoring borders, food security/early warning systems, and so forth [50]. In general, the Sentinel-2 bands are narrower than their predecessors, limiting the atmospheric influence on light waves' reception. The near-infrared band ρ_{8A} is specially designed with a "narrow" width to avoid contamination due to water vapour present in the atmosphere; nevertheless, it can describe the plateau of the vegetation spectral curve in the infrared and is sensitive to iron oxides in the soil. The ρ_1, in the blue window of the spectrum, is necessary for the precise correction of the deformations induced by the atmospheric aerosol.

Table 3.1. Sentinel-2 bands. The 10 m bands are highlighted in blue. In red are the six 20 m bands to be super-resolved. The remaining are 60 m bands.

Bands	B1	B2	B3	B4	B5	B6	B7	B8	B8a	B9	B10	B11	B12
Center wavelength [nm]	443	490	560	665	705	740	783	842	865	945	1380	1610	2190
Bandwidth [nm]	20	65	35	30	15	15	20	115	20	20	30	90	180
Spatial resolution [m]	60	10	10	10	20	20	20	10	20	60	60	20	20

3.1.1 Orbit and Revisit Time

The two satellites operate on the same orbit, inclined 98.62^o (for the equator), sun-synchronous, at an average elevation of 786 km, out of phase by 180.. In this case, to minimize the impact of the clouds, the choice of the sun-synchronous orbit was made. In fact, with this orbit, the shadows are minimized.

The orbit is kept stable by a dedicated propulsion system and by measurements from a dual-frequency receiver. The twin Sentinel-2 satellites are designed to obtain a global World coverage with a revisit time of five days at the equator, and this value decreases due to the overlap between swaths from adjacent orbits, as shown in Fig.3.3.

In addition to the normal acquisition areas, at regular intervals, acquisitions are performed on specific areas, such as Dome-C in Antarctica, to calibrate the instruments.

3.1.2 Instruments

Sentinel-2 satellites are equipped with a multispectral sensor, MSI, a push-broom instrument that acquires data lines perpendicular to the swath and uses the satellite's forward movement to acquire new data strings.

As already mentioned, the incident radiation is deflected by a "beam-splitter" system on various focal planes according to the filters applied. The swath is 290km.

The MSI's radiometric resolution is 12 bits with the ability to acquire

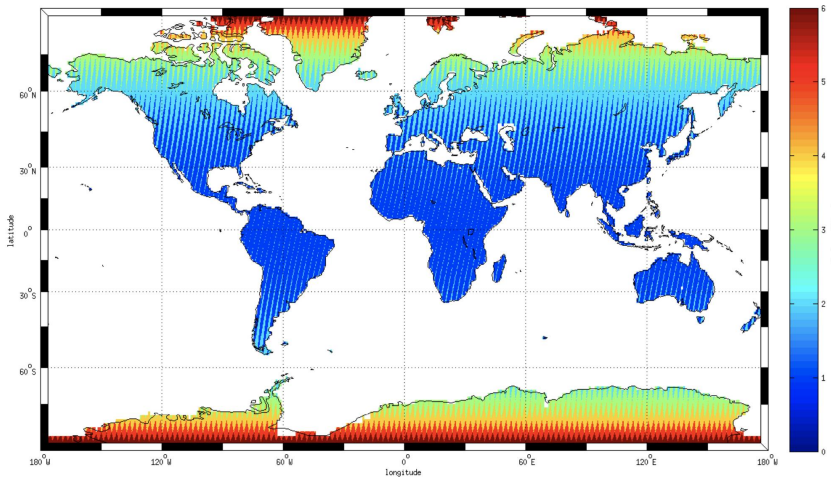


Figure 3.3. Overlap between adjacent orbits (source: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/revisit-coverage)

images in a range of light intensity values ranging from 0 to 4095.

3.1.3 Data

Sentinel-2 data is available in various processing levels described in the following:

- Level 1C provides orthorectified Top-Of-Atmosphere (TOA) reflectance. At this level, cloud and land/water masks are included in the product.

- Level 2A includes orthorectified products with Bottom-Of-Atmosphere (BOA) reflectance information and a "refined" scene classification map (including cloud, vegetation, soils,deserts, water, snow, classes). This kind of products is generated by atmospheric corrections of 1C products (Top-Of-Atmosphere, TOA). At levels 1C and 2A, there are "granules" (also called tiles) of 100 km^2;

Level-0, Level-1A and Level-1B are not distributed to users.

3.2 Super Resolution: Motivations and Challenges

The 10 and 20 m S2 bands are commonly employed for land-cover or water mapping, agriculture or forestry, estimation of biophysical variables, and risk management (floods, forest fires, subsidence, and landslide). In contrast, lower-resolution 60 m bands can be used for monitoring of water vapour, aerosol corrections, pollution monitoring, cirrus clouds estimation, and so forth [44, 50]. Specifically, beyond land-cover classification, S2 images can be useful in such diverse applications as the prediction of growing stock volume in forest ecosystems [51], the estimation of the Leaf Area Index (LAI) [52, 53], the retrieval of canopy chlorophyll content [54], the mapping of the extent of glaciers [55], the water quality monitoring [56], the classification of crop or tree species [57], and the built-up areas detection [58].

The quality, free availability and world-wide coverage make S2 an important tool for current and future Earth observation, and motivates several research teams to propose solutions to super-resolve Sentinel-2 images, rising 20 m and/or 60 m bands up to 10 m resolution. This goal would partially revert the spectral-spatial compromise typical of MS images. Besides, the advantage of using super-resolved S2 images is tested in several applications such as water mapping [], fire detection [59], urban mapping [60], and vegetation monitoring [61]. According to the taxonomy suggested by Lanaras et al. [44] resolution enhancement techniques can be gathered in three main groups: (i) pansharpening and related adaptations; (ii) imaging model inversion; and (iii) machine learning. In addition to this category, it is also worth mentioning the matrix factorization approaches (e.g., [62, 63]), which are more suited to the fusion of low-resolution hyperspectral images with high-resolution multispectral ones. In fact, the spectral

variability becomes a serious concern to be handled carefully through unmixing oriented methodologies [64, 65]. In general, the super resolution methods can be also distinguished on the basis of the information used to increase spatial resolution, as follows: Single Image Super-Resolution (SISR) and Super-Resolution Data Fusion (SRDF) methods. The SISR methods do not use additional information from other sources, and they rely on the original image's spatial features to increase its resolution, such as, for example, a bicubic or a more general polynomial interpolation.. On the other hand, SRDF methods (for instance, pan-sharpening) are based on the idea that the spatial information from other sources is helpful to improve the spatial resolution of the original image [66].

The first category refers to the classical pansharpening, where the super-resolution of low-resolution bands is achieved by injecting spatial information from a single spectrally-overlapping higher-resolution band. This is the case for many remote sensing systems such as Ikonos, QuickBird, GeoEye, WorldView, and so forth. The so-called component substitution methods [67, 68], the multi-resolution analysis approaches [69, 70], or other energy minimization methods [71–73] belong to this category. A recent survey on pansharpening can be found in [39]. Pansharpening methods can also be extended to Sentinel-2 images in different ways, although S2 bands at different resolutions present a weak or negligible spectral overlap, as shown by several works [74–78].

The second group refers to methods that face the super-resolution as an inverse problem under the hypothesis of a known imaging model. The ill-posedness is addressed utilizing additional regularization constraints encoded in a Bayesian or a variational framework. Brodu's super-resolution method [42] separates band-dependent from cross-band spectral information, ensuring the consistency of the "geometry of scene elements" while preserving their overall reflectance. Lanaras et al. [43] adopted an observation model with a per-band point spread functions that accounts for the

convolutional blur, downsampling, and noise. The regularization consists of two parts, a dimensionality reduction that implies a correlation between the bands and a spatially varying, contrast-dependent penalization of the (quadratic) gradients learned from the 10 m bands. In a similar approach, Paris et al. [79] employed a patch-based regularization that promotes self-similarity of the images. The method proceeds hierarchically by first sharpening the 20 m bands and then the coarser 60 m ones.

The last category casts machine learning, and notably deep learning (DL) approaches, which have recently gained great attention from the computer vision and signal processing communities and nearby fields, including remote sensing. In this case, contrarily to the previous categories, no explicit modelling (neither exact nor approximated) of the relationship between high and low-resolution bands is required, since it is directly learned from data. Deep networks allow in principle to mimic very complex nonlinear relationships provided that enough training data are available. In this regard, it is also worth recalling that the pansharpening of multi-resolution images is somewhat related to the unmixing of multi-/hyper-spectral images [64, 65], since in both cases, the general aim is to derive the different spectral responses covered by a single, spatially coarse observation. However, in these two problems, expectations are considerably different: spectral unmixing is a pertinent solution when the interest is focused on surface materials, hence requiring high precision on retrieving the corresponding spectral responses without the need to improve their spatial localization. In pansharpening, the focus is mainly on spatial resolution enhancement preserving the sources' spectral properties. Moreover, no specific information discovery about the radiometry of materials is typically expected. In fact, traditional pansharpening methods try to model spectral diversity, for example, by means of the modulation transfer function of the sensor [69, 70], instead of using radiative transfer models associated with the possible land covers. In any case, from the deep learning perspective, it makes

little difference once the goal is fixed and, more importantly, a sufficiently rich training dataset is provided, as the knowledge (model parameters) will come from experience (data).

The super-resolution approach, proposed in this work, is a data fusion method because we consider to fuse spatial and spectral information from different bands, and to better understand the data fusion nature of the proposed approach, we report a brief taxonomy of the data fusion methods given in [80]. Data fusion processes deal with data from multiple sources to achieve improved information for decision-making. It can be grouped into three main categories:

- *pixel*-level: the pixel values of the source images are jointly processed [81–84];
- *feature*-level: features like lines, regions, keypoints, maps, and so on, are first extracted independently from each source image and subsequently combined to produce higher-level cross-source features which may represent the desired output or be further processed [85–92];
- *decision*-level: the high-level information extracted independently from each source is combined to provide the final outcome, for example using fuzzy logic [93, 94], decision trees [95], Bayesian inference [96], Dempster-Shafer theory [97], and so forth.

In remote sensing, fusion methods can be roughly gathered in the following categories linked to the source date:

- multi-*resolution*: concerns a single sensor with multiple resolution bands. One of the most frequent applications is pansharpening [2, 81, 98], although many other tasks can be solved under a multi-resolution paradigm, such as segmentation [99] or feature extraction [74], to mention a few.

- multi-*temporal*: is one of the most investigated forms of fusion in remote sensing due to the rich information content hidden in the temporal dimension. In particular it can be applied to strictly time-related tasks, like prediction [88], change detection [100–102], co-registration [103], and general-purpose tasks, like segmentation [82], despeckling [104], feature extraction [105–107], which do not necessarily need a joint processing of the temporal sequence but can benefit from it.

- multi-*sensor*: is gaining an ever growing importance due both to the recent deployment of many new satellites, and to the increasing tendency of the community to share data. It also represents the most challenging case because of the several sources of mismatch (temporal, geometrical, spectral, radiometric) among involved data. Like for other categories, a number of typical remote sensing problems can fit this paradigm, such as classification [85, 91, 108–110], coregistration [90], change detection [111], feature estimation [112–115].

- *mixed*: the above cases may also occur jointly, generating mixed situations. For example, hyperspectral and multiresolution images can be fused to produce a spatial-spectral full-resolution datacube [84, 116]. Likewise, low-resolution temporally dense series can be fused with high-resolution but temporally sparse ones to simulate a temporal-spatial full-resolution sequence [117]. The monitoring of forests [96], soil moisture [118], environmental hazards [87], Moreover, other processes can also be carried out effectively by fusing SAR and optical time series. Finally, works that mix all three aspects, resolution, time, and sensor can also be found in the literature [86, 97, 119].

To the best of our knowledge, the first notable example of DL applied to the super-resolution of remote sensing images is the pansharpening con-

volutional neural network (PNN) proposed by Masi et al. [83]. However, this method can not be applied to S2 images without architectural network adaptation and retraining. Examples of convolutional networks conceived for Sentinel-2 are proposed in [44]. Lanaras et al. [44] collected a vast training dataset which has been used to train two much deeper super-resolution networks, one for the 20 m subset of bands and the other for the remaining 60 m bands, achieving state-of-the-art results. In related problems, for example the single-image super-resolution of natural images or other more complex visual tasks such as object recognition or instance segmentation,data sharing has represented a critical enabling factor in these cases allowing researchers to compete with each other or reproduce others' models.

In light of this consideration, we believe that Sentinel-2 is an exciting case thanks to the free access to data that can serve as a playground for a larger scale research activity on remote sensing super-resolution or other tasks. In the same spirit, Lanaras et al. [44] pushed on the power of the data by collecting a relatively large dataset to get good generalization properties. On the other hand, complexity is also an issue that end-users care about. In this regard, the challenge of our contribution is to design and train a relatively small and flexible network capable of achieving competitive results at a reduced cost on the super-resolution of the 20 m S2 bands, exploiting spatial information from the higher-resolution 10 m S2/VNIR bands. Indeed, the proposed network being lightweight, apart from enabling the use of the method on cheaper hardware, allows quickly fine-tuning it when the target data are misaligned from the training data for some reason.

3.3 The shallow Convolutional Neural Network

In the first part of the conducted research, although Sentinel-2 provides six bands at the lowest resolution, the super-resolution of the 20-m Short-Wave Infrared (SWIR) band is only considered because the SWIR has proven to be very useful for several purposes [120]. Here, according to this line of research, we start from the use of a CNN-based approach similar to [83] which have proved to be very successful in pansharpening very high-resolution data like Ikonos, GeoEye, or WorldView.

In the problem at hand we dispose of companion bands which are coregistered with the target and carry complementary spatial information. Therefore, it would be effective to use pansharpening-like methods [2, 39, 83] meant to fuse a low-resolution multispectral (MS) image with a high-resolution single panchromatic (PAN) band, to raise the resolution of the MS to that of the PAN. In our case, the SWIR band would play the MS role, while one or more higher resolution companion bands would replace the single PAN. A deep learning approach has already been proposed in the fusion of multispectral and hyperspectral images [121].

In this first part, we have developed three CNN models corresponding to three different input combinations. In the simplest case (M1), we feed the network only with the objective band ρ_{11}, without high-resolution guiding bands (pure super-resolution). Then, we developed pansharpening-like algorithm (M2) feeding the network with also the most correlated band (specifically, near infra-red, NIR) to the target band, as shown in Tab.3.2. Eventually, we also extended to all 10-m resolution bands (M5).

The development of a deep learning super-resolution method suited for a given remote sensing imagery involves at least three key steps, with some iterations among them:

a. Selection/generation of a suitable dataset for training, validation and test;

Table 3.2. Estimation of Correlation Index between all the analyzed bands at 10-m (blue) and at 20-m (green). The grey intensity in the cells corresponds to different level of correlation between all bands.

ρ	$\rho_{\mathbf{2}}$	$\rho_{\mathbf{3}}$	$\rho_{\mathbf{4}}$	$\rho_{\mathbf{5}}$	$\rho_{\mathbf{6}}$	$\rho_{\mathbf{7}}$	$\rho_{\mathbf{8}}$	$\rho_{\mathbf{8A}}$	$\rho_{\mathbf{11}}$	$\rho_{\mathbf{12}}$
$\rho_{\mathbf{2}}$	1	0.9632	0.8802	0.7734	0.5436	0.4934	0.4857	0.4659	0.6027	0.6628
$\rho_{\mathbf{3}}$	0.9632	1	0.9417	0.8793	0.691	0.6437	0.6352	0.6147	0.7262	0.7626
$\rho_{\mathbf{4}}$	0.8802	0.9417	1	0.953	0.7727	0.7311	0.7252	0.7134	0.8434	0.8713
$\rho_{\mathbf{5}}$	0.7734	0.8793	0.953	1	0.8996	0.8657	0.849	0.852	0.9198	0.9119
$\rho_{\mathbf{6}}$	0.5436	0.691	0.7727	0.8996	1	0.9946	0.9822	0.9875	0.9036	0.8311
$\rho_{\mathbf{7}}$	0.4934	0.6437	0.7311	0.8657	0.9946	1	0.9885	0.9951	0.8869	0.8035
$\rho_{\mathbf{8}}$	0.4857	0.6352	0.7252	0.849	0.9822	0.9885	1	0.9886	0.8789	0.7913
$\rho_{\mathbf{8A}}$	0.4659	0.6147	0.7134	0.852	0.9875	0.9951	0.9886	1	0.8906	0.7993
$\rho_{\mathbf{11}}$	0.6027	0.7262	0.8434	0.9198	0.9036	0.8869	0.8789	0.8906	1	0.9715
$\rho_{\mathbf{12}}$	0.6628	0.7626	0.8713	0.9119	0.8311	0.8035	0.7913	0.7993	0.9715	1

b. Design and implementation of one or more DL models;

c. Training and validation of the models (b) using the selected dataset (a).

By following this rationale, for ease of presentation, in this section, we first present the datasets and their preprocessing (a), and then we describe design (b) and training (c) of the proposed model.

3.3.1 Datasets and Labels Generation

Regardless of its complexity and capacity, a target deep learning model remains data-driven machinery whose ultimate behaviour heavily depends on the training dataset, notably on its representativeness of real-world cases. Hence, we provide here detailed information about our datasets and their preprocessing.

Except for some cases where unsupervised learning strategies can be applied, a sufficiently large dataset containing input-output examples is usually necessary to train a deep learning model. This is also the case for super-resolution or pansharpening. In our case, we fused 10 m ($\mathbf{z}$) with 20 m ($\mathbf{x}$) bands to enhance the resolution of $\mathbf{x}$ by a factor of 2 (resolution

ratio). This means that we should have examples of the kind $((\mathbf{x}, \mathbf{z});)$, being the desired (super-resolved) output corresponding to the composite input instance $(\mathbf{x}, \mathbf{z})$. In rare cases, one can rely on referenced data, for example thanks to ad hoc missions to collect full-resolution data to be used as a reference, whereas in most cases, referenced samples are unavailable.

Under the latter assumption, many deep learning solutions for super-resolution or pansharpening have been developed (e.g., [1, 44, 83, 122–124]) through a proper schedule for generating referenced training samples from the same no-reference input dataset. It consists of a resolution downgrade process that each input band undergoes and involves two steps:

(i) band-wise low-pass filtering; and

(ii) uniform $R \times R$ spatial subsampling, being R the target super-resolution factor.

This is aimed to shift the problem from the original *full*-resolution domain to a *reduced*-resolution domain. In our case, $R = 2$ while the two original input components, $\mathbf{x}$ and $\mathbf{z}$, will be transformed in corresponding variables $\mathbf{x}_{\downarrow}$ and $\mathbf{z}_{\downarrow}$, respectively, lying in the reduced-resolution space, with associated reference $\mathbf{r}_{\downarrow}$ trivially given by $\mathbf{r}_{\downarrow} = \mathbf{x}$. How to filter the bands before subsampling is an open question. Lanaras et al. [44] pointed out that with deep learning, one does not need to specify sensor characteristics (for instance, spectral response functions) since sensor properties are implicit in the training data. Contrarily, Masi et al. [83] asserted that the resolution scaling should be done accounting for the sensor Modulation Transfer Function (MTF) in order to properly generalize when applied at full resolution. Such a position follows the same rationale of the so-called Wald's protocol, a procedure commonly used for generating referenced data for objective comparison of pansharpening methods [39]. Actually, this controversial point cannot be resolved by looking at the performances in the reduced-resolution space since a network learns from

training data the due relationship whatever preprocessing has been performed on the input data. On the other hand, in full-resolution domain, no objective measures can be used because of the lacking referenced test data. In this work, we follow the approach proposed in [83] making use of sensor MTF. The process for the generation of a training sample is summarized in Fig. 3.4. Each band undergoes different low-pass filtering before being downsampled, whose cut-off frequency is related to the sensor MTF characteristics. Additional details can be found in [125].

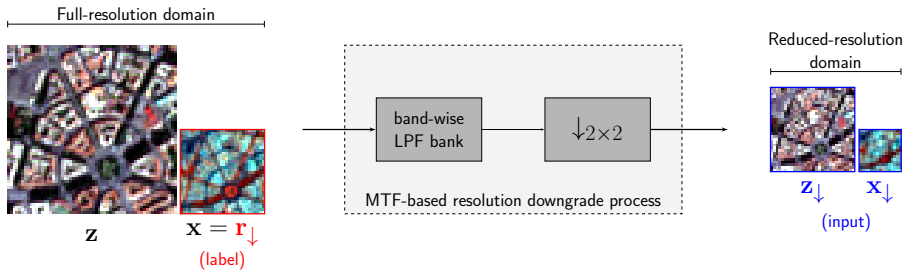


Figure 3.4. Generation of a training sample $((\mathbf{x}_{\downarrow}, \mathbf{z}_{\downarrow}); \mathbf{r}_{\downarrow})$ using Wald's protocol. All images are shown in false-color RGB using subsets of bands for ease of presentation. Each band is low-lass filtered with a different cut-off frequency according with the sensor MTF characteristics.

Another rather critical issue is the training dataset selection, as it impacts the trained models' capability to generalize well on unseen data. In the computer vision domain, a huge effort has been devoted to collecting vast datasets to support the development of deep learning solutions for such diverse problems as classification, detection, semantic segmentation, tracking video and so forth (notable examples are ImageNet and Kitty datasets). Instead, in remote sensing, there are no examples of datasets that are as large as ImageNet or Kitty. This is due to several obstacles, among which the cost of the data and the related labelling, which requires domain experts and the data sharing policy usually adopted in the past years by the remote sensing community. Luckily, for super-resolution, one

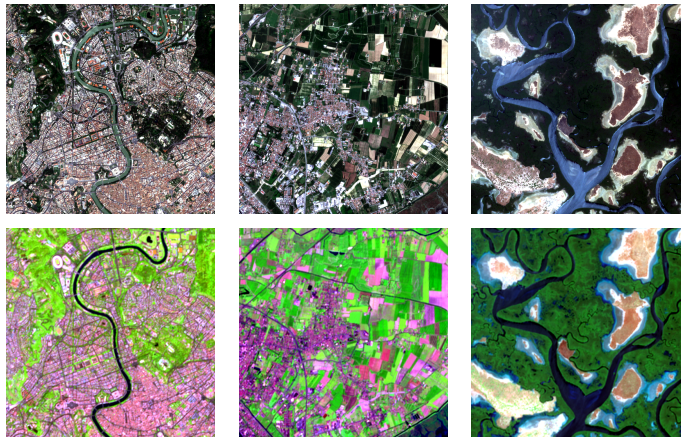
can at least rely on the above-described fully-automated resolution downgrading strategy to avoid labelling costs. Due to the scarcity of data, most deep-learning models for resolution enhancement applied to remote sensing have been trained on a relatively small dataset, possibly taken from a few large images, from which non-overlapping sets for training, validation and testing are singled out [83, 124, 126]. The generalization limits of a pansharpening model trained on too few data have been stressed in [123] for both cross-image and cross-sensor scenarios, where a fine-tuning stage has been proposed to cope with the scarcity of data. In particular, it was shown that for a relatively small CNN that integrates a residual learning module, a few training iterations (fine-tuning) on the reduced-resolution version of the target image allow quickly recovering the performance loss due to the misalignment between training and test sets. For Sentinel-2 imagery, thanks to the free access guaranteed by the Copernicus program, larger and more representative datasets can be collected, as done by Lanaras et al. [44], aiming for a roughly even distribution on the globe and variety in terms of climate zone, land-cover and biome type.

In this study, we opted for a lighter and flexible solution with a relatively small number of parameters to learn and a (pre-)training dataset of relatively limited size. This choice is motivated by the experimental observation that in the actual application, the tuning of the parameters is still recommendable even if larger datasets have been used in training, making appealing lighter solutions that can be quickly tuned if needed.

To compare our results with those presented by Lanaras et al. [44], we decided to keep their setting by using Sentinel-2 data without atmospheric correction (L1C product) for our experiments. For training and validation, we referred to three scenes (see Fig. 3.5), corresponding to different environmental contexts: Venice, Rome, and Geba River.

In particular, we randomly cropped 18,996 square tiles of size 33×33 (at 20 m resolution) from the three selected scenes to be used for train-

ing (15,198) and validation (3898). Besides, we have chosen four more scenes for testing, namely Athens, Tokyo, Addis Abeba, and Sydney, which present different characteristics, allowing for more robust validation of the proposed model. From such sites, we singled out three 512 × 512 crops at 10 m resolution, for a total of twelve test samples.



Rome Venice Geba River

Figure 3.5. Examples of images used for training. (Top row) RGB-composite images using 10 m bands B4(R), B3(G) and B2(B), subset of **z**; and (Bottom row) corresponding 20 m RGB subset of **x**, using B5(R), B8a(G) and B11(B).

3.4 Convolutional neural networks

Before describing the specific solutions for S2 super-resolution, in this Section we provide some basic notions and terminology about Convolutional Neural Networks.

In the last few years, CNNs have been successfully applied to many classical image processing problems, such as denoising [127], segmentation [128], object detection [129, 130], change detection [100], classification [3, 92, 131, 132]. The main strengths of CNNs are (i) an extreme versatility

that allows them to approximate any sort of linear or non linear transformation, including scaling or hard thresholding; (ii) no need to design handcrafted filters, replaced by machine learning; (iii) high-speed processing, thanks to parallel computing. On the downside, for proper training, CNNs require the availability of a large amount of data with ground-truth (examples from which the CNNs learn some complex non-linear function between input and output). In our specific case, the data availability is not the main problem, given the quantity of cloud-free Sentinel-2 time-series that can potentially be downloaded from the web repositories. However, using large datasets has a cost in terms of complexity, and may lead to unreasonably long training times. A CNN is a chain[1] of different layers, like convolution, non-linearities, pooling, deconvolution. In this work, only convolutional layers interleaved with non-linear activations are employed.

The generic l-th convolutional layer, with N-band input $\mathbf{x}^{(l)}$, yields an M-band stack $\mathbf{z}^{(l)}$ computed as

$$\mathbf{z}^{(l)} = \mathbf{w}^{(l)} * \mathbf{x}^{(l)} + \mathbf{b}^{(l)},$$

whose m-th component can be written in terms of ordinary 2D convolutions

$$\mathbf{z}^{(l)}(m,\cdot,\cdot) = \sum_{n=1}^{N} \mathbf{w}^{(l)}(m,n,\cdot,\cdot) * \mathbf{x}^{(l)}(n,\cdot,\cdot) + \mathbf{b}^{(l)}(m).$$

The tensor $\mathbf{w}$ is a set of M convolutional $N \times (K \times K)$ kernels, with a $K \times K$ spatial support (receptive field), while $\mathbf{b}$ is a M-vector bias. These parameters, compactly, $\Phi_l \triangleq \left(\mathbf{w}^{(l)}, \mathbf{b}^{(l)}\right)$, are learnt during the training phase. If the convolution is followed by a point-wise activation function $g_l(\cdot)$, then, the overall layer output is given by

$$\mathbf{y}^{(l)} = g_l(\mathbf{z}^{(l)}) = g_l(\mathbf{w}^{(l)} * \mathbf{x}^{(l)} + \mathbf{b}^{(l)}) \triangleq f_l(\mathbf{x}^{(l)}, \Phi_l). \tag{3.1}$$

[1] Parallels, loops or other combinations are also possible.

Due to the good convergence properties it ensures [3], the Rectified Linear Unit (ReLU), defined as $g(\cdot) \triangleq \max(0,\ \cdot)$, is a typical activation function of choice for input or hidden layers.

Assuming a simple L-layer cascade architecture, the overall processing will be

$$f(\mathbf{x}, \Phi) = f_L(f_{L-1}(\dots f_1(\mathbf{x}, \Phi_1), \dots, \Phi_{L-1}), \Phi_L), \tag{3.2}$$

where $\Phi \triangleq (\Phi_1, \dots, \Phi_L)$ is the whole set of parameters to learn. In this chain, each layer l provides a set of so-called *feature maps*, $\mathbf{y}^{(l)}$, which activate on local cues in the early stages (small l), to become more and more representative of abstract and global phenomena in subsequent ones (large l).

3.4.1 Proposed CNN-based method

For all super-resolution proposals that we present, the use of a relatively shallow CNN architectures is considered. In this case, the proposed CNN is a cascade of $L = 3$ convolutional layers interleaved by Rectified Linear Unit (ReLU) activations that ensure fast convergence of the training process [3]. Let $\mathbf{x} \triangleq (\rho_{11}, \rho^{\mathrm{HR}_1}, \dots, \rho^{\mathrm{HR}_B})$ be the input to the network[2], and $\mathbf{y} \triangleq \widehat{\rho}_{11}$ be the network output that is the sharpened SWIR band. The relationship between input and output of $l-th$ generic layer can be schematized as follows:

$$\mathbf{y}^{(l)} \triangleq f_l(\mathbf{x}^{(l)}, \Phi_l) = \begin{cases} \max(0, \mathbf{w}^{(l)} * \mathbf{x}^{(l)} + \mathbf{b}^{(l)}), & l < L \\ \mathbf{w}^{(l)} * \mathbf{x}^{(l)} + \mathbf{b}^{(l)}, & l = L \end{cases}$$

[2]In CNN-based super-resolution or pansharpening, it is custom to preliminary upsample the lower resolution components in input with a standard interpolator, e.g. bicubic, to align the input stack. For the sake of simplicity, we keep the notation ρ_{11} for this interpolated band which actually feeds the net.

whose concatenation gives the overall CNN function

$$\mathbf{y} = f(\mathbf{x}, \Phi) = f_L(f_{L-1}(\dots f_1(\mathbf{x}, \Phi_1), \dots, \Phi_{L-1}), \Phi_L)$$

where $\mathbf{x} = \mathbf{x}^{(1)}$, $\mathbf{y} = \mathbf{y}^{(L)}$, and $\Phi \triangleq (\Phi_1, \dots, \Phi_L)$ is the whole set of parameters to learn.

The network hyper-parameters are summarized in Tab.3.3. Model M1 corresponds to the "pure" super-resolution of ρ_{11}, without using any additional "guiding" band. M2 uses only the most correlated band as a guide, while M5 uses all available high-resolution bands. The last column reported the scope of the overall network function readily obtained as the cumulative convolutional spread, as the nonlinear ReLU is a punctual operator that does not increase the scope. These hyperparameters were selected among several alternative configurations as the optimal choice in terms of complexity and accuracy. It is worth notice that the patch size is relatively small (17×17) compared to that of the CNN pansharpening method [83]. This should not surprise as [83] is conceived for a double super-resolution ratio, requiring in principle major efforts to get the same performances.

Table 3.3. Hyper-parameters of the proposed networks.

Model	**input bands**	**kernel size (# features)**			**Interaction range**
		$l = 1$	$l = 2$	$l = 3$	
M1	ρ_{11}	3×3 (48)	3×3 (32)	3×3 (1), $\widehat{\rho}_{11}$	7×7
M2	ρ_{11},	3×3 (48)	3×3 (32)	3×3 (1), $\widehat{\rho}_{11}$	7×7
M5	ρ_{11}, ρ_2, ρ_3, ρ_4	3×3 (48)	3×3 (32)	3×3 (1), $\widehat{\rho}_{11}$	7×7

Training

In order to train the network's parameters, Φ, a sufficiently large number

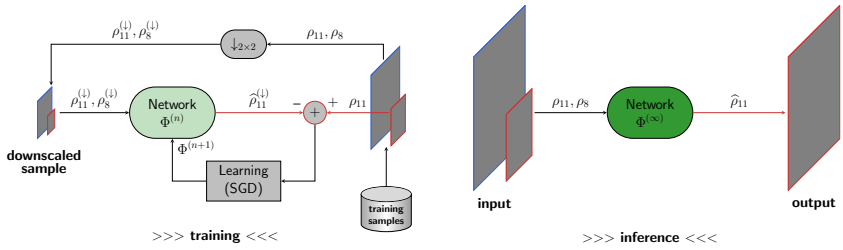


Figure 3.6. Top-level training (left) and inference (right) workflows for model M2.

of input-output examples and the choice of a suitable cost function are required. In this context we used the Stochastic Gradient Descent (SGD) with momentum that is adopted for the first time in neural networks training in [133].

To generate examples for training we considered Wald's protocol as approached in [83]. The training based on Wald's Protocol consists of using inputs properly downsampled PAN-MS pairs and taking as corresponding output the original MS. In our method, we considered a simple adaptation that consist in replacing the PAN component with the high-resolution bands. A high-level description of the training process for model M2 is given in Fig. 3.6 (left). The downgraded bands are marked with a downward arrow superscript. Once the network has reached a convergence condition, the current parameters $\Phi^{(\infty)}$ are frozen and ready to be used to perform the super-resolution of the target images (right part of Fig. 3.6). The training phase is carried out offline once and takes a few hours using GPU cards, while the test can be done in real-time. Several losses can be found in the literature, like L_n norms, cross-entropy, negative log-likelihood. The choice depends on the domain of the output, and affects the convergence properties of the networks [134]. Our experiments have shown the L_1-norm (Eq. 3.3) to be more effective than other options for training in case of small errors (i.e. $<< 1$), therefore we keep this choice which proved to be effective also in other generative problems.

Specifically, the loss is computed by averaging the difference between the target and the CNN's output over a suitable set (mini-batch) of training examples at each updating step of the SGD process:

$$L(\Phi^{(n)}) = \mathrm{E}\left[\left\|\rho_{11} - \widehat{\rho}_{11}^{(\downarrow)}(\Phi^{(n)})\right\|_1\right].$$

A whole scan of the training set is called an *epoch*, and training a deep network may require dozens of epochs, for simpler problems like handwritten character recognition [135], to thousands of epochs for complex classification tasks [3]. Accuracy and speed of training depend on both the initialization of Φ and the setting of hyperparameters like learning rate α and momentum μ, with α being to most critical, impacting heavily on stability and convergence time. In particular we have experimentally found optimal values for these parameters which are $\alpha = 0.5 \cdot 10^{-3}$ and $\mu = 0.9$.

3.5 Advances

In the following, we propose an improved version of our method [136] with the aforementioned integrations. The proposed solution is summarized in Fig.3.7. At the core is a three-layer CNN block whose internal architecture is nearly the same as the CNN described in the previous section, the main difference being the output shape which counts here six bands rather than just the single SWIR. Further, along with the 10-m bands, we included all 20-m Sentinel-2 images in the input stack, improving the previous work where we only considered one MS band in input and one in output.

For the sake of clarity, let us first recall the main characteristics of the 13 spectral bands of Sentinel-2, gathered in Tab. 3.1, and clarify symbols and notations that are used in the following with the help of Tab. 3.4.

First, it can be noticed the additional batch normalization (BN) layer,

Table 3.4. Notations and symbols.

Symbol	Meaning
$\mathbf{x}$	Stack of six S2 spectral bands (B5, B6, B7, B8A, B11, B12) to be super-resolved.
$\mathbf{z}$	Stack of four high-resolution S2 bands (B2, B3, B4, B8).
$\mathbf{x}^{\mathrm{hp}}$, $\mathbf{z}^{\mathrm{hp}}$	High-pass filtered versions of $\mathbf{x}$ and $\mathbf{z}$, respectively.
$\hat{\mathbf{x}}$	Super-resolved version of $\mathbf{x}$.
	Full-resolution *reference* (also referred to as *ground truth* or *label*), usually unavailable.
$x, \hat{x}, r$	generic band of $\mathbf{x}, \hat{\mathbf{x}},$, respectively.
$\widetilde{\mathbf{x}}, \widetilde{x}, \widetilde{\mathbf{x}}^{\mathrm{hp}}$	Upsampled (via bicubic) versions of $\mathbf{x}, x, \mathbf{x}^{\mathrm{hp}}$, respectively.
$\overline{z}$	Single (average) band of $\mathbf{z}$.
$\mathbf{x}_{\downarrow}$, $\mathbf{z}_{\downarrow}$, $\mathbf{r}_{\downarrow}$,...	Reduced-resolution domain variables associated with $\mathbf{x}$, $\mathbf{z}$, ,..., respectively. Whenever unambiguous subscript $\downarrow$ will be dropped.

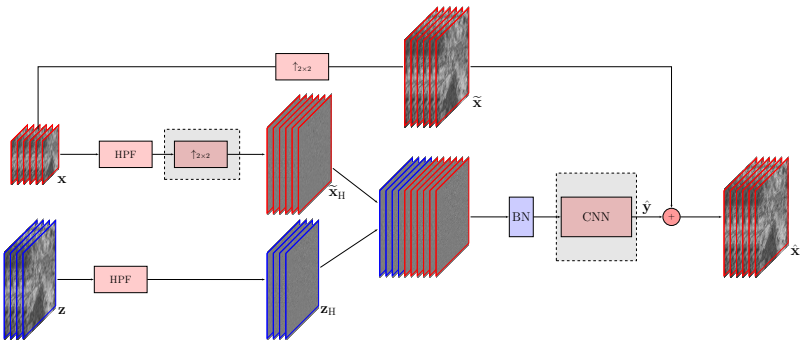


Figure 3.7. Top-level workflow of the proposed super-resolution method for 20-m bands of Sentinel-2. Only dashed boxes are used in [136].

which processes the input stack $(\mathbf{z}^{\mathrm{hp}}, \mathbf{x}^{\mathrm{hp}})$ that feeds the network in order to make the learning process robust concerning the statistical fluctuations of the training dataset [137].

The BN block takes a concatenation stack as input which comprises both the high pass filtered (HPF) 10-m resolution bands of Sentinel-2, $\mathbf{z}^{\mathrm{hp}}$, and the upscaled (via bicubic interpolation) HPF version of the target 20-m resolution bands, $\mathbf{x}^{\mathrm{hp}}$. The use of HPF bands in place of the original ones has been introduced in the context of pansharpening in [126] based on the intuition that low spatial frequencies do not carry relevant information about spatial details. Therefore they can be neglected, eventually facilitating the training process.

Another relevant trick generally useful when dealing with deep con-

volutional networks, and effective for super-resolution and pansharpening as well, is the use of residual learning [138], which is here implemented through the skip connection (on top) that directly links the input (upscaled) to the output ($\hat{\mathbf{x}} = \widetilde{\mathbf{x}} + \hat{\mathbf{y}}$). Intuitively speaking, since the low-frequency content of the desired output is already comprised in the low-resolution input bands, it is sufficient for the network to learn only how to predict the complementary detail component $\hat{\mathbf{y}}$ to be combined with the input $\widetilde{\mathbf{x}}$ in order to get the full-resolution product. The practical consequence of this is that the network learns much faster, which is very useful to finetune the network on the dataset.

3.5.1 Proposed Multi-loss Function

The last improvements of the proposed solution take inspiration from two state-of-the-art CNN models for pansharpening, namely PanNet [126] and the target-adaptive version [123] of PNN [83], both conceived for very high-resolution sensors such as Ikonos or WorldView-2/3. Both methods rely on a residual learning scheme, while the main differences concern loss function, input preprocessing, and overall network backbone shape and size.

Fig. 3.8 shows the top-level flowchart of the proposed method. As we deal with Sentinel-2 images, differently from [126] and [123], we have 10 input bands, six lower-resolution ones ($\mathbf{x}$), to be super-resolved, plus four higher-resolution bands ($\mathbf{z}$). Let us preliminarily point out that we train a single (relatively small) network for each band x to be super-resolved, as represented at the output in Fig. 3.8. However, the deterministic preprocessing bounded by the dashed box is a shared part, while the core CNN, with fixed hyper-parameters, changes from one band to another to be super-resolved. This choice presents two main advantages. The first is that whenever users need to super-resolve only a specific band, they can use a lighter solution with computational advantages. The second reason

is related to the experimental observation that training separately the six networks allow reaching the desired loss levels more quickly than using a single wider network. This feature is particularly desirable if users need to fine-tune the network on their dataset. Turning back to the workflow, observe that both input subsets, $\mathbf{x}$ and $\mathbf{z}$, are high-pass filtered (HPF) as also done in the previous section. This operation relies on the intuition that the missing details that the network is asked to recover lie in the input image's high-frequency range. Next, the HPF component $\mathbf{x}^{\text{hp}}$ is upsampled ($R \times R$) using a standard bicubic interpolation, yielding $\widetilde{\mathbf{x}}^{\text{hp}}$, in order to match the size of $\mathbf{z}^{\text{hp}}$ with which to be concatenated before feeding the actual CNN. The single-band CNN output $f_\Phi(\mathbf{x}, \mathbf{z})$ is therefore combined with the upsampled target band $\widetilde{x}$ to provide its super-resolved version $\hat{x} = \widetilde{x} + f_\Phi(\mathbf{x}, \mathbf{z})$. This last residual learning combination, obtained through a skip connection that retrieves the low-resolution content of $\hat{x}$ directly from the input, is proven to speed-up the learning process [126].

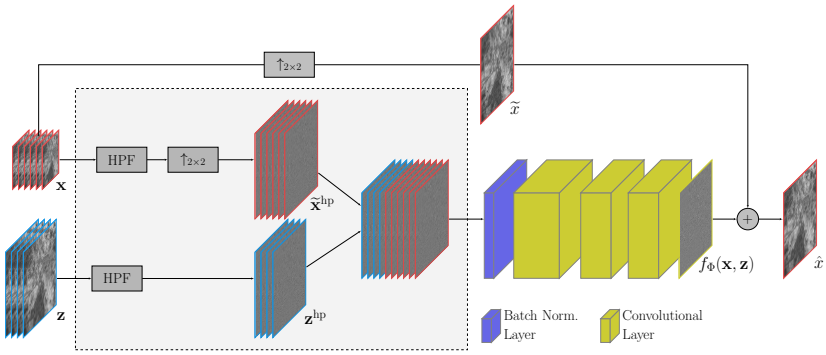


Figure 3.8. Top-level workflow for the super-resolution of any 20 m band of Sentinel-2. The dashed box gathers the shared processing which is the same for all predictors.

The CNN architecture is shallower than to PanNet [126], using just four convolutional layers. PanNet uses ten layers, each singling out 32 features (except for the output layer). However, we add an additional convolu-

tional layer concerning the previous work. Moreover, a batch normalization layer operating on the input stack precedes the convolutional ones. This has proven to make the learning process robust concerning the statistical fluctuations of the training dataset [137]. In Table 3.5, the network hyper-parameters of the convolutional layers are summarized.

Table 3.5. Hyper-parameters of the convolutional layers for the proposed CNN model.

	ConvLayer 1	**ConvLayer 2**	**ConvLayer 3**	**ConvLayer 4**
Input channels	10	48	32	32
Spatial support	3×3	3×3	3×3	3×3
Output channels	48	32	32	1
Activation	ReLU	ReLU	ReLU	tanh

Training

Once the training dataset and model are fixed, a suitable loss function to be minimized needs to be defined for the learning process. L_2 or L_1 norms are typical choices [1, 44, 83] due to their simplicity and robustness, with the latter being more effective to speed-up the training, as observed in [44, 123]. However, these measures do not account for structural consistency as they are computed on a pixel-wise basis and, therefore, assess only spectral dissimilarity. To cope with this limitation, an option is to resort to a so-called *perceptual* loss [139], which is an indirect error measurement performed in a suitable feature space generated with a dedicated CNN. In [126], structural consistency is enforced by working directly on detail (HPF) bands. In the proposed solution, in addition to the use of HPF components, we also define a combined loss that explicitly accounts for spectral and structural consistency. In particular, inspired by the variational approach [140], we make use of the following loss function:

$$\mathscr{L} = \lambda_1 \mathscr{L}_{\text{Spec}} + \lambda_2 \mathscr{L}_{\text{Struct}} + \lambda_3 \mathscr{L}_{\text{Reg}} \tag{3.3}$$

where three terms, corresponding to fidelity, or spectral consistency ($\mathscr{L}_{\text{Spec}}$), structural consistency ($\mathscr{L}_{\text{Struct}}$) and regularity ($\mathscr{L}_{\text{Reg}}$), are linearly combined. The weights were tuned experimentally using the validation set as $\lambda_1 = 1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$.

By following the intuition proposed in [44, 123], we decided to base the fidelity term on the L_1 norm, that is

$$\begin{aligned} \mathscr{L}_{\text{Spec}} &= \left\{ \left\| \hat{x}_{\downarrow} - r_{\downarrow} \right\|_1 \right\} \\ &= \left\{ \left\| f_{\Phi}(\mathbf{x}_{\downarrow}, \mathbf{z}_{\downarrow}) + \widetilde{x}_{\downarrow} - r_{\downarrow} \right\|_1 \right\} \end{aligned}$$

where the expectation $\{\cdot\}$ is estimated on the reduced-resolution training minibatches during the gradient descent procedure. $f_{\Phi}(\cdot)$ stands for the CNN function (including preprocessing) whose parameters to learn are collectively indicated with Φ. This loss term, as well as the other two, refers to a single band ($x_{\downarrow}$) super-resolution whose ground-truth is $r_{\downarrow} = x$. As the training is performed in the reduced-resolution domain, in the reminder on this section, we drop the subscript $\downarrow$ for the sake of simplicity.

The structural consistency term is given by

$$\mathscr{L}_{\text{Struct}} = \left\{ \sum_{i=1}^{4} \left\| G_i(\hat{x} - r) \right\|_{1/2} \right\},$$

where the operator $G = (G_1, \ldots, G_4)$ generalizes the gradient operator including derivatives in the diagonal directions that help to improve quality, as shown in [140]. It has been shown that the gradient distribution for real-world images better fits with a heavy-tailed distribution such as a hyper-Laplacian ($p(x) \propto e^{-k|x|^p}$, $0 < p < 1$). Accordingly, we make use of a L_p-norm with $p = 1/2$, which we believe can be more effective [140]. This term penalizes discontinuities in the super-resolved band $\hat{x}$ if they do not occur, with the same orientation, in the panchromatic band. As the dynamics of these discontinuities are different, an additional prior regular-

ization term that penalizes the total variation of $\hat{x}$ helps to avoid unstable behaviors:

$$\mathscr{L}_{\text{Reg}} = \{\|\nabla \hat{x}\|_1\} = \{\|\nabla f_\Phi(\mathbf{x}, \mathbf{z}) + \nabla \widetilde{x}\|_1\}\,.$$

Eventually, the network parameters were (pre-)trained by means of the Adaptive Moment Estimation (ADAM) optimization algorithm [141] applied to the above-defined overall loss (Equation (3.3)). In particular, we have set the ADAM default hyper-parameters, which are learning rate, $\eta = 0.002$, and decay rate of the first and second moments, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively [142]. The training was run for 200 epochs, being an epoch a single pass over all minibatches (118) in which the training set has been split, with each minibatch composed of 128 33×33 input–output samples.

This page intentionally left blank.

Chapter 4

Deep Learning for Sentinel 1 Segmentation

The objective of this section is to propose a set of solutions to estimate a target optical feature (or land use conditions) at a given date from multispectral images acquired at close dates, and/or from the temporally closest SAR image. Such different solutions also reflect the different operating conditions found in practice. Firstly, deep learning (Convolutional Neural Networks-based) algorithms are investigated in generative problem, which means the reconstruction of the most common multispectral vegetation index, the Normalized Difference Vegetation Index (the beforementioned NDVI). The reconstruction of the NDVI is possible because just as the NDVI index provides chlorophyll information, the radar is sensitive to variations in vegetation scattering. These phenomena are correlated in the vegetation monitoring, and thus our CNN-based method is consistent in the investigation context. The discussion is different when information about the mineral content or the water quality is required since there are no physical connections, and so this would make any regression and index estimation algorithm impossible.Then, we propose a method to obtain a

segmentation map through CNNs feeded by Sentinel-1 input. In particular, in [143], innovative DL methodologies to estimate semantic segmentation maps from the L2A product of Sentinel-2 at a given date from the dual-polarizations Interferometric Wide swath mode (IW) data are presented. Radar can be collected with both along ascending and descending orbit. Specifically, the purpose is to identify three classes (vegetation, water e bare soil) from the Sentinel-1 information. Summarizing, first, the ability of a shallow CNN architecture to extrapolate the NDVI is evaluated [144]. Then, the CNN-based solution is improved thanks to the innovations listed below [143]:

a. the objective changes from a generative solution to a discriminative one, that means to recognize the classes mentioned earlier;

b. the input stack only includes the Sentinel-1 polarizations, and at single target date. This is a crucial point because in the previous solution we needed the same training conditions for test phase, for instance contemporary acquisitions of S1 and S2;

d. a set of more complex deep learning architectures is compared;

e. the scene classification map provided by the L2A product that included many invalid pixels is used as target.

Then, further innovations have been introduced in [145]:

a. a novel deep learning architecture, called W-Net because of its W-shaped structure;

b. the deep learning data fusion approach to the case of multi-temporal data, in particular in this case the two closest SAR images are not associated with the closest free-cloud Multispectral images, as in the first work, but is associated to the target SAR image;

c. a different segmentation map as a reference, obtained by using the L2A product. This segmentation map is obtained by means of a thresholds-based classification from some spectral indices.

The rest of the Chapter is organized as follows. First, more details about the Sentinel 1 dataset's characteristics are described. Then, an "ad hoc" architecture is proposed.

4.1 Sentinel-1

The Sentinel-1 mission consists of a constellation of two satellites equipped with radar instrumentation. The Sentinel-1A satellite was launched on April 3, 2014, and the S1-B satellite on April 25, 2016; both arrived in orbit thanks to a Russian Soyuz rocket that left the ESA launch centre in French Guiana. The specific characteristics of the Sentinel-1 data, used in these works, are provided in Tab. 4.1.

Table 4.1. General information about the considered SAR data.

Specifications	**Sentinel-1A Data**
Acquisition orbit	Descending
Imaging mode	IW
Imaging frequency	C-band (5.4 GHz)
Polarization	VV, VH
Data Product	Level-1 GRDH
Spatial Resolution	10-m

This first series of ESA satellites cover almost all the general application fields proposed by the Copernicus program. In particular, they extend to oceans and ice monitoring, land cover/use monitoring, and so forth. Furthermore, the Sentinel 1 allows environmental applications such as the study of subsidence or the monitoring of landslides or volcanoes with millimetre accuracy and is particularly suitable for rapid response in environmental emergencies due to its ability to penetrate beyond the clouds.

Figure 4.1. Sentinel-1 satellite.

4.1.1 Orbit and geographical coverage

The two satellites share the same solar-synchronous quasi-polar orbit out of phase by 180 degrees; both satellites operating a cycle that lasts six days (12 days for a single satellite). Each cycle consists of 175 orbits for each satellite. Since the spacing between the satellites' tracks on the ground varies with the latitude, the revisiting time is much lower at high latitudes rather than at the equator.

The satellites orbit 693 km from the Earth. The positioning must be exact, especially for interferometric applications. The orbit is maintained with variations of less than $\pm$ 50m concerning the predetermined orbit thanks to the control instruments. From these variations, it is possible to define an "orbital tube" of a 50 m radius around the nominal orbit.

4.1.2 Instruments

Sentinel-1 satellites are equipped with a C-band synthetic aperture radar. Thanks to a transmission system (interchangeable from H to V) and two parallel receiver systems for H and V polarizations, this satellite

supports dual-polarization operations (HH + HV, VV + VH).

There are four acquisition methods of which the main characteristics are shown in Tab. 4.2. The different acquisition modes differ in the angle

Table 4.2. The main characteristics of the Sentinel-1 acquisition modes

Mode	Incidence Angle	Spatial Resolution	Swath	Polarizations
Stripmode	20°-45°	5 × 5	80 km	HH+HV, VH+VV,HH, VV
Interferometric Wide Swath	29°-46°	5 × 20	250 km	HH+HV, VH+VV,HH, VV
Extra Wide Swath	19°-47°	20 × 40	400 km	HH+HV, VH+VV,HH, VV
Wave	22°-38°	5 × 5	20 × 20 km	HH, VV

of the ray's incidence to the ground, spatial resolution, polarization and swath. With this last term, we mean the width of the portion of the earth's surface acquired by the sensor when the satellite rotates around the Earth (ground trace of the array sensor). The main acquisition modes are the Interferometric Wide swath (IW) and the Wave mode (WV). The first is the default mode for acquiring land and satisfies almost all service needs, and the second is used at sea.

The IW mode consists of data acquisition with a swath of 250 km divided into three sub-swaths, thus producing an image for each sub-swath and one for each polarization channel. For a total of three images in single or double polarization.

The spatial resolution is 5 × 20 m.

The images can then be merged with a merging operation into a single one covering a large area to ensure each sub-swath has a minimum overlap area (2 km).

On the other hand, the WV mode provides a series of images (single polarization VV or HH) at alternating angles of attack and a distance of 100km. Each image covers 20 × 20 km of the earth's surface with a resolution of 5 × 5 m. There are two other acquisition modes, as reported in Tab. 4.2, the Stripmode (SM) and Extra Wide Swath (EW) mode, which support normal operations. The first generates a series of six images with a resolution of 5 × 5 m and a swath of 80 km, the second works in a

similar way to IW mode using 5 sub-swaths, obtaining a spatial resolution of 20×40 m.

4.1.3 Data

On ground, data are prepared by a network of processing and archive centres and published online. Each data acquisition mode can generate products at different levels:

- Level 0, are the raw data, uncompressed and uncalibrated, the basic data from which all the other products come, in IW mode are about 1 GB.
- Level 1 Single Look Complex, are calibrated and geo-referenced data using orbit and attitude data from the satellite, they are 8 GB data (IW mode double polarization), or 4 GB (if in a single polarization).
- Level 1 Ground Range Detected, data calibrated, geo-referenced and projected to the ground using an ellipsoid model. They are available at multiple resolution levels.
- Level 2, are products of geophysical interest obtained from SAR data and are used for applications in the ocean, wind, waves and currents.

4.2 Sentinel 1 and Sentinel 2 data fusion: Motivations and Challenges

Global World Monitoring is nowadays supported by a huge number of satellites [146] providing precious data in several applications, such as agriculture, biodiversity, and hydrology [147]. Many works used information from a single satellite with daily basis global information and coarse spatial resolution is utilized (MODIS: 250 m; Sentinel-5: 8 km, NOAA -

AVHRR: 1 km). Other researchers exploit data from satellites with finer spatial resolution and coarse temporal resolution (i.e., Landsat TM: 16 days; SPOT/HRV: 26 days) [148–151]. In this context, the bottleneck of using single sensor is an inherent trade-off between spatial and temporal resolution. Thus, in order to obtain a pixel-wise land cover monitoring with a dense revisit time, the fusion of the features extracted by both the multi-spectral and SAR data is approached. In this scenario, the ESA Copernicus program developed Sentinel missions, specifically for RS applications, aiming to ensure a continuity of data, never seen before, with the huge number of launched satellites. In particular, the launch of Sentinel-1/Sentinel-2 satellites opened unprecedented opportunities for both industrial and institutional end-users, and poses new challenges to remote sensing research community. Besides this, the Sentinel constellation's technical features make it a precious tool for a wide range of remote sensing applications. With revisit time ranging from two days to about a week, depending on the geographic location, spatial resolution from 5 to 60 meters, and wide coverage of the spectrum, from visible to shortwave infrared ($\sim 440 - 2200$ nm), Sentinel data may impact decisively on several Earth monitoring applications, such as climate change monitoring, map updating, agriculture and forestry planning, flood monitoring, ice monitoring, and so forth.

Especially precious is the diversity of information guaranteed by SAR and MS sensors, a key element for boosting the constellation's monitoring capability. The information conveyed by the Sentinel-2 (S2) multiresolution optical sensor depends on the spectral reflectivity of the target illuminated by sunlight, as described in the previous Chapter. The backscattered signal acquired by the Sentinel-1 (S1) SAR sensor depends on both target's characteristics and the illuminating signal. The joint processing of optical and radar temporal sequences offers the opportunity to extract the information of interest with augmented accuracy. This scenario poses

intriguing scientific challenges.

As first, we focus on estimating the Normalized Difference Vegetation Index (NDVI) in critical weather conditions, fusing the information provided by temporal sequences of S1 and S2 images. Remote sensing images can be continuously utilized to monitor land cover/land use changes around the world with extremely accurate precision. Optical images (i.e., Sentinel-2) allow to obtain rich information about vegetation, water, and so on. In fact, the most common processing pipelines of many land monitoring applications is based on a single date or a time-series of MS data, and thei related indexes, such as the NDVI [152–154], and the Normalized Difference Water Index (NDWI) [155, 156]. However, these features are unusable in cloudy weather conditions. The commonly adopted solution consists of interpolating between temporally adjacent images where the target feature is present.

However, given the availability of weather-insensitive SAR data, it makes sense to pursue fusion-based solutions, exploiting SAR images that may be temporally very close to the target date. Indeed, it is well known that radar images can provide precious information on vegetation monitoring [113, 118, 157–162]. However, the integration of S2 and SAR data is still an open problem. To address this problem, benefiting from deep learning methods' powerful learning capability, we designed and trained a three-layer convolutional neural network (CNN), to account for both temporal and cross-sensor dependencies. Note that the same approach, with minimal adaptations, could be extended to estimate many other spectral indices commonly used for water, soil, and so on. However, in the following, we will directly extract the class of water, soil or other, without going through the spectral indices. Therefore, besides solving the specific problem, we also demonstrate the potential of deep learning for this kind of data fusion in remote sensing.

Considering multi-sensor SAR-optical fusion for vegetation monitor-

ing, several contributions can be found in the literature [86, 91, 96, 113, 163]. In [86] ALOS POLSAR and Landsat time-series were combined at feature level for forest mapping and monitoring. The same problem was addressed in [96] through a decision-level approach. In [163], the fusion of single-date S1 and simulated S2 was presented for classification. In [113], instead, RADARSAT-2 and Landsat-7/8 images were fused, through an artificial neural network, to estimate soil moisture and leaf area index. The NDVI obtained from the Landsat source was combined with different SAR polarization subsets for feeding *ad hoc* artificial networks. A similar feature-level approach, based on Sentinel data, was followed in [91] for land cover mapping. To this end, the texture maps extracted from the SAR image were combined with several indices drawn from the optical bands. Although some fusion techniques have been proposed for Spatio-temporal NDVI super-resolution [117], or prediction [88], they use exclusively optical data. None of these papers attempts to directly estimate features that can be related to multispectral indices, like NDVI, from SAR data. In most cases, the fusion, occurring already at the feature level, is intended to provide high-level information, like the classification or detection of some physical item. Conversely, we can register some notable example of indices directly related to physical items of interest, like soil moisture or the area leaf index, estimated by fusing SAR and optical data [112, 113]. This work proposes several CNN-based algorithms to estimate a vegetation index through the fusion of optical and SAR Sentinel data. We acquired temporal sequences of S1 SAR data and S2 optical data about a specific case study, covering the same time-lapse, with the latter partially covered by clouds. Both temporal and cross-sensor (S1-S2) dependencies are used to obtain the most effective estimation protocol.

4.3 Estimating Vegetation Index using Sentinel 1 and Sentinel 2 data

The first explored approach is the reconstruction of an NDVI-like, in an available target image partially or totally cloudy. However, one may also consider the case in which the feature is built and used on a date for which no image is available.

Regarding Sentinel images, the NDVI is obtained at a 10 m spatial resolution by combining pixel-by-pixel two bands, near-infrared (NIR, 8th band) and red (Red, 4th band), as:

$$\text{NDVI} \triangleq \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \in [-1, 1] \tag{4.1}$$

All proposed solutions are based on a simple three-layer architecture (already described in the previous section), and differ only in the input layer, as different combinations of input bands are considered.

Once the architecture has been chosen, its parameters are learned by means of appropriate optimization strategy. In particular, the stochastic gradient descent (SGD) algorithm, specifying the cost function to be minimized over a properly selected training dataset, is chosen. Accuracy and speed of training depend on both the initialization of CNN's weights (Glorot inizialitation) and the setting of hyperparameters like learning rate α and momentum μ, with α being to most critical, impacting heavily on stability and convergence time. In particular, in this case, we have experimentally found optimal values for these parameters which are $\alpha = 0.5 \cdot 10^{-2}$ and $\mu = 0.9$.

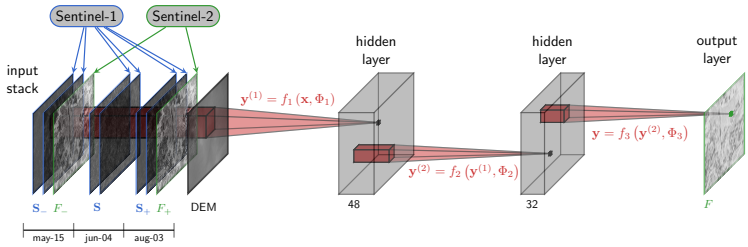


Figure 4.2. Proposed CNN architecture. The depicted input corresponds to the Optical-SAR+ case. Other cases use a reduced set of inputs.

4.4 Proposed prediction architectures

In the following developments, with reference to a given target S2 image acquired at time t, we consider the items defined below:

- F: unknown feature (NDVI in this work) at time t;
- F_- and F_+: feature F at previous and next useful times, respectively;
- $\mathbf{S} \triangleq (S^{VV}, S^{VH})$: double polarized SAR image closest to F (within ± 5 days for our dataset);
- $\mathbf{S}_-$ and $\mathbf{S}_+$: SAR images closest to F_- and F_+, respectively;
- D: DEM.

Apart from the composition of input stack and also the input layer, the CNN architecture is depicted in Fig. 4.2, with hyper-parameters summarized in Tab. 4.3. This relatively shallow CNN is characterized by a rather small number of weights (as CNNs go), counted in Tab. 4.3, and hence can be trained with a small amount of data. Moreover, we have already proven that similar architectures achieve state-of-the-art performance in closely related super-resolution application.

The number b_x of input bands depends on the specific solution and will be made explicit below. In order to provide output values falling in

Table 4.3. CNN hyper-parameters: # of features, M; kernel shape for each feature $N\times(K\times K)$; # of parameters to learn for each layer given by MNK^2 (for $\mathbf{w}$) + M (for $\mathbf{b}$). In addition, in the last row it is shown an example of feature layer shape for a sample input $\mathbf{x}$ of size $b_x\times(33\times33)$.

	ConvLayer 1	$g_1(\cdot)$	ConvLayer 2	$g_2(\cdot)$	ConvLayer 3
M	48		32		1
$N\times(K\times K)$	$b_x\times(9\times9)$	ReLU	$48\times(5\times5)$	ReLU	$32\times(5\times5)$
Learning rate	$0.5\cdot10^{-2}$		$0.5\cdot10^{-2}$		$0.5\cdot10^{-2}$
Momentum	0.9		0.9		0.9
# parameters	$\sim 3888\cdot b_x$		$\sim$38400		$\sim$800
Shape of $\mathbf{y}^{(i)}$	$48\times(25\times25)$		$32\times(21\times21)$		$1\times(17\times17)$

the compact interval [-1,1], as required by the NDVI semantics (Eq. 4.1), one can include a suitable nonlinear activation, like $\tanh(\cdot)$, to complete the output layer. In such a case, it is customary to use a cross-entropy loss for training. As an alternative, one may remove the nonlinear output mapping altogether and take the result of the convolution, which can be optimized using, for example, a L_n-norm. Obviously, in this case, a hard clipping of the output is still needed. This additional transformation does not participate in the error backpropagation. Hence it should be considered external to the network. Through preliminary experiments, we have found this latter solution more effective than the former, for our task, and therefore we train the CNN considering a linear activation in the last layer, $g_3(\mathbf{z}^{(3)}) = \mathbf{z}^{(3)}$.

We now describe briefly the different solutions considered here, which depend on the available input data, as already mentioned before, and the required response time.

We considered estimation based on optical-only, SAR-only, and optical+SAR data. When using SAR images, we will also test the DEM's inclusion, which may convey relevant information on them. Instead, the DEM is useless and hence neglected when only optical data are used. All

Table 4.4. Proposed models. The naming reflects the input stacking, explicited on the right. "SAR" refers to S1 images and "Optical" to S2 products ($F_{\pm}$). "+" marks the inclusion of the DEM. Moreover "C" stands for causal.

		Input Bands		
Model name	b_x	Optical	SAR	DEM
SAR	2		$\mathbf{S}$	
SAR+	3		$\mathbf{S}$	D
Optical/C	1	F_-		
Optical-SAR/C	5	F_-	$\mathbf{S}_-, \mathbf{S}$	
Optical-SAR+/C	6	F_-	$\mathbf{S}_-, \mathbf{S}$	D
Optical	2	F_-, F_+		
Optical-SAR	8	F_-, F_+	$\mathbf{S}_-, \mathbf{S}, \mathbf{S}_+$	
Optical-SAR+	9	F_-, F_+	$\mathbf{S}_-, \mathbf{S}, \mathbf{S}_+$	D

these cases are of interest for the following reasons.

- The *optical-only* case allows for a direct comparison, with the same input data, between the proposed CNN-based solution and the current baseline, which relies on temporal linear interpolation. Therefore, it will provide us with a measure of the net performance gain guaranteed by deep learning over conventional processing.

- Although SAR and optical data provide complementary information, the occurrence of a given physical item, like water or vegetation, can be detected through both scattering properties and spectral signatures. The analysis of the *SAR-only* case will allow us to understand if significant dependencies exist between the NDVI and SAR images. Moreover, if a reasonable quality can be achieved even when only this source is used for estimation. To this aim, we do not count on temporal dependencies. In this case, trying to estimate an S2 feature from the closest S1 image only.

- The *optical-SAR* fusion is the case of highest interest in this work.

Given the complete set of relevant input and an adequate training set, the proposed CNN will synthesize expressive features and is expected to provide a high-quality NDVI estimate.

Except for the SAR-only case, we distinguished between "nearly" causal estimation, in which only data already available at time t, for example, D, F_-, $\mathbf{S}_-$, or shortly later[1] (it can be the case of $\mathbf{S}$), can be used, and non-causal estimation, when the whole time series is supposed to be available and so future images (F_+ and/or $\mathbf{S}_+$) are involved.

- *Causal estimation* is of interest when a specific application requires a response in a short time, for example, early warning systems for food security. Hereinafter, we will refer to this "nearly" causal case as Causal for short;
- On the other hand, in the absence of temporal constraints, all relevant data should be taken into account to obtain the best possible quality, therefore using *non-causal estimation*.

Tab. 4.4 summarizes all these different solutions.

4.5 Semantic Segmentation task

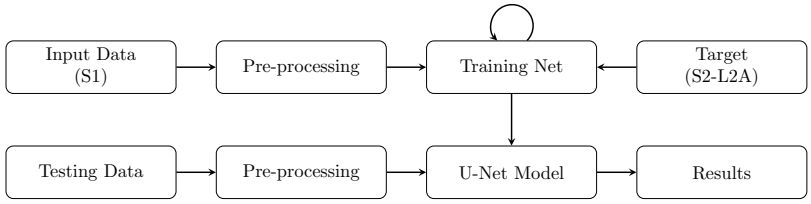


Figure 4.3. General workflow of the implemented method.

[1] In our experiments this happens only in two dates out of five, May 15th (3 days delay) and September 2nd (1-day delay).

Once the NDVI estimation is assessed, a direct extraction of useful land cover information without using spectral indices, but directly a specific scene classification map provided by Sentinel-2 data (L2A product) is at hand. Thus, from a generative problem we must face to a discrimative problem.

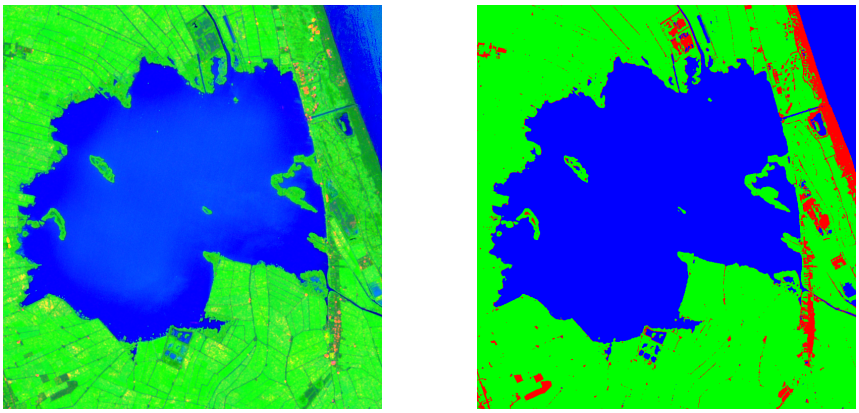


Figure 4.4. The false RGB colour of the lake under investigation (R: VH polarisation, G: NDVI, B: MNDWI), on the left, and the segmentation maps provided by L2A product (R: Bare soil, G: Vegetation, B: Water), on the right. Both in a specific date: August 24th, 2019.

4.5.1 Area of Interest

The study area shown in Figure 4.4 is a wetland area located in the Natural Park of Albufera, in the vicinity of Valencia, Spain. In Figure

Table 4.5. Different input stack considered in training phase

Configurations	No. Input bands	Description
1	1	VH
2	1	VV
3	2	VV, VH

4.4 we used the World Geodetic System (WGS84) ellipsoid as geographic reference system; latitude and longitude coordinates are reported in the image caption. The Albufera region is a wetland with a rich biodiversity, where rice farmers, fishermen, ecologists, fishes and migratory birds share the access to water. In this context, water volume and quality are affected not only by seasonality and climatic conditions, but also by human activities. In particular rice farmers pump or transfer water in accordance with their cultivation needs. This complex space- and time-variant environment constitutes an excellent benchmark for land cover mapping research. In addition, this area is particularly interesting from the applicative point of view, because many stakeholders are trying to use remote sensing tools for water management issues with an increasing need of technical algorithms to improve the usability of the data. In this area due to the rapid vegetation dynamics the fusion of SAR and optical data became very useful. In fact, the importance of combining these sources is highly helpful when cirrus and cloud presence reduces the usability of data from passive sensor. In our method, we used S1 and S2 data, whose main characteristics are reported in the Tab.4.6. The information from these sensors comes from different interaction between electromagnetic fields and the considered scene. In particular, in Figure 4.5 a S2 and S1 (with its typical speckle noise) are compared. The speckle effect is the consequence of the small objects presence (at the wavelength scale) and of the coherent nature of illumination in the resolution cell. Due to the microscopic details of these objects' shapes and positions the signals may interfere constructively or destructively. Thus, the overall effect is a "salt-and-pepper" noise on the S1 images.

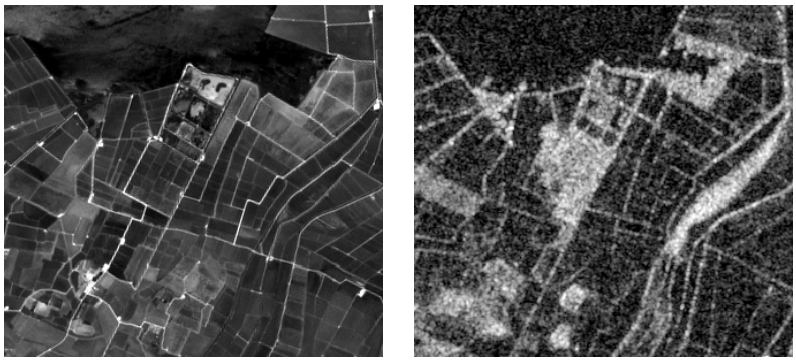


Figure 4.5. Visual comparison between Sentinel 2, B2 band on the left, and Sentinel 1, VV polarization on the right.

4.6 Deep Learning for Semantic Segmentation using Sentinel-1 data

In this context, our proposed approach is even based on a supervised learning using state-of-the-art architectures. We fed the proposed architectures with combinations of VV and VH polarizations as input stack, as reported in Tab. 4.5, and three out of eleven classes of Sentinel-2 classification scene (S2-L2A) as output. This method take advantage of the scene classification map included in L2A product (available in 12-18 hours through the web repositories), also used in mask clouds, water bodies, and snow. In particular, we only select 3 classes of the S2-L2A, specifically water, vegetation and bare soil (or not vegetated in SC map), useful for our study area. Further we want to show that the scene classification information of the S2-L2A product can be extracted by the S1 SAR data. However this optical-based classification (provided by L2A product) is affected by errors and a lot of invalid pixels. These problems determine a strong limitations in the training phase. Of course in order to further improve the performance of the deep learning approaches the target must

be more accurate. We consider this task as a further and future work. However, in this works we only want to explore the capability of S1 to determine this kind of information. Anyway we limit the invalid pixels to obtain better results. In literature the image segmentation is addressed using different approaches: (i) superpixel segmentation methods [164, 165], (ii) watershed segmentation methods [166, 167], (iii) level set segmentation methods [168, 169], and (iv) deep learning segmentation methods [170]. In particular, in the last decade an increasing interest in Deep Learning (DL) has incentivized the use of the methods based on the Convolutional Neural Networks in segmentation of remote sensing (RS) images. In a rising number of works, the segmentation is approached by using the Deep Learning (DL) architectures because of their effectiveness and appeal in computer vision and accordingly in RS context [171, 172]. Specifically, we investigate the use of the Sentinel-1 data (Tab. 4.1) in the Natural Park of Valencia, Albufera, in Spain, to monitor the presence of water, rice or bare soil. The rice growing dynamics have regularly to be monitored and can take advantages of: (i) the high accuracy of land cover mapping, (ii) an improved spatial resolution of the remote sensing data, and (iii) even better revisit time from the RS products. The area under investigation and the data on stage are described in the next chapters with more details.

Table 4.6. Summary of the used satellite sensors. The * symbol indicates that we only considered 6 days revisit time around the considered S2 date.

Data	Type	Satellite	Spatial Resolution	# Images	Minimum Revisit Time	Considered Revisit Time	Polarization Bands
Satellite Images	Synthetic Aperture Radar (SAR)	S-1	10 m	36	6 days	6 days*	VV + VH
	Multi-Spectral	S-2	10 m	12	5 days	1 month	ρ_8, ρ_{11}, ρ_4

4.7 Proposed Deep Learning Approach

Firstly, we proposed the U-Net architecture (Fig. 4.6) that is commonly used in semantic segmentation applications [173]. We considered

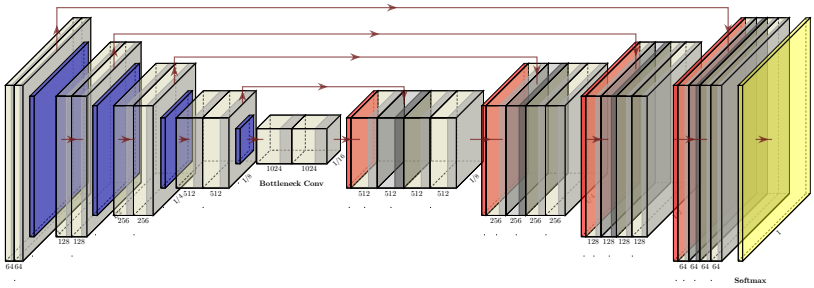


Figure 4.6. General illustration of the architecture of the used CNN U-Net.

the MobileNetV2 as backbone that gives a good compromise between computational time and performances. The U-Net took the idea from the architecture of the fully Convolutional Neural Network (CNN), followed by an additional part with upsampling that increase the features size. It is organized in a contracting and an expansive path, as reported in [174] and represented in Fig. 4.6. The first contracting part is an encoder, composed of blocks that include batch normalization layers, 3×3 convolutional layers interleaved by rectified linear unit (ReLU) and max pooling 2×2 layers. As described in the Chapter2, the max-pooling is useful to preserve the main target information and to reduce the number of parameters. Each max-pooling layer corresponds to a doubling in the number of feature maps. The second expansive path is the decoder part. Convolutional and deconvolution layers compose it. The downsampling of the first part and the upsampling of the second one is performed 4 times, and so the last feature maps obtained by the upsampling path have the same size as the input images. The U-Net performs a concatenation of the feature maps from the encoder path to the corresponding feature maps from the decoder part, and this concatenation is very helpful in recovery information lost during the convolutional and max pooling operations. The last layer is a 1×1 convolutional layer that matches the number of feature maps to the desired number of classes and uses a softmax activation function.

The configuration of this U-Net based semantic segmentation approach also requires supervised learning, so we have to realize input-target (x,y) samples to begin the training phase. We consider the S1 SAR data with different combinations of VV and VH polarizations as input and the scene classification maps from the S2-L2A product as the target.

4.7.1 Training

Couples of S1 data and S2 segmentation maps are used to train the model. To train the architecture properly, the S2 slave data are reprojected into the native reference system of the S1 master images. To be more effective in error back-propagation, we consider an objective function based on the Intersection over Union (IoU) [175]. In a plethora of segmentation works, the IoU is used as metrics in the classification task, and this clarifies why we used it in our specific context. The IoU function can be defined as:

$$IoU = \frac{I}{U} = \frac{y \cap \hat{y}}{y \cup \hat{y}} \tag{4.2}$$

where I and U are Intersection and Union, respectively, and y is the reference map and $\hat{y}$ is the predicted one. The IoU can be also expressed in terms of True Positive (TP), False Positive(FP), and False Negative (FN):

$$\frac{I}{U} = \frac{TP}{(TP + FP + FN)}. \tag{4.3}$$

Specifically, the IoU loss is computed on the objective function by averaging over the mini-batches samples at each updating step of the learning process:

$$\mathscr{L}_{IoU} = 1 - IoU = 1 - \frac{1}{N} \sum_{n=1}^{N} \frac{y_n \cap \hat{y_n}}{y_n \cup \hat{y_n}} \tag{4.4}$$

where N is the batch size in the training phase, y_n is the n-th target and $\hat{y_n}$ is the n-th predicted map, dependent on all the parameters of the network.

Then, we consider this IoU loss into the objective function, defined as in the following form:

$$argmin_w = 1 - IoU \tag{4.5}$$

where w is the weights of the convolutional layers. Further, we adopt the Adam optimizer implementation of Tensorflow, and we consider a 0.001 learning rate lower to the default value, while the other parameters are configured as in [141], that means for $\beta_1 = 0.9$, and $\beta_2 = 0.999$. In networks with a large number of convolutional layers, a good initialization of the weights is crucial. A good initialisation of the weights is crucial in networks with a large number of convolutional layers. In fact, we start the training by weights tuned with the Imagenet dataset [176]. Because of the scarcity of data for the training phase, we have tested that the network weights' initialisation is better than a random choice. This specific training phase is significant to obtain remarkable performance, as reported in the 5.5.2.

4.8 "Ad hoc" architecture

In this section, we introduce a multi-temporal input stack using the same datasets, i.e. S1 and S2. Further, as mentioned in the introduction, a threshold-based segmentation obtained by the multispectral indices replaced the scene classification map as output. The general workflow is shown in Fig.4.7 This is different in terms of samples from which the network learn a complex relationship between the S1 data and the specific classes under investigation. In this section, we describe the S1 and S2 data used in this subsequent approach.

The Sentinel-2 images allow obtaining rich information about vegetation, water, and so on. The most common land monitoring applications are based on temporal series of indices, for instance the Normalized Difference Vegetation Index (NDVI) [152], the Normalized Difference Water Index (NDWI) [155], and much more. The NDVI is defined as before in

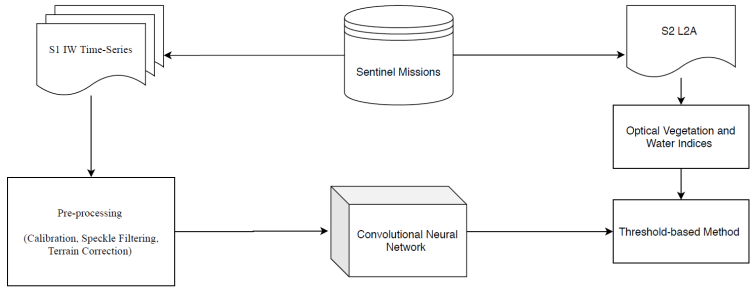


Figure 4.7. The general workflow.

the Chapter, and the NDWI is defined as in [177]:

$$\text{NDWI} \triangleq \frac{\rho_8 - \rho_{11}}{\rho_8 + \rho_{11}} \in [-1, 1] \tag{4.6}$$

where ρ_8 is the Near Infrared (NIR) band, centered at the wavelength of 842 nm; while ρ_{11} is the SWIR band, centered at 1610 nm.

Thus, we use a simple threshold on these indices to provide a segmentation map. In particular, the threshold is set to zero for both indices. After establishing the thresholds, we defined 3 classes using the mentioned thresholds-based rule: (i) *water*, when the values of NDWI are higher than zero, (ii) *vegetation* for values of NDVI higher than zero, and (iii) *bare soil* for values lower than zero both in NDVI and NDWI. This 3 classes segmentation map is the target of neural network in exam.

Sentinel-1 Analysis

We focus our data analysis on the entire 2019 year. The considered dataset obtained by twin Sentinel-1 satellites (details in Tab. 4.1) includes 12 acquisitions, one for each month. In the same time interval 12 S2 acquisitions have been acquired. In Figure 4.8 we summarize the temporal profiles per VV and VH polarizations of the considered classes: bare soil,

water and rice. As expected the water and vegetation classes are more easily distinguishable, instead the bare soil class is more challenging. In Figure 4.8 it is equally noticed that the bare soil class is distinctive in winter seasonal condition, and in summer assumes values that can be easily misclassified with other two considered classes. The information provided by VH is similar in terms of multi-temporal trend and different in terms of the range of intensity values. The combination of these two polarizations are surely effective in the segmentation context. Observing the multi-temporal trends, from the combination of these two polarisations, we can see that some specific configurations allow us to distinguish one class from the others. For instance, the water class is separable from the vegetation using dual polarization information, since the σ_0 in VV and VH for vegetation is lower than for water. Thus, in Figure 4.8, we even found the vegetation trend in the top of the water, because we showed $-10 \cdot log_{10}(\cdot)$. But, the bare soil class is more misleading because its average range of values is wider both for the VV and VH polarisations.

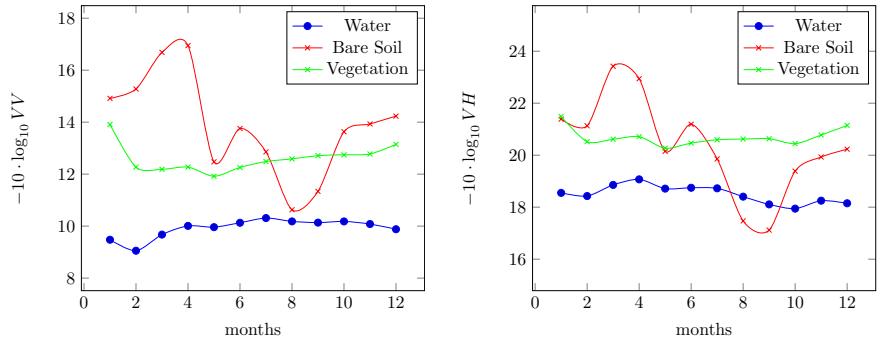


Figure 4.8. Multi-temporal information about the considered VV and VH polarisations, month by month, where with VV and VH we denote the σ_0 of these two polarisations.

As demonstrated in results, 12-day S1 information (three S1 dates closest to the target date) in the input stack further increases segmentation

performance.

4.8.1 A novel W-shape Architecture (W-Net)

Our proposed method is based on a supervised deep learning algorithm, whose general workflow is drawn in Figure 4.7. The core of the considered workflow is the Convolutional Neural Network that is fed with four combinations of VV and VH polarisations as input stack, as reported in Tab. 4.7. As output, we consider the above-described S2 segmentation map, obtained by threshold-based technique.

Table 4.7. Different input stack considered in training phase.

Configurations	No. Input Bands	Description	Considered Times
I	1	VH_i	1
II	1	VV_i	1
III	2	VV_i, VH_i	1
IV	6	VV_i, VH_i	0, 1, 2

In the multi-temporal input configuration, we consider three dates: one is the closest to the target date, and the others are the next closest dates, before and after the target date. The main innovation of this work is in using a deep learning architecture called W-Net. The CNN-based solutions are commonly used in many remote sensing applications [178, 179]. But, this specific architecture is used in a remote sensing context for the first time in this work. The structure is drawn in Figure 4.9, and we can see here that the architecture basically consists of two concatenated U-Nets. The W-Net, designed as an evolution of the U-Net architecture [173], is inspired by [180, 181]. The W-shaped design is identical to [180], but the two networks differ in the interconnection between the two twin U-Nets and as well as in the application point of view. In fact, in our case, this design is performed in supervised training to have a large number of convolutional levels, keeping a low number of trainable parameters given the

presence of max pooling, while in [180] this design is considered to perform unsupervised training. We limited the number of the features because this limitation gives a better compromise between processing time and accuracy of results. In fact, the number of kernels for any convolutional layers increases from 8 to 128 in the contracting path and goes back up to 8 in the expansive path. The second expansive path is the decoder part. Convolutional and upsampling layers compose it. The downsampling of the first part and the upsampling of the second one is performed four times, and so the last feature maps obtained by the upsampling path have the same size as the input images. After the last concatenation achieved by the

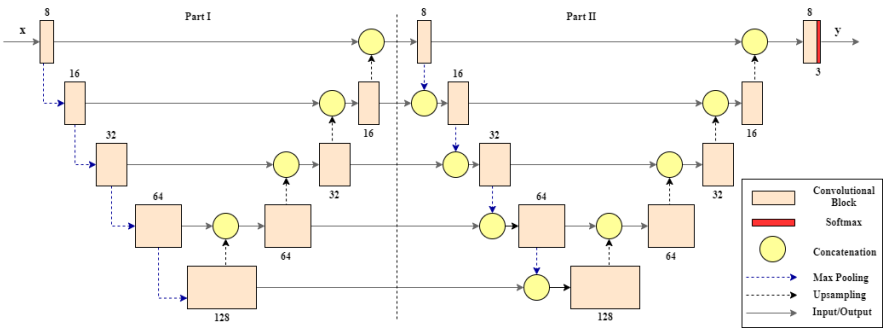


Figure 4.9. The proposed W-Net architecture.

first U-Net-like part, another sub-architecture has a U-Net structure. This second part is equal to the previous, but the max-pooling layers' output is concatenated with the blocks' output at the same level in the previous U-Net part. All these concatenations fed the basic encoder blocks of this second part of the W-Net. In the second U-Net, after the last concatenation between the first block of the encoder path and the upsampling of the last block of the decoder path, there is an additional block equal to all others. As in the single U-Net case, the last layer is a 1×1 convolutional layer that matches the number of feature maps to the desired number of classes and uses a softmax activation function. This W-Net based segmen-

tation approach also requires supervised learning, so we have to realize input-target (x, y) samples to begin the training phase. We consider the S1 SAR data with different multi-temporal combinations of VV and VH polarisations as input (see Tab. 4.7) and the indices-based segmentation maps from the S2-L2A product as the target. The (x,y) samples are used to train the model. In more details, we again adopt the ADAM algorithm as optimiser, and we set the ADAM default values, for the learning rate $\lambda = 0.02$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, as reported in [142]. In this case, we consider a larger dataset, and so we start the training by the Glorot initialization [182]. With larger training dataset, we have verified that this random initialization of the network weights gives considerable results.

Chapter 5

Results

One of the main challenges of DL methods is to acquire the capability of obtaining good performances on heterogeneous dataset. Such an aspect is not always adequately stressed and analysed when DL methods are presented. Firstly, in Section 5.1, we describe the results of the first proposed super-resolution approach (M5) in terms of water mapping and active fire monitoring. Then, in Section 5.3, we describe the general results of the last proposed approach (FUSE) from visual and numerical inspection using some classic metrics in pansharpening and super-resolution context. Thereafter, in Section 5.4, the experimental analysis of NDVI estimation using CNN is shown. In Section 5.5, the CNN-based semantic segmentation workflow is described.

5.1 Super Resolution (M5) and its Applications

Numerical results discussed below show that the quality of the first simple super-resolution method compares favourably against different pansharpening-like alternatives according to several indicators. In addition we have also tested the proposal M5 from the user's perpective by detecting water basins

through the MNDWI computed at 10-m resolution using the upsampled SWIR component and active fire through a set of spectral indices. In these first analyses, we want to underline the importance of using all 10-m bands in the training phase with respect to the "pure" super-resolution or the use of a single more correlated 10-m band. This super-resolution approach is particularly beneficial in water monitoring at a fine-scale through the MNDWI that, in some cases, provides more accurate results than the NDWI because of the spectral response of water in different conditions.

To train the three-layer CNN and build a sufficiently general dataset for training we have chosen three images portraiting rather different scenes: Guinea, Tunisia, and Italy (Venice). 450×450 clips have been selected for testing, and 17×17 patches in all the scenes in the remaining segments for training. Overall, 19k patches were collected and randomly grouped in 128-size mini batches for the implementation of the SGD-based training. Additional patches were also extracted for the purpose of validation completing the partition in 70% (training), 15% (validation), and 15% (test). The average numerical results obtained for the three scenes of interest are gathered in Tab. 5.1 (left part). The proposed pure super-resolution method M1 is compared to the standard bicubic interpolator in the top part of the Table. As average figures, Q-index and ERGAS do not evidence the gain provided by M1 as it does HCC which deals with high frequency components that are affected by the super-resolution and are mostly localized on boundaries. Moving to model M2, which takes the additional input band ρ_8, it compares favourably against classical pansharpening methods adapted to the Sentinel-2/SWIR problem as suggested in [74].In the last row it is given the performance of the proposed method when all four high-resolution bands are added to the input stack. As it can be seen the three additional, although less correlated with ρ_{11} (as shown in Tab. 3.2), provide an additional gain.

The proposed models are also tested from the application point of view

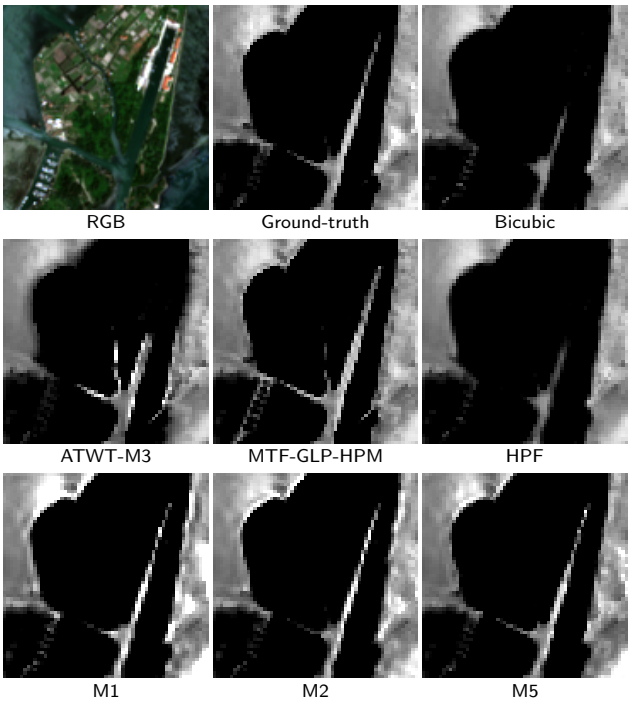


Figure 5.1. MNDWI estimations over a sample detail (from Venice image). In order to have a reference ground-truth we applied Wald's protocol (downgraded resolution).

by detecting water basins through the computation of the MNDWI index (I) defined as

$$I = \frac{\rho_3^{(\downarrow)} - \rho_{11}}{\rho_3^{(\downarrow)} - \rho_{11}} \quad \text{or} \quad \widehat{I} = \frac{\rho_3 - \widehat{\rho}_{11}}{\rho_3 - \widehat{\rho}_{11}}$$

at resolution of 20-m or 10-m, respectively. Once water (W) is detected by suitably thresholding the MNDWI ($W = I > \alpha$), the classification error rate on the whole image (CER) and locally to boundaries[1] (L-CER) is computed and reported on the right-hand side of Tab. 5.1. These figures provide a further confirm of the superiority of the proposed method.

[1]Boundaries are detected using morphological gradient.

Table 5.1. Accuracy of $\widehat{\rho}_{11}$ (Q-index, ERGAS, HCC) and water maps (CER, L-CER) at 20-m.

Methods (ideal value)	**Q-index** (1)	**ERGAS** (0)	**HCC** (1)	**CER** (0)	**L-CER** (0)
Bicubic	0.9914	4.992	0.5366	0.0166	0.1876
M1 (proposed)	**0.9970**	**3.036**	**0.7090**	**0.0086**	**0.0909**
ATWT-M3 [70]	0.9873	5.949	0.5828	0.0160	0.1762
MTF-GLP-HPM [183]	0.9823	7.245	0.4509	0.0207	0.1370
HPF [69]	0.9922	4.688	0.5832	0.0138	0.1680
M2 (proposed)	**0.9975**	**2.830**	**0.7718**	**0.0064**	**0.0637**
M5 (proposed)	**0.9983**	**2.354**	**0.8500**	0.0066	**0.0594**

To conclude this section we show in Fig. 5.1 some sample results (at 20-m with reference) which further confirm the effectiveness of the proposed method. An example of full-resolution (10-m) estimation of $\hat{\rho}_{11}$ is shown in Fig. 3.1.

5.1.1 Filter Size Comparison

In this analysis, we compare the proposed $M5$ method with a heavier $M5^{\uparrow}$ method. In the $M5^{\uparrow}$, we consider more weights than in the 3-layer CNN $M5$ architecture, as shown in Tab. 5.2, but this is not beneficial in the training phase, as we can see in Fig. 5.2. Further confirmation of the superiority of $M5$ method is shown in Tab. 5.3. The $M5$ superiority is because, with a larger filter size, we compute the loss function on a smaller number of pixels, and thus we learn from a more minor valid part of the considered patch in the batches. This aspect gives better capability in the learning phase to the $M5$ method to its heavier version $M5^{\uparrow}$.

5.2 Super Resolution in Active Fire Monitoring

Sentinel-2 data provide useful information in the burnt areas and active fire monitoring, using several algorithms [184]. Many algorithms are based

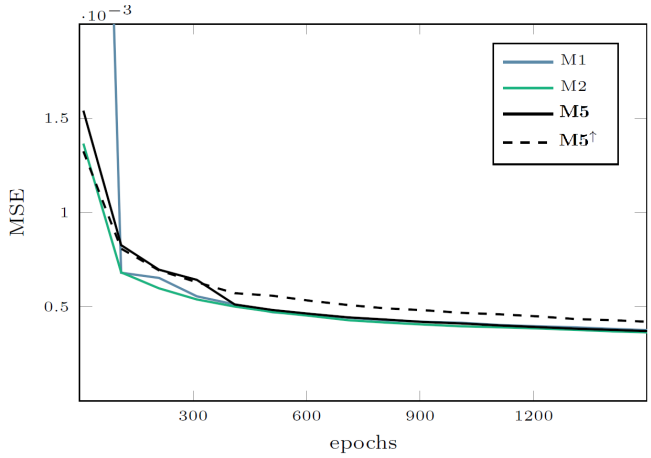


Figure 5.2. Comparison of Losses for Different Configurations.

Table 5.2. Different filter Size.

Model	**kernel size**			time [s]
	$l = 1$	$l = 2$	$l = 3$	
M5↑	9×9	5×5	5×5	7605
M5	3×3	3×3	3×3	4794

on the threshold of spectral indices involving Near-Infrared (NIR) and Short-Wave Infrared (SWIR) bands [59, 185, 186] that Sentinel-2 provides at the spatial resolution of 10 m and 20 m, respectively. Therefore, it is common to resort to the 20-m resolution indices by just downscaling the NIR band from 10 m to 20 m. However, following this approach, spatial information from the NIR band would be lost. An alternative approach to enhance the AFD method using the Sentinel-2 images is to produce the Active Fire Indices (AFIs) by upscaling the SWIR bands from 20 m to 10 m. The main issue is the method's choice to improve the spatial resolution of the SWIR bands. To produce the 10-m AFIs from Sentinel-2 bands with SRDF methods, the use of all the highest spatial resolution bands is not benefiting because of their smoke-sensitivity. The major contribution

Table 5.3. Comparisons in terms of filter size.

Methods (ideal value)	**Q-index** (1)	**ERGAS** (0)	**HCC** (1)	**CER** (0)	**L-CER** (0)
M5$^{\uparrow}$ (proposed)	0.9947	4.294	0.8432	0.0142	0.0946
M5 (proposed)	**0.9983**	**2.354**	**0.8500**	**0.0066**	**0.0594**

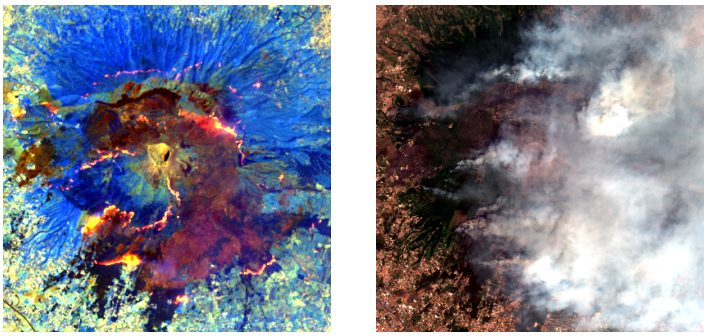
is derived from the NIR, which is the only band we consider in the SRDF approaches.

5.2.1 Study Area and Dataset

The area under investigation is located at the Vesuvius (in Fig. 5.3), a volcano close to Naples, Italy. We are motivated by choice of the study area since the presence of a natural park with a huge variety of flora and fauna considering its limited size [187]. At the beginning of July 2017, hundreds of wildfires ignited and damaged the Italy country, whose the most serious was at Vesuvius. Fires had been interesting the Vesuvius area for several days, and the situation quickly became more dangerous due to adverse climatic conditions (winds and dry weather) [188]. The considered dataset is the Sentinel-2 Level-1C product acquired *on 12th July 2017*. As we can see in Fig. 5.3, the area under investigation is mainly covered by heavy smoke (Fig. 5.3-(b)), which reduces the usability of 10-m spectral information.

The proposed M5 model is evaluated, from the application point of view, by monitoring active fires through the computation of three different spectral indices (in Fig. 5.4), mainly used to this aim in literature [189–191] because of their ease computing. The AFIs [189, 190] are defined on Sentinel-2 data as follows:

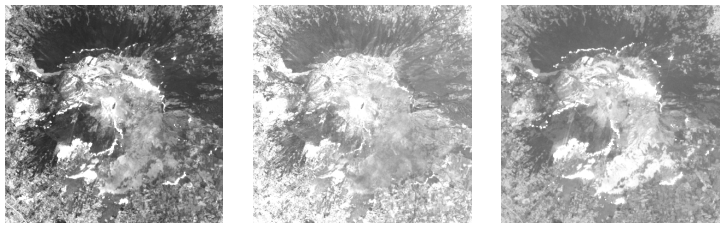
$$AFI_1 = \frac{\rho_{12}}{\rho_8} \qquad AFI_2 = \frac{\rho_{11}}{\rho_8} \qquad AFI_3 = \frac{\rho_{12}}{\rho_{11}}$$



(a) (b)

Figure 5.3. (a) false colour composite (ρ_{12}, ρ_{11} and ρ_8 bands) and (b) RGB image of Vesuvius .

where ρ_8 is the 10-m spatial resolution NIR band, centered at the wavelength of 0.834 μm; while ρ_{11} and ρ_{12} are the 20-m SWIR bands, centered at 1.610 μm and 2.190 μm, respectively. All of these bands represent the radiance data at top of the atmosphere. The choice of these indices is based on their physical properties. In fact, we exploit the sensitivity of the SWIR band to the emissive component of fires in this SWIR spectral band. However, due to the reflected solar component that dominates the background, there is a risk of losing the reflected active fire components, so the NIR band is exploited to separate the background from the active fire contribution. This is possible because the ρ_{NIR} band is related to the ρ_{SWIR} band in the absence of fires, but is not in presence of the fire [192]. As happens for thermal anomalies in the mid-thermal infrared bands, in the same way for the ρ_{SWIR} bands the presence of fires produces an anomaly in the SWIR reflectance with respect to solar reflection. Furthermore, the visible bands are unusable since they are sensitive to smoke, unlike ρ_{NIR} and ρ_{SWIR}, as shown in Fig. 5.4.Specifically, the conditions $AFI_1 > 1$ and $AFI_3 > 1$ often occur in active fire; while the condition $AFI_2 < 1$ is verified near the fire fronts [191].



AFI_1 AFI_2 AFI_3

Figure 5.4. Active Fire Indices related to Vesuvius.

In this application, we test the improvement of both SWIR bands' spatial resolution using the M5 method.

5.2.2 Accuracy Metrics

In order to evaluate the performances of the presented methods with a full resolution analysis, we used the active fire monitoring application as benchmark, and we compare all the methods in terms of binary classification. To this end, we need to define a ground truth and the main classification metrics. In this context, such ground truth is obtained with a differential multi-temporal approach, based on a thresholding of the difference between two cloud-free realizations of Normalized Difference Vegetation Index (NDVI) in two different dates (before and after the fire event). This ground truth (GT) is affected by noise (or small bright pixels), and so we have used a morphological operator (opening) to erase this undesired noisy effect.

Thus, we compare this GT with the active fire maps obtained by thresholding the above-mentioned spectral indices. *In the results, we compare different super-resolution approaches on the super-resolved AFIs and the corresponding AFDs.* In particular, we consider different thresholds on each of AFIs to match the active fires' best detection. To evaluate the quality of the obtained binary maps, we have considered some metrics

typically used in the classification task:

- Precision (P) is the ratio between the correctly predicted positive observations and the total predicted positive ones ;
- Recall (R) is the ratio between the correctly predicted positive observations and all actually positive observations;
- Intersection over Union (IoU) is the ratio between the overlapping area and the union area. The intersection and the union are computed on the predicted positive observations and the positives from the GT.

It is worthwhile to remember that a high precision corresponds to a low false-positive rate. In other words, the higher the percentage of correctly predicted positive over the total predicted positive, the higher the precision. Instead, a high recall corresponds to a low false-negative rate.

5.2.3 Training Phase

Given the limited number of available input-output samples in the present context, we start from a pre-trained M5 solution [45] to train the network's parameters Φ^n. In [45] a super-resolution technique is only considered for ρ_{11} band ($\mathbf{x} = \rho_{11}$). Here we propose an equivalent solution for the ρ_{12} band ($\mathbf{x} = \rho_{12}$). In particular, to create a pre-trained solution, we use the dataset considered in [45]. Then, we fine-tune the CNN's weights from the considered Vesuvius zone on two different dates, close to the target date (specifically June 27th and July 27th). This can be considered a geographical fine-tuning because we adapt CNN's weights on the geometric features of the study area. Then, we test this fine-tuned solution, hereafter $M5_+$, on the date under investigation (July 12th, 2017). Once left apart from the target date for testing, 17×17 patches for training are uniformly extracted from two dates, as mentioned earlier in the remaining segments. Overall, 10k patches are collected from the considered dates and randomly

partitioned in 80% for the training phase and 20% for the validation phase. The 8k training patches are grouped in 32 pixel size mini-batches for the implementation of the ADAM-based training. The fine-tuned solution is considered better than the solution from scratch when a large amount of data for the training phase is not available or when the computing power is not sufficient [193]. Eventually, we minimize the L1-norm cost function, defined in Chapter 3, on the training examples using the ADAM learning algorithm. Thus, we set the ADAM default values $\eta = 0.002$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, as reported in [194]. In this specific case, the training phase requires 200 epochs (32×200 weight updates) performed in a few minutes using GPU cards, while can do the test in real-time.

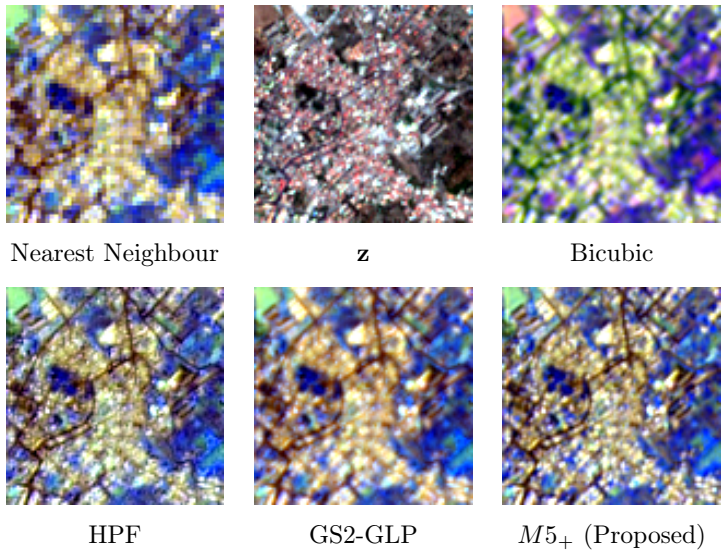


Figure 5.5. Detail of the study area obtained by several super-resolution techniques and our proposal to underline the improvement in terms of spectral distortion. In the middle of the first row: **z** is only composed by RGB bands.

5.2.4 Discussion: Active Fire Detection

In this section, $M5_+$ is compared to the pre-trained CNN-based method (M5), three popular SRDFs adapted to the Sentinel-2/SWIR problem, including GS2-GLP [195], HPF [69], and PRACS [195], and even the SISRs, that is the Nearest Neighbour (NNI) and bicubic interpolation techniques. The numerical results obtained for the area of interest are reported in the left part of the Tab. 5.4. In the results, we consider an average of the metrics on both SWIR bands. In the top part of the Table, the $M5_+$ is compared to the SISR techniques, and the improvement is very remarkable in the HCC metric, because high-frequency components are affected by the super-resolution and are mostly localized on boundaries. Observing the Tab. 5.4, the proposed $M5_+$ method compares favourably against classical fusion methods, which take information from the additional input band ρ_8. In the penultimate row, the pre-trained M5 model is performed, and even in this case, the $M5_+$ performs slightly better results in terms of all the metrics at the 20-m resolution. As it can be seen the additional fine-tuning, although having few training patches, provide a further gain. To conclude this section, we show in Figures 5.6-5.5 some sample results (at 10-m without reference) which further confirm the effectiveness of the proposed method.

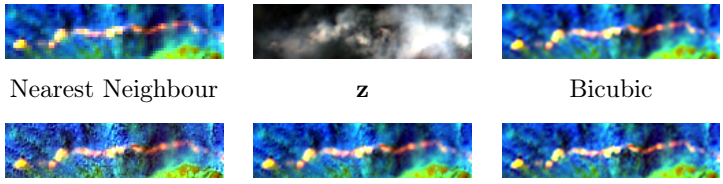


Figure 5.6. Detail of the area under investigation obtained by several super-resolution techniques and our proposal to underline the improvement in terms of spectral distortion. In the middle of the first row: **z** is only composed by RGB bands that are affected by smoke presence (in the CNN input **z** is also composed by ρ_8 band).

Once active fire (AF) is detected by considering the followed rules: $AFD_k = AFI_k > \alpha_k$, where $k \in \{1, 2, 3\}$, the performance is computed in terms of Precision, Recall and IoU and reported on the right-hand side of Tab. 5.4. The numerical results confirmed the effectiveness of the proposal, and the Fig.5.7-5.8 further confirm the superiority of the proposed method. As we can see in Tab. 5.4, the $M5_+$ method has the best performance in terms of the precision metric. In particular, its values are much greater than those of the classic techniques, demonstrating that it benefits from the joint information obtained from the visible bands. The low false alarm rate is well visible in Fig.5.8, where an urban area is shown. The Fig.5.8 only refers to AFI_2, but similar results are provided by the other analysed indices. The proposed method ($M5_+$) has worse performance both in terms of Recall and IoU metrics. We suppose this is mainly due to the ground truth used in validation, which probably over-estimates the areas interested by fires. In fact, as we can see in the central column of the Fig. 5.7, the $M5_+$ AFIs better define these areas, resulting in lighter and thinner than the ones obtained by other techniques. Furthermore, we can observe from visual inspection that the boundaries are more evident considering ρ_{12} and ρ_{11} than ρ_{12} and ρ_8. In general, even though this determines a low detection rate on the maps obtained by the AFI_3 concerning AFI_1.

5.2.5 SAM-based Active Fires Monitoring

We further stress the use of super-resolved bands, using an alternative method, based on the use of the Spectral Angle Mapper (SAM), for the detection of active fires. We take advantage of the super-resolved bands to detect the active fire better, using SAM-based approach with respect to AFI-based. To understand better the goodness of this SAM-based method, we further consider other classical classification metrics:

- F1 score (F-measure) is the harmonic average of precision and recall

Table 5.4. In the left part of the Table: average results in terms of main metrics (at 20-m), typically used in pansharpening and super-resolution context. In the right part of the Table: average results in terms of classification metrics.

Methods	SAM (0)	Q-index (1)	ERGAS (0)	HCC (1)	Precision (1)	Recall (1)	IoU (1)
NNI	0.001960	0.9182	9.353	0.1355	0.8329	0.5773	0.5309
Bicubic	0.001964	0.9515	7.155	0.471	0.8387	0.5900	0.5471
HPF	0.064590	0.9405	8.150	0.2826	0.7799	0.5991	0.5476
PRACS	0.001979	0.9535	7.057	0.5117	0.7993	0.5987	0.5497
GS2-GLP	0.050190	0.9540	7.043	0.4694	0.8008	**0.6131**	**0.5571**
M5	0.001963	0.9688	5.943	0.6246	0.8373	0.5649	0.5158
$M5_+$ (Proposed)	**0.001956**	**0.9743**	**5.425**	**0.6334**	**0.8414**	0.5642	0.5157

(defined earlier in section 5.2.2);

- Accuracy (p_A) is the ratio between the correctly predicted observations and all actually observations.

Also for these metrics the predicted observations are compared to the ground truth (GT, described before in section 5.2.2). Hereafter, we only considered the AFI as the ratio between the two SWIR bands provided by Sentinel-2:

$$AFI = \frac{\rho_{12}}{\rho_{11}}$$

An alternative option to detect the active fire is a method based on the SAM metric. This metric is usually applied in unsupervised classification based on the spectral angles between image and reference pixels [196]. The considered SAM is defined as follows [197]:

$$SAM = \arccos\left(\frac{\mathbf{x^T} \cdot \mathbf{y}}{||\mathbf{x}||||\mathbf{y}||}\right)$$

where $\mathbf{x}$ and $\mathbf{y}$ are respectively the vector related to the measured reflectance in the two considered SWIR bands for the image pixels and the

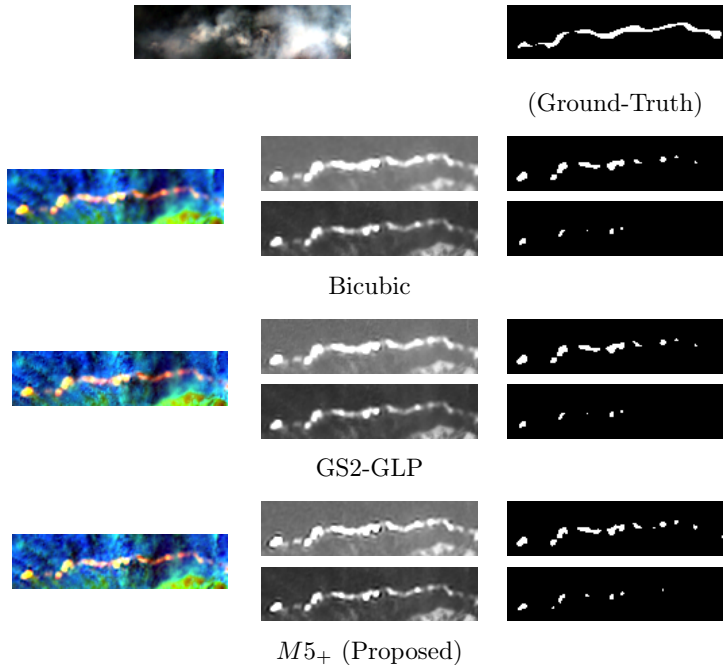


Figure 5.7. In the first row the RGB image in which we can observe the presence of the smoke and the ground truth. Then, from the second row to the bottom: in the first column false-RGB, in the second AFI_1 and AFI_3, in the third the respective Maps.

vector for the reference pixels. In our case, the reference is obtained by a linear Least Square (LS) regression of one band using the information from the other. The LS regression has been suggested by the observed linear relationship between the SWIR bands in no-fire condition (as shown in Fig. 5.9 for June 27th). Furthermore, the different behaviour in presence of active fires stimulated the usage of this metric in the AFD (Jul 12th in Fig. 5.9). Thus, starting from the linear relationship between the SWIR bands in no-fire condition, it is noticed a rotation in the cloud of points when the fire is active. The spectral angle variation is caused by different natural events, and in our case the main rotation with respect

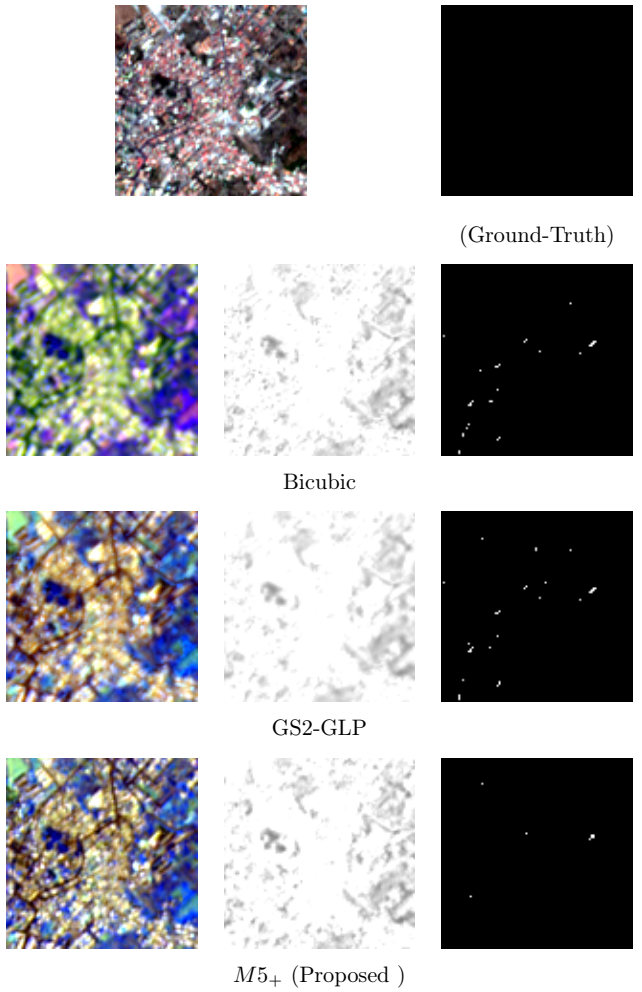


Figure 5.8. In the first row the RGB image in which we can observe the absense of the smoke and the ground truth. Then, from the second row to the bottom: in the first column false-RGB, in the second AFI_2, and in the third the respective Map.

to the natural behaviour is caused by the presence of active fire. In this case, we exclusively observed that a technique based on distance between the pixel values in (ρ_{11}, ρ_{12}) space is not sufficient, as in AFI-based tech-

niques. The observed scatter plot suggest us that higher angular distortion implicate higher probability of active fire presence, and thus we estabilish a threshold-based rule, similarly to AFI-based approaches [59].

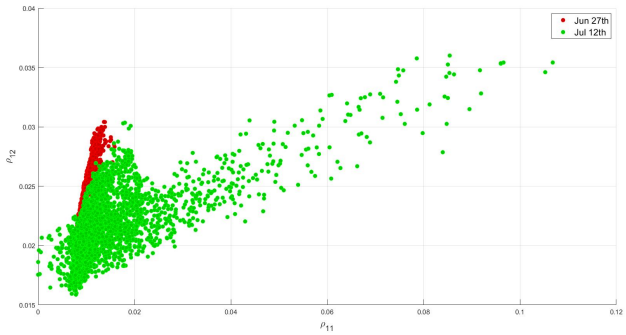


Figure 5.9. Scatter plot of the Vesuvius image in (ρ_{11}, ρ_{12}) space.

5.2.6 Comparison between SAM-based and AFI-based Detection

In Fig.5.10 it is shown the same previously analysed area. In this Figure, the AFI-based and the proposed SAM-based Detection are compared. According to poor image quality and imprecision of the detection technique the results are not reliable in terms of Recall and IoU (Tab. 5.5). This example allows us to underline the goodness of the SAM-based approach on thin anomalies detection, since the result is not reliable when we used the AFI-based approach. In this specific case, the anomalies are mainly represented by the presence of active fires. This technique is much better in terms of detection, and further confirmation came from the numerical results in Tab.5.5. The results are only considered on this small area, but similar consideration were confirmed in the whole area. Furthermore, we have only shown the numerical results of $M5_+$ since we have coherent considerations related to the others. However the SAM-based technique can

not separate different vegetative conditions from active fires, this provokes poor precision, that represents the capability of the method to detect the real positive values.

Table 5.5. The results are related to $M5_+$ on the small selected area.

	Precision	Recall	IoU	F-Measure	Accuracy
AFI-based	**0.9489**	0.2834	0.2791	0.4364	0.9100
SAM-based	0.6116	**0.7382**	**0.5026**	**0.6690**	**0.9102**

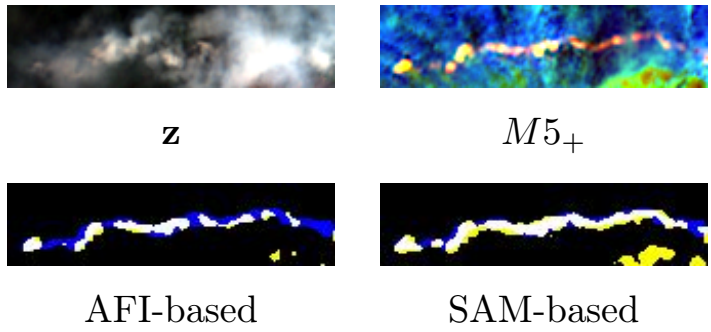


Figure 5.10. In the first row: Cloudily-sensitive RGB image, and False color RGB, in the second row the detection based on AFI and SAM. The colours of the maps have the same meaning as in Fig. 5.7.

5.3 Fast Upscaling Sentinel-2 (FUSE)

In this section, after a brief recall of the accuracy evaluation metrics (Section 5.3.1) and of the comparative methods (Section 5.3.2), we provide and discuss numerical and visual results (Section 5.3.3) of the final proposed FUSE method.

5.3.1 Accuracy Metrics

The quality assessment of pansharpening algorithms can be carried out in two frameworks, with or without ground-truth. Since the ground-truth

is often unavailable, the application of Wald's protocol [125] is applied, which is the same process used for the generation of training samples, as described in Section 3.3.1. Therefore, this evaluation frame, hereinafter referred to as *reference-based*, applies in the reduced-resolution domain and allows one to provide objective quality measurements. Because of the resolution shift (downgrade), the reference-based evaluation approach has a limited extent and it is therefore custom to complement it with a full-resolution assessment, referred to as the *no-reference* one, aimed to give qualitative measurements at full resolution. Experimental results demonstrate that the proposed solution can achieve better performance with respect to most of the state-of-the-art methods, including other deep learning based ones with a considerable saving of computational burden.

To evaluate the performance when the target image is available (in our case, at 20-m spatial resolution), the proposed method is compared to alternative methods using four reference metrics, commonly used for pansharpening [198]:

- Spectral Angular Mapper (SAM) the spectral distortion between pixel of reference image and estimated one [199];
- Universal Image Quality Index (UIQI, or Q-index), an image quality indicator introduced in [200];
- Relative Dimensionless Global Error (as known as ERGAS) which reduces to the root mean square error (RMSE) in case of single band [66];
- High-frequency Correlation Coefficient (HCC), the correlation coefficient between the high-pass components of two images [45].

As these indicators require reference, likewise for the training data, reduced resolution test data are produced through Wald's protocol. On the other hand, as no-reference metrics, we use the following [39, 201]:

- Spectral Distortion (D_λ) measures the spectral distance between the bicubic upscaling of the image component to be super-resolved, $\tilde{\mathbf{x}}$, and its super-resolution, $\hat{\mathbf{x}}$.
- Spatial Distortion (D_S) is a measurement of the spatial consistency between the super-resolved image $\hat{\mathbf{x}}$ and the high-resolution component $\mathbf{z}$.
- Quality No-Reference (QNR) index is a combination of the two above indexes that accounts for both spatial and spectral distortions.

For further details about the definition of the above metrics, the reader is referred to the associated articles.

5.3.2 The FUSE competitors

The FUSE method is compared with several state-of-the-art solutions.

On the one side there are classical approaches for pansharpening, properly generalized to the case of Sentinel-2, such as the following:

- Generalized Intensity Hue Saturation (GIHS) method [68];
- Brovey transform-based method [202];
- Indusion [203];
- Partial Replacement Adaptive Component Substitution (PRACS) [204];
- A Troús Wavelet Transform-based method (ATWT-M3) [70];
- The High-Pass Filtering (HPF) approach [69];
- Generalized Laplacian Pyramid with High Pass Modulation injection (MTF-GLP-HPM) [205];

- Gram-Schmidt algorithm with Generalized Laplacian Pyramid decomposition (GS2-GLP) [205].

Detailed information about these approaches can be found in the survey work of Vivone et al. [39].

On the other hand, some following deep learning approaches native for Sentinel-2 images, including two ablations of our proposal is considered in the comparison:

- the first CNN-based method (M5) proposed in [136], extended (training from scratch) to all six 20 m bands, a shallow architecture of the CNN that take as input the 20 m bands and the other bands provided at higher spatial resolution (SR), and the SR of the 20 m bands is only improved;
- The CNN model (DSen2) proposed in [44], which is much deeper than ours and has been trained on a very large dataset. this solution improved to 10 m both the 20 m and the 60 m bands, and needed several hours to learn from the selected samples [44];
- The advance of M5 where High-Pass filtering on the input and other minor changes have been introduced (HP-M5) [49], which represents a first insight on the improvements proposed in this work;
- FUSE with only three layers instead of four;
- FUSE trained using the L1 norm without regularization and structural loss terms.

5.3.3 Numerical and Visual Results

To assess the performance of the proposed method, we collected twelve 512×512 images (at 10 m resolution) from four larger images taken over

Athens, Adis Abeba, Sydney and Tokyo, respectively, from which no training or validation samples were extracted.

Numerical figures were computed for all compared methods on each test image. The average measures over the dataset are gathered in Tab. 5.6. Reference-based accuracy indicators shown on the left-hand side of the Table are computed in the reduced-resolution space and provide objective measurements of the reconstruction error. Overall, we can see that the proposed FUSE method performs slightly better than DSen2 and outperforms all compared solution on three out of four indicators. On the other hand, M5 and ATWT-M3 show a slightly better spectral preservation compared to FUSE according to the Spectral Angle Mapper indicator.

As reduced-resolution data do not fully reproduce statistical fluctuations that may occur in the full resolution context, a common choice is to complement the low-resolution evaluation with a full-resolution assessment that, however, does not rely on objective error measurements. In particular, we resort to three well-established indicators that are usually employed in the pansharpening context: the spectral and spatial distortions, D_λ and D_S, respectively, and their combination, the QNR. According to these indicators, shown on the right-hand side of Tab. 5.6, the proposed method, again, outperforms the competitors. A slightly better spectral preservation is given by HP-M5, M5 and ATWT-M3.

Let us now look at some sample results starting from the full-resolution context. Figures 5.11 and 5.12 show some of the 512×512 images used for test, associated with urban and extra-urban contexts, respectively. For the sake of visualization, we use RGB false-colour subsets of $\mathbf{z}$ and $\mathbf{x}$. In particular, we use three out of four bands of $\mathbf{z}$ (ρ_2, ρ_3 and ρ_4), and three out of six bands of $\mathbf{x}$ (ρ_5, ρ_{8A} and ρ_{11}—see Tab. 3.1). The input components $\mathbf{z}$ and $\widetilde{\mathbf{x}}$ are shown on the left and middle columns, and the super-resolution $\hat{\mathbf{x}}$ obtained with the proposed method is shown on the right.

Although these results look positive at first glance, a different observa-

Table 5.6. Accuracy assessment of several super-resolution methods. On top are model-based approaches and DL methods are on the bottom, including the proposed FUSE method.

	Reference-Based				No-Reference		
Method	**Q**	**HCC**	**ERGAS**	**SAM**	**QNR**	$\mathbf{D_\lambda}$	$\mathbf{D_S}$
(**Ideal**)	**(1)**	**(1)**	**(0)**	**(0)**	**(1)**	**(0)**	**(0)**
HPF	0.9674	0.6231	3.054	0.0641	0.8119	0.1348	0.0679
Brovey	0.9002	0.6738	4.581	0.0026	0.6717	0.2382	0.1241
MTF_GLP_HPM	0.8560	0.6077	19.82	0.2813	0.7802	0.1678	0.0643
GS2_GLP	0.9759	0.6821	2.613	0.0564	0.8129	0.1367	0.0647
ATWT-M3	0.9573	0.6965	3.009	**0.0019**	0.8627	0.0947	0.0473
PRACS	0.9767	0.7284	2.274	**0.0019**	0.8800	**0.0847**	0.0395
GIHS	0.8622	0.6601	5.336	0.0579	0.6112	0.2999	0.1444
Indusion	0.9582	0.6273	3.314	0.0425	0.8424	0.1311	0.0321
M5	0.9883	0.8432	1.830	**0.0019**	0.8715	0.0942	0.0389
HP-M5	0.9895	0.8492	1.720	0.0282	0.8779	0.0931	0.0329
DSen2	0.9916	0.8712	1.480	0.0194	0.8684	0.1028	0.0330
FUSE (3 layers)	0.9931	0.8602	1.631	0.0020	0.8521	0.1082	0.0474
FUSE (L1 loss)	0.9930	0.8660	1.681	0.1963	0.8570	0.1081	0.0410
FUSE (full version)	**0.9934**	**0.8830**	**1.354**	0.0184	**0.8818**	0.1002	**0.0203**

tion scale would help to gain insight into the behaviour of the compared solutions. Therefore, in Figure 5.13, we show some zoomed details with the corresponding super-resolution obtained by the selected methods. In particular, for the sake of simplicity, we have restricted the visual inspection to the most representative approaches according to both reference-based and no-reference indicators reported in Tab. 5.6. A careful inspection reveals that some model-based approaches provide higher detail enhancement compared to DL methods. However, it remains difficult to appreciate the different methods' spectral preservation capability due to the lack of objective references.

Actual errors can be visualized in the reduced-resolution domain, while Fig. 5.14 shows a few meaningful details processed in such a domain.

For each sample, the composite input $(\widetilde{\mathbf{x}}_\downarrow, \mathbf{z}_\downarrow)$ is shown in the leftmost

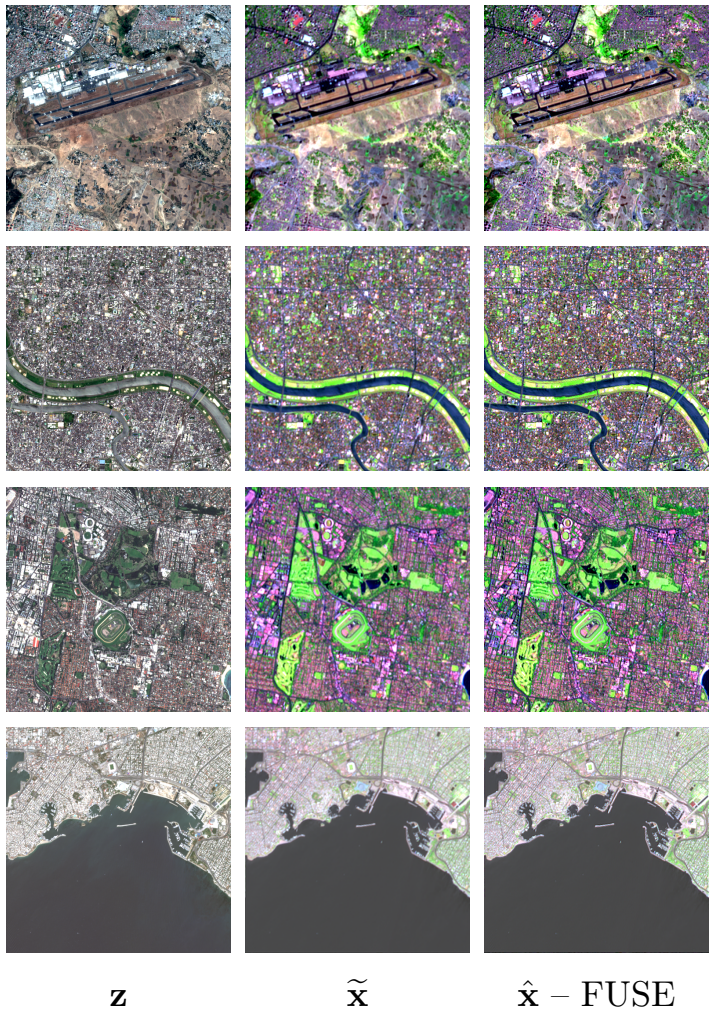


Figure 5.11. Super-resolution of the test images—Urban zones. From top to bottom: Adis Abeba, Tokyo, Sydney, and Athens. From left to right: High-resolution 10 m input component $\mathbf{z}$, low-resolution 20 m component $\widetilde{\mathbf{x}}$ to be super-resolved, and super-resolution $\hat{\mathbf{x}}$ using the FUSE algorithm.

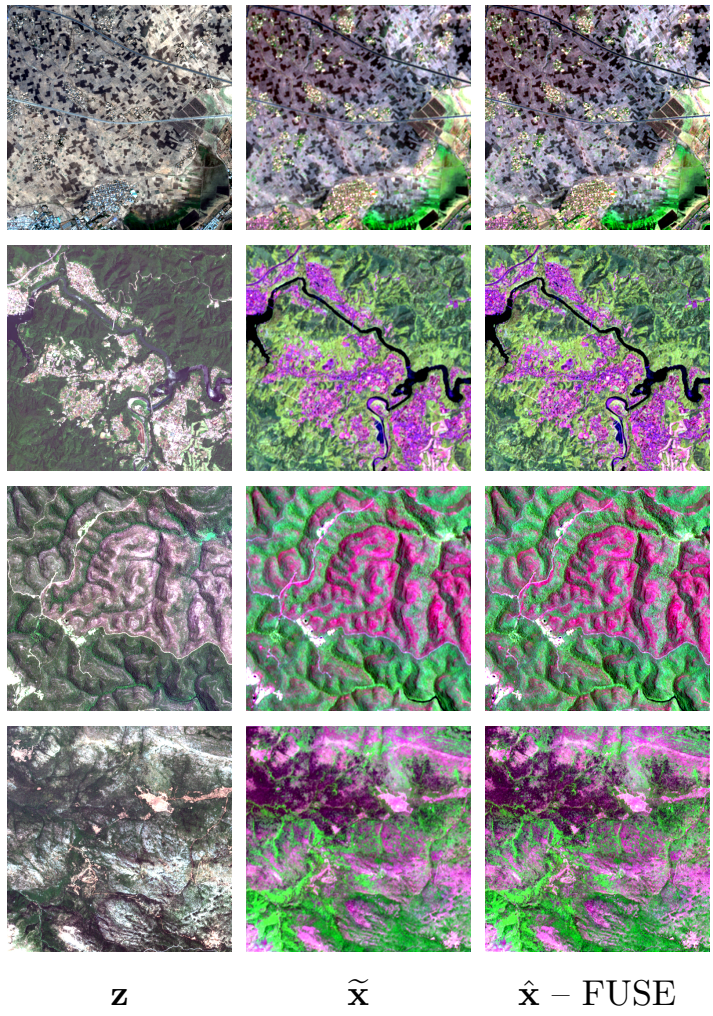


Figure 5.12. Super-resolution of the test images—Extra-urban zones. From top to bottom: Adis Abeba, Tokyo, Sydney, and Athens. From left to right: High-resolution 10 m input component $\mathbf{z}$, low-resolution 20 m component $\widetilde{\mathbf{x}}$ to be super-resolved, and super-resolution $\hat{\mathbf{x}}$ using the FUSE algorithm.

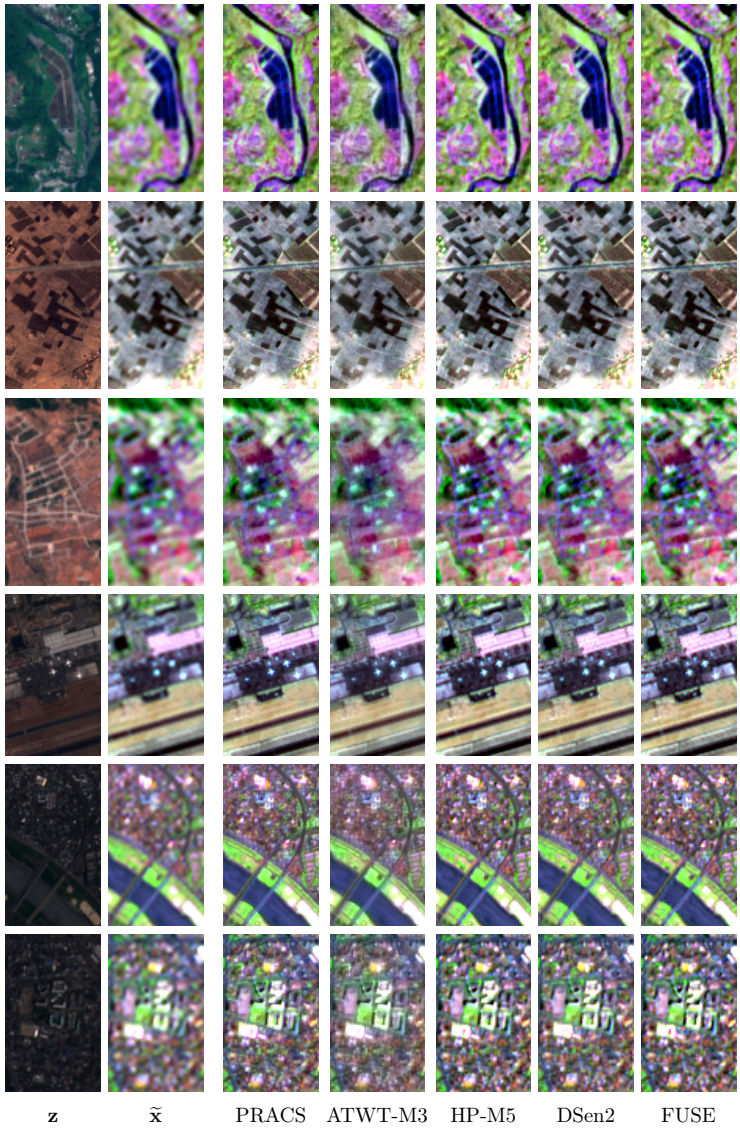


Figure 5.13. Full-resolution results for selected details. For each detail (row) from left to right are shown the two input components to be fused, followed by the corresponding fusions obtained by compared methods.

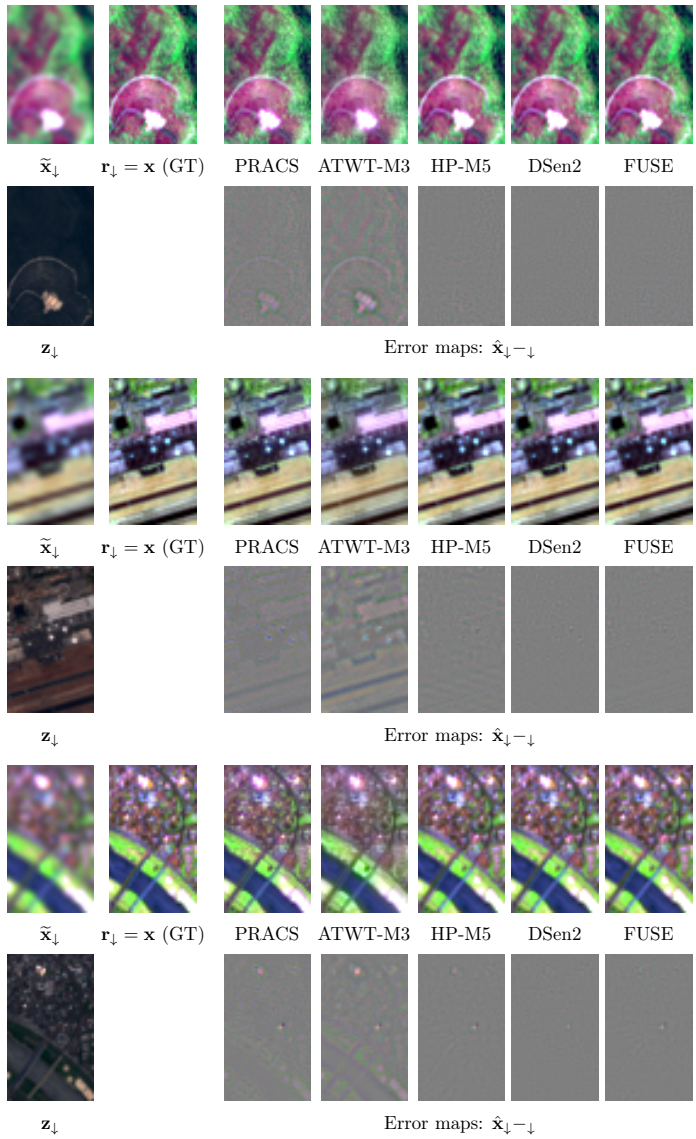


Figure 5.14. Reduced-resolution samples. Bottom images (Columns 3–7) show the difference with the ground-truth (GT).

column, followed by the reference ground-truth $\mathbf{r}_{\downarrow}$. Then, Columns 3–7 show a few selected solutions (odd rows) with the corresponding error maps (even rows) obtained as the difference between the super-resolved image and the reference, $\hat{\mathbf{x}}_{\downarrow} - \mathbf{r}_{\downarrow}$. As it can be seen, the DL methods perform better in comparison with model-based approaches as the error map is nearly constant grey, whereas, for PRACS and ATWT-M3, visible piece-wise colour shifts are introduced. This observation does not contrast with the good values of SAM obtained by PRACS, since this indicator accounts for the relative colour/band proportions but not for their absolute intensity (some "colourful" error maps in Fig. 5.14 are partially due to the band-wise histogram stretching used for the sake of visualization). Overall, by looking at numerical accuracy indicators and visual results, in both reduced- and full-resolution contexts, the proposed method provides state-of-the-art results on our datasets, as does DSen2.

5.3.4 Discussion

To assess the impact of the proposed changes concerning the baseline HP-M5, that are the additional convolutional layer and the composite loss that adds a regularization term and a structural term to the basic spectral loss (L1-norm), we also carried out an ablation study. In particular, we have the three-layer scaled version of FUSE and the four-layer version trained without regularization and structural loss terms. These two solutions are also reported in Tab. 5.6. As can be seen, except for the SAM index, the full version of FUSE outperforms consistently both scaled versions, with remarkable gains on ERGAS, in the reference-based framework and on the spatial distortion D_S, in the no-reference context. Focusing on the two ablations, the use of the composite loss has a relatively better impact than the network depth increase. This consideration is particular evident looking at the SAM indicator.

The experimental evaluation presented above confirms the great poten-

tial of the DL approach in the context of the data fusion problem at hand, as already seen for pansharpening [83] and single-image super-resolution of natural images [1] a few years ago. The numerical gap between DL methods and the others is consistent and confirmed by visual inspection. In particular, we observe that the additional structural loss term, the most relevant change concerning our previous models M5 and HP-M5, allowed us to reach and slightly overcome the accuracy level of DSen2. Beside accuracy assessment, it is worth focusing on the related computational burden. DL methods, in fact, are known to be computationally demanding, hence potentially limited for large-scale applicability. Thus, we focused from the beginning on relatively small CNN models. Indeed, the proposed model involves about 28K parameters in contrast to DSen2 has 2M parameters. In Tab. 5.8, we gather a few numbers obtained experimentally on a single GPU Quadro P6000 with 24 GB of memory. For both the proposed and DSen2, we show the GPU memory load and the computational time for the inference concerning the image size.

Table 5.7. Computational burden of FUSE and DSen2 at test time for different image sizes.

	GPU Memory (Time)				
Im. Size	**512 × 512**	**512 × 1024**	**1024 × 1024**	**1024 × 2048**	**2048 × 2048**
DSen2	6.6 GB (3.4 s)	8.7 GB (4.3 s)	9.2 GB (7.4 s)	17.4 GB (9.8)	out of memory
FUSE	391 MB (6×0.45 s)	499 MB (6 × 0.47 s)	707 MB (6 × 0.50 s)	1.1 GB (6 × 0.55 s)	1.9 GB (6 × 0.60 s)

As the proposed model is replicated, with different parameters, for each of the six bands to be super-resolved, we assume either a sequential GPU usage (as done in the Tab.5.8) or a parallel implementation 6× memory usage but also 6× faster processing. In any case, to have a rough idea of the different burden, it is sufficient to observe that, by using about one-third of the memory necessary for DSen2 to super-resolve a 512 × 512 image, FUSE can super-resolve a 16× larger image (2048 × 2048) in the same time slot. Besides, it also has to be considered that, in many applications, the user may be interested in super-resolving a single band,

hence saving additional computational and/or memory load. Finally, this picture does not consider the less critical training phase or an eventual fine-tuning stage, highlighting the advantage of using a smaller network. We recall that, according to [44], DSen2 was trained in about three days on an NVIDIA Titan Xp 12 GB GPU, whereas the training of our model took about 3 h.

5.4 Sentinel 1 and Sentinel 2 integration

The experimental analysis of NDVI estimation using CNN-based is described. It is worth underlining that no peculiar property of the NDVI was exploited, and therefore these results have a broader significance, suggesting that other image features can be better estimated by cross-sensor CNN-based fusion.

5.4.1 Estimating NDVI using Sentinel 1 and Sentinel 2 data

The area under study is located in the province of Tuy, Burkina Faso, around the commune of Koumbia. This area is particularly representative of West African semiarid agricultural landscapes, for which the Sentinel missions offer new opportunities in monitoring vegetation, notably in the context of climate change adaptation and food security. The use of SAR data in conjunction with optical images is particularly appropriate in these areas, since most of the vegetation dynamics take place during the rainy season, especially over the cropland, as smallholder rainfed agriculture is dominant. This strongly reduces the availability of usable optical images in the critical phase of vegetation growth, due to the significant cloud coverage [206] from which SAR data are only loosely affected. The 5253×4797 pixels scene is monitored from May 5th to November 1st 2016, that corresponds to a regular agricultural season in the area.

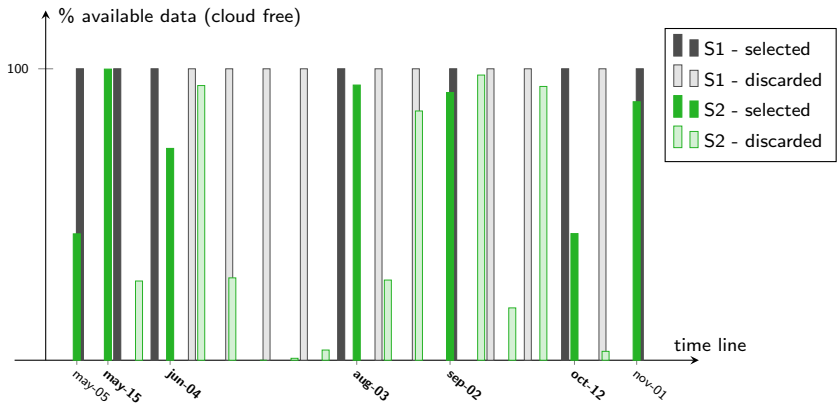


Figure 5.15. Available S1 (black) and S2 (green) images over the period of interest. The bar height indicates the fraction of usable data. Solid bars mark selected images, boldface date mark test images.

Fig. 5.15 indicates the available S1 and S2 acquisitions in this period. In the case of S2 images, the bar height indicates the percentage of data which are not cloudy. It is clear that some dates provide little or no information. Note that, during the rainy season, the lack of sufficient cloud-free optical data may represent a major issue, preventing the extraction of spatio-temporal optical-based features, like time-series of vegetation, water or soil indices, and so on. S1 images, instead, are always completely available, as SAR data are insensitive to meteorological conditions.

For the purpose of training, validation and testing of the proposed methods, we kept only S2 images which were cloud-free, or such that the spatial distribution of clouds did not prevent the selection of sufficiently large training and test areas. For the selected S2 images (solid bars in Fig. 5.15) the corresponding dates are indicated on the x-axis. Our dataset was then completed by including also the S1 images (solid bars) which are temporally closest to the selected S2 counterparts. The general idea of the proposal is to use the closest cloud-free S2 and S1 images to estimate the desired feature on the target date of interest. Therefore, among the

seven selected dates, only the five inner ones are used as targets. Observe, also, that the resulting temporal sampling is rather variable, with intervals ranging from ten days to a couple of months, allowing us to test our methods in different conditions.

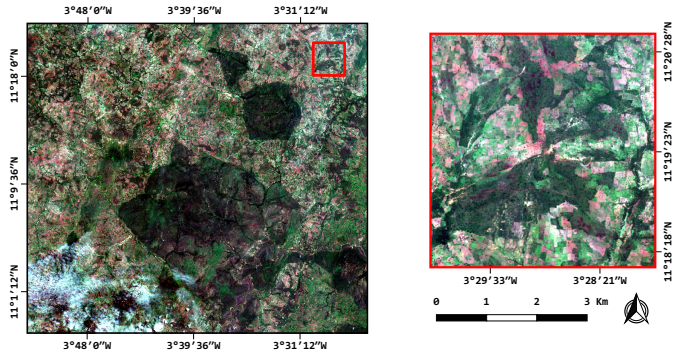


Figure 5.16. RGB representation of the 5253×4797 S2-Koumbia dataset (August 3rd, 2016), with a zoom on the area selected for testing.

To allow temporal analyses, we chose a test area, of size 470×450, which is cloud-free in all the selected dates, and hence with available reference ground-truth for any possible optical feature. Fig. 5.16 shows the RGB representation of a complete image of the Koumbia dataset (August 3rd), together with a zoom of the selected test area. Even after discarding the test area, a quite large usable area remains, from which a sufficiently large number of small (33×33) cloud-free patches are randomly extracted for training and validation.

For this work, we used Sentinel-1 data acquired in Interferometric Wide swath (IW) mode, in the high-resolution Ground Range Detected (GRD) format as provided by ESA. Such Level-1 products are generally available for most data users, and consist of focused SAR data detected in magnitude, with a native range by azimuth resolution estimated to 20×22 meters and a 10×10 meter pixel spacing. A proper multi-looking and ground range

projection is applied to provide the final GRD product at a nominal 10 m spatial resolution. On our side, all images have been calibrated (VH/VV intensities to sigma nought σ_0) and terrain corrected using ancillary data, and co-registered to provide a 10 m resolution, spatially coherent time series, using the official European Space Agency (ESA) Sentinel Application Platform (SNAP) software [207]. No optical/SAR co-registration has been performed, assuming that the co-location precision provided by the independent orthorectification of each product is sufficient for the application. Sentinel-2 data are provided by the French Pole Thématique Surfaces Continentales (THEIA) [208] and preprocessed using the Multi-sensor Atmospheric Correction and Cloud Screening (MACCS) level 2A processor [209] developed at the French National Space Agency (CNES) to provide surface reflectance products as well as precise cloud masks.

In addition to the Sentinel data, we assume the availability of two more features, the cloud masks for each S2 image, and a Digital Elevation Model (DEM). Cloud masks are obviously necessary to establish when the prediction is needed and which adjacent dates should be involved. The DEM is a complementary feature that integrates the information carried by SAR data, and may be useful to improve estimation. It was gathered from the Shuttle Radar Topographic Mission (SRTM) 1 Arc-Second Global, with 30 m resolution resampled at 10 m to match the spatial resolution of Sentinel data.

For an effective training of the networks, a large cloud-free dataset is necessary, with geophysical properties as close as possible to those of the target data. This is readily guaranteed whenever all images involved in the process, for example F_-, F and F_+, share a relatively large cloud-free area. Patches will be extracted from this area to train the network which, afterwards, will be used to estimate F also on the clouded area, obtaining a complete coverage at the target date.

For our relatively small networks ($\sim 7{\cdot}10^4$ weights to learn in the worst

case – see Tab. 4.3), a set of 19k patches is sufficient for accurate training, as already observed for other generative tasks like super-resolution [1] or pansharpening [83] addressed with CNNs of similar size. With our patch extraction process, this number requires an overall cloud-free area of about 1000×1000 pixels, namely, about 4% of our 5253×4797 target scene (Fig. 5.16). If the unclouded regions are more scattered, this percentage may somewhat grow, but remains always quite limited. Therefore, a perfectly fit training set will be available most of the times (always, in our experiments). However, if the scene is almost completely covered by clouds at the target date, one may build a good training set by searching for data spatially and/or temporally close characterized by similar landscape dynamics, or resorting to data collected in other similar sites. This case will be discussed in more detail with the help of a temporal transfer learning example in Sec. 5.4.3. In the present case, instead, for each date a dataset composed of 15200 33×33 examples for training, plus 3800 more for validation, was created by sampling the target scene with a 8-pixel stride in both spatial directions, always skipping test area and cloudy regions. Then, the whole collection was shuffled to avoid biases when creating the 128-examples mini-batches used in the SGD algorithm.

To conclude this section we present in Fig. 5.17 some preliminary results about the evolution of the loss computed on the validation dataset during the training process for a sample proposed architecture and for some deviations from it. Although the L1 loss (or mean absolute error) has not been directly considered for the accuracy evaluation presented in the next section which refers to widespread measures of quality, it is strictly related to them and can provide an rough preview of the performance. For the sake of simplicity, we gather in Fig. 5.17 only a subset of meaningful orthogonal hyperparameter variations. The first observation is that after 500 training epochs all models are about to converge and doubling such number would provide a negligible gain as tested experimentally. Decreasing the number

of layers w.r.t. the reference architecture implies a considerable performance drop. On the other side, increasing the network complexity with an additional layer does not bring any gain. The number of features is also a factor that can impact on accuracy. Fig. 5.17 reports the cases when the number of features for the first layer is changed from 48 (proposed) to either 32 or 64. In this case, however, the losses are very close to each other, with the proposed and the 64-feature case almost coincident at the end of the training. The last two plots show the impact of the learning rate α, and again the proposed setting ($5 \cdot 10^{-3}$) is "optimal" if compared with neighbouring choices (10^{-3} and 10^{-2}). It is also worth underlining that using an higher learning rate, *e.g.* 10^{-2}, one can induce a steep decay in the early phase of training which can be paid with a premature convergence.

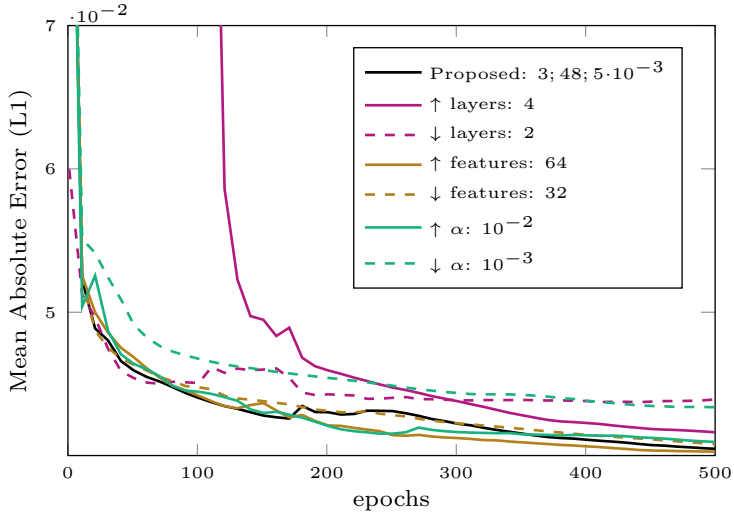


Figure 5.17. Loss functions for the validation dataset of August 3th. The proposed Optical-SAR model (with 3 layers, 48 features in the 1st layer, and $\alpha = 5{\cdot}10^{-3}$) is compared to several variants obtained by changing one hyper-parameter at time.

Besides accuracy, complexity is also affected by architectural choices.

For the same variants compared in Fig. 5.17, we report the average training time in Tab. 5.8, registered using a NVIDIA GPU, GeForce GTX TITAN X. The test time is instead negligible in comparison with that of training and is therefore neglected. For all models the total cost for training is in the order of one hour. However, as expected, increasing the number of network parameters adding layers or features impacts on the computational cost. Eventually the proposed architecture is the result of a tradeoff between accuracy and complexity.

Table 5.8. Training time in seconds for a single epoch and for the overall training (500 epochs), for different hyperparameter settings.

	Proposed	↑ layers	↓ layers	↑ features	↓ features	↑ α	↓ α
Time per epoch	6.548	7.972	4.520	7.224	5.918	6.526	6.529
Overall	3274	3986	2260	3612	2959	3263	3264

5.4.2 Experimental results

In order to assess the accuracy of the proposed solutions, we consider two reference methods for comparison, a deterministic linear interpolator (temporal gap-filling) which can be regarded as the baseline, and affine regression, both in causal and non-causal configurations. Temporal gap filling was proposed in [206] in the context of the development of a national-scale crop mapping processor based on Sentinel-2 time series, and implemented as a remote module of the Orfeo Toolbox [210]. This is a practical solution used by analysts [206] to monitor vegetation processes through NDVI time-series. Besides being simple, it is also more generally applicable and robust than higher-order models which require a larger number of points to interpolate and may overfit the data. Since temporal gap filling is non-causal, we add a further causal interpolator for completeness, a simple zero-order hold. Of course, deterministic interpolation does not take into account the correlation between available and target data, which

can help performing a better estimate and can be easily computed based on a tiny cloud-free fraction of the target image. Therefore, for a fairer comparison, we consider as a further reference the affine regressors, both causal and non-causal, optimized using the least square method. If suitable, post-processing may be included for spatial regularization, both for the reference and proposed methods. This option is not pursued here. In summary the following alternatives are considered for comparison:

$$\widehat{F} = \begin{cases} F_- & \text{Interpolator/C} \\ \frac{\Delta_+}{\Delta_-+\Delta_+}F_- + \frac{\Delta_-}{\Delta_-+\Delta_+}F_+ & \text{Interpolator ([206])} \\ a_-F_- + b & \text{Regressor/C} \\ a_-F_- + a_+F_+ + b & \text{Regressor} \end{cases}$$

where Δ_- and Δ_+ are the left and right temporal gaps, respectively, and a_-, a_+ and b satisfy

$$(a_-, (a_+), b) = \arg\min \mathrm{E}\left[\| F - \widehat{F} \|^2\right].$$

Table 5.9. Correlation index, $\rho \in [-1, 1]$.

		may-15	**jun-04**	**aug-03**	**sep-02**	**oct-12**	**average**
	gaps (before/after)	10/20	20/60	60/30	30/40	40/20	
Cross-sensor	**SAR**	0.8243	0.8161	0.5407	0.4219	0.4561	0.6118
	SAR+	0.8254	0.7423	0.3969	0.4963	0.6428	0.6207
Causal	**Interpolator/C**	0.9760	0.8925	0.6566	0.6704	0.6098	0.7611
	Regressor/C	0.9760	0.8925	0.6566	0.6704	0.6098	0.7611
	Optical/C	0.9811	0.9407	0.7245	0.7280	0.7302	0.8209
	Optical-SAR/C	0.9797	**0.9432**	0.7716	**0.7880**	0.7546	0.8474
	Optical-SAR+/C	**0.9818**	0.9424	**0.7738**	0.7855	**0.7792**	**0.8525**
Non-causal	**Interpolator**	0.9612	0.8915	0.7643	0.7288	0.8838	0.8459
	Regressor	0.9708	0.9004	0.7618	0.7294	0.8930	0.8511
	Optical	**0.9814**	0.9524	0.8334	0.758	0.9115	0.8874
	Optical-SAR	0.9775	**0.9557**	**0.8567**	0.8194	0.9002	0.9019
	Optical-SAR+	0.9781	0.9536	0.8550	**0.8220**	**0.9289**	**0.9075**

Table 5.10. Peak signal-to-noise ratio (PSNR) [dB].

		may-15	**jun-04**	**aug-03**	**sep-02**	**oct-12**	**average**
	gaps (before/after)	10/20	20/60	60/30	30/40	40/20	
Cross-sensor	**SAR**	24.30	19.52	12.34	17.30	10.70	16.83
	SAR+	23.49	17.96	14.78	16.12	19.01	18.27
Causal	**Interpolator/C**	30.11	19.48	10.62	17.70	14.59	18.50
	Regressor/C	30.86	22.60	18.30	20.39	20.02	22.44
	Optical/C	30.85	24.92	18.74	21.01	21.22	23.35
	Optical-SAR/C	31.24	**25.07**	**19.96**	21.56	20.71	23.71
	Optical-SAR+/C	**32.81**	24.90	19.79	**21.76**	**21.91**	**24.24**
Non-causal	**Interpolator**	27.91	21.97	19.12	17.41	23.61	22.00
	Regressor	30.26	22.86	20.01	21.14	24.67	23.79
	Optical	**32.61**	26.09	21.41	21.53	24.74	25.28
	Optical-SAR	29.72	**26.29**	**22.01**	**22.48**	23.89	24.88
	Optical-SAR+	31.62	25.65	21.84	22.30	**25.24**	**25.33**

Table 5.11. Structural similarity measure (SSIM) [-1,1].

		may-15	**jun-04**	**aug-03**	**sep-02**	**oct-12**	**average**
	gaps (before/after)	10/20	20/60	60/30	30/40	40/20	
Cross-sensor	**SAR**	0.5565	0.4766	0.3071	0.3511	0.2797	0.3942
	SAR+	0.5758	0.4534	0.3389	0.3601	0.3808	0.4218
Causal	**Interpolator/C**	0.9128	0.7115	0.3481	0.6597	0.6335	0.6531
	Regressor/C	0.9168	0.7364	0.4161	0.6425	0.6001	0.6624
	Optical/C	0.9557	0.8583	0.6057	0.7265	0.6671	0.7627
	Optical-SAR/C	0.9543	0.8600	0.6280	0.7539	0.6918	0.7776
	Optical-SAR+/C	**0.9565**	**0.8602**	**0.6365**	**0.7545**	**0.6989**	**0.7813**
Non-causal	**Interpolator**	0.8801	0.6798	0.6696	0.7177	0.8249	0.7544
	Regressor	0.9067	0.7330	0.6693	0.7218	0.8032	0.7668
	Optical	**0.9589**	0.8788	0.7623	0.7618	0.8470	0.8418
	Optical-SAR	0.9541	**0.8835**	**0.7780**	**0.7841**	0.8339	0.8467
	Optical-SAR+	0.9571	0.8788	0.7757	0.7834	**0.8559**	**0.8502**

The numerical assessment is carried out on the basis of three commonly used indicators, the correlation coefficient (ρ), the peak signal-to-noise ratio (PSNR), and the structural similarity measure (SSIM). These are gathered in Tables 5.9, 5.10 and 5.11, respectively, for all proposed and reference methods and for all dates. The target dates are shown in the

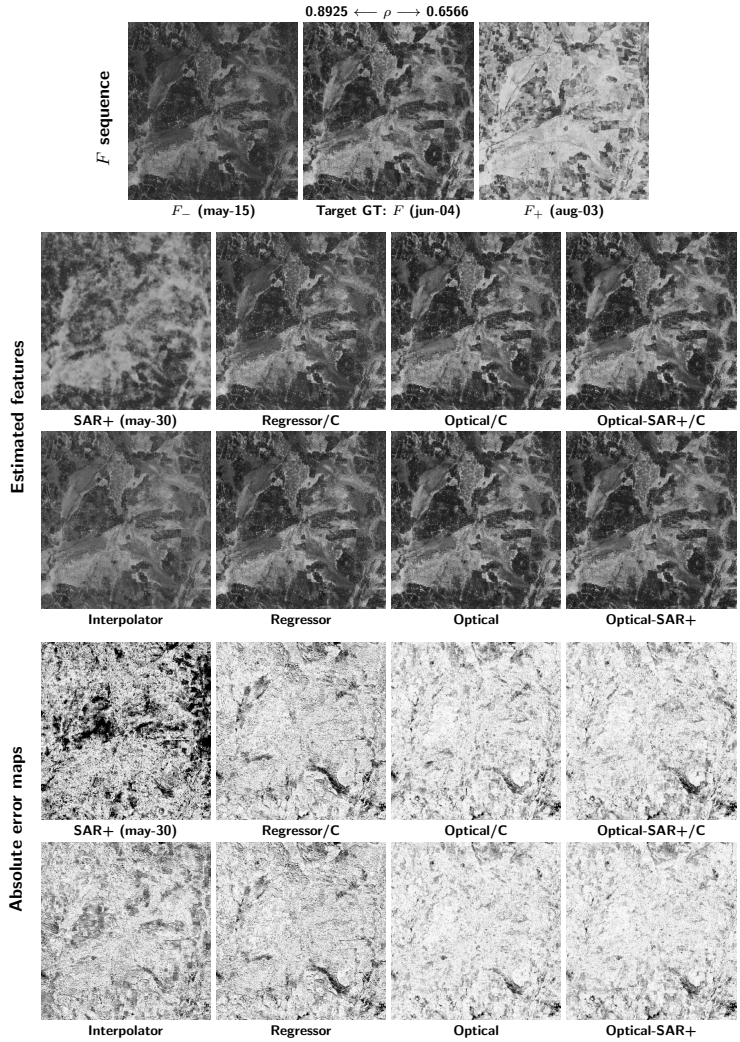


Figure 5.18. Sample results for the jun-04 target date. Top row: previous, target, and next NDVI maps of the crop selected for testing. Second/third rows: NDVI maps estimated by causal/non-causal methods. Last two rows: corresponding absolute error images.

first row, while the second row gives the temporal gaps (days) between the target and the previous and next dates used for prediction, respectively.

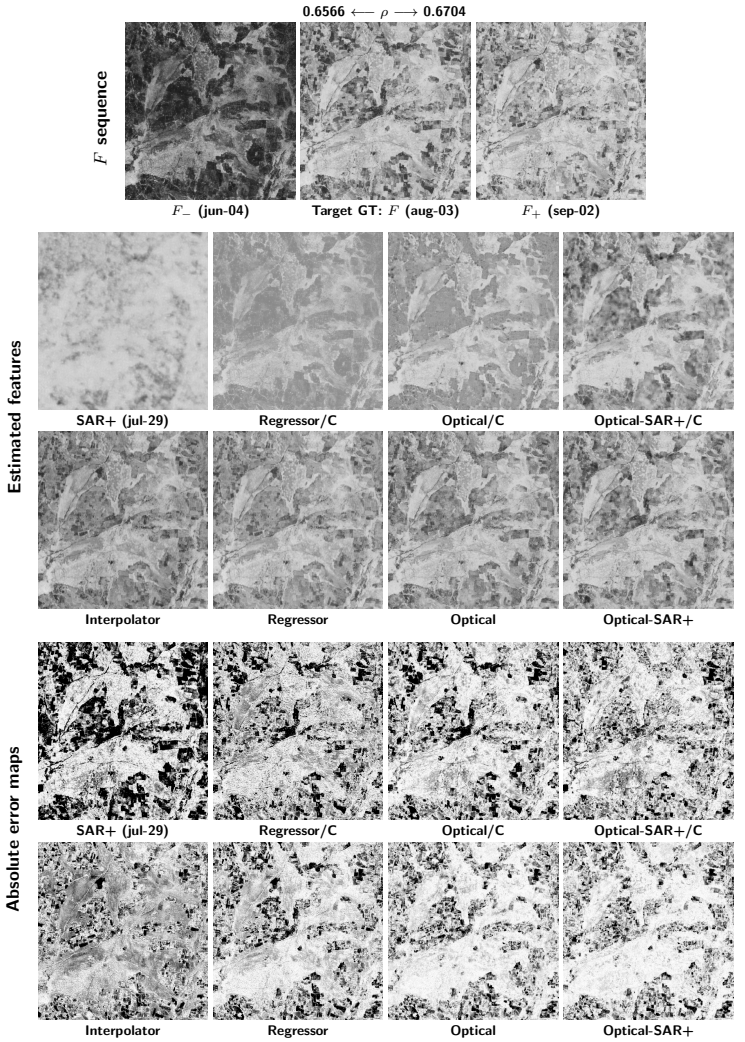


Figure 5.19. Sample results for the aug-03 target date. Top row: previous, target, and next NDVI maps of the crop selected for testing. Second/third rows: NDVI maps estimated by causal/non-causal methods. Last two rows: corresponding absolute error images.

The following two lines show results for fully cross-sensor, that is, SAR-only estimation, while in the rest of the Table we group together all causal (top)

and non-causal (bottom) models, highlighting the best performance in each group with blue text. For a complementary subjective assessment by visual inspection some meaningful sample results are shown in Figures 5.18 and 5.19.

5.4.3 Discussion and future perspective

In this section we will discuss about the accuracy of the proposed methods both objectively, through the numerical results gathered in Tabb. 5.9-5.11, and subjectively by visually inspecting Figures 5.18 and 5.19. Then we conclude the section discussing critical conditions when training data cannot be retrieved from the target.

Let us start with the numerical evaluation focusing for the time being on the ρ Tab. 5.9, and in particular on the last column with average values, which accounts well for the main trends. First of all, the fully cross-sensor solutions, based on only-SAR or SAR+DEM data, respectively, are not competitive with methods exploiting optical data, with a correlation index barely exceeding 0.6. Nonetheless, they allow one to obtain a rough estimate of the NDVI in the absence of optical coverage, proving that even a pure spectral feature can be inferred from SAR images, thanks to the dependencies existing between the geometrical and spectral properties of the scene. Moreover, SAR images provide information on the target which is not available in optical images, and complementary to it. Hence, their inclusion can help boosting the performance of methods relying on optical data.

Turning to the latter, we observe, as expected, that non-causal models largely outperform the corresponding causal counterparts. As an example, for the baseline interpolator, ρ grows from 0.761 (causal) to 0.846 (non-causal), showing that the constraint of near real-time processing has a severe impact on estimation quality.

However, even with the constraint of causality, most of this gap can

be filled by resorting to CNN-based methods. By using the very same data for prediction, that is, only F_-, the Optical/C model reaches already $\rho = 0.821$. This grows to 0.847 (like the non-causal interpolator) when also SAR data are used, and to 0.852 when also the DEM is included. Therefore, both the use CNN-based estimation and the inclusion of SAR data guarantee a clear improvement. On the contrary, using a simple statistical regressor is of little or no[2] help. Looking at the individual dates, a clear dependence on the time gaps emerges. For the causal baseline, in particular, the ρ varies wildly, from 0.610 to 0.976. Indeed, when the previous image is temporally close to the target, like for May-15, and hence strongly correlated with it, even this trivial method provides a very good estimation, and more sophisticated methods cannot give much of an improvement. However, things change radically when the previous available image is acquired long before the target, like for the Aug-03 or Oct-12 dates. In these cases, the baseline does not provide acceptable estimates anymore, and CNN-based methods give a large performance gain, ensuring a ρ always close to 0.8 even in the worst cases.

Moving now to non-causal estimation we observe a similar trend. Both reference methods are significantly outperformed by the CNN-based solutions working on the same data, and further improvements are obtained by including SAR and DEM. The overall average gain, from 0.851 to 0.907 is not as large as before, since we start from a much better baseline, but still quite significant. Examining the individual dates, similar considerations as before arise, with the difference that now two time gaps must be taken into account, with previous and next images. As expected, the CNN-based methods provide the largest improvements when both gaps are rather large, that is, 30 days or more, like for the Aug-03 and Sep-02 images.

The very same trends outlined for the ρ are observed also with reference

[2]The causal interpolator and regressor have identical ρ by definition.

to the PSNR and SSIM data, shown in Tab. 5.10 and Tab.5.11. Note that, unlike ρ and SSIM, the PSNR is quite sensitive to biases on the mean, which is why, in this case, the statistical affine regressor provides significant gains over the linear interpolator. In any case, the best performance is always obtained using CNN-based methods relying on both optical and SAR data, with large improvements with respect to the reference methods.

Further insight into the behavior of the compared methods can be gained by visual inspection of some sample results. To this end we consider two target dates, June 4th and Aug 3rd, characterized by significant temporal changes in spectral features with respect to the closest available dates. In the first case, a high correlation exists with the previous date $\rho = 0.8925$ but not with the next $\rho = 0.6566$. In the second, both correlation indexes are quite low, 0.6566 and 0.6704, respectively. These changes can be easily appreciated in the images, shown in the top row of Fig.5.18 and Fig.5.19, respectively. In both figures, the results of most of the methods described before are reported, omitting less informative cases for the sake of clarity. To allow easy interpretation of results, images are organized for increasing complexity from left to right, with causal and non-causal versions shown in the second and third row, respectively. As only exception, the first column shows results for SAR+ and non-causal interpolator. Moreover, in the last two rows, the corresponding absolute error images are shown, suitably magnified, with the same stretching and reverse scale (white means no error) for better visibility.

For jun-04, the estimation task is much simplified by the availability of the highly correlated may-15 image. Since this precedes the target, causal estimators work almost as well as non-causal ones. Moderate gradual improvements are observed going from left to right. Nonetheless, by comparing the first (interpolator) and last (Optical-SAR+) non-causal solutions, a significant accumulated improvement can be perceived, which becomes obvious in the error images. In this case, the SAR-only estimate is also

quite good, and the joint use of optical and SAR data (fourth column) provides some improvements.

For the aug-03 image, the task is much harder, no good predictor images are available, especially the previous image, 60 days old. In these conditions, there is clear improvement when going from causal to non-causal methods, even more visible in the error images. Likewise, the left-to-right improvements are very clear, both in the predicted images (compare for example the sharp estimate of Optical-SAR+ with the much smoother output of the regressor) and in the error images, which become generally brighter (smaller errors) and with fewer black patches. In this case, the SAR-only estimate is too noisy, while the joint solution (fourth column) provides a sensible gain over the others.

Table 5.12. Temporal transfer learning results for model "Optical-SAR+". (i, j) Table entry corresponds to the accuracy (ρ) obtained on the j-th date (column) when training is carried out on the i-th date (row).

	may-15	**jun-04**	**aug-03**	**sep-02**	**oct-12**
may-15	**0.9781**	0.9111	0.5782	0.4907	0.6199
jun-04	0.9542	0.9536	0.8461	0.6612	0.5285
aug-03	0.9055	**0.9661**	**0.8550**	0.8602	0.5728
sep-02	0.5535	0.6892	0.6748	0.8220	**0.9387**
oct-12	0.3357	0.5090	0.3966	**0.8981**	0.9289

To conclude this discussion let us now focus on the learning related issues. In particular a fundamental question is how to proceed when no training data can be collected from the target image at a given time (fully cloudy condition). A first idea can be to explore the possibility to generate a vegetation index from the radar data directly. However, this approach introduces an increase in time-consuming (the SNAP tool is very slow, and other tools are not very reliable) because the pre-processing of radar index requires a primary CPU and GPU resources occupation. Further, in our analysis, we want to include the estimated NDVI in a classical reliable segmentation framework based on multi-temporal NDVI. Because multi-

temporal requirements and time-consuming increase, the radar index is not suggested in this specific context. So, another idea can be to utilize a machine learning model trained elsewhere, but to what extent we can use a CNN elsewhere? This is a key problem in machine learning, and is very relevant for a number of remote sensing applications, such as coregistration [211] or pansharpening [2]. In [211] it has been underlined the importance of selecting training data which are homogeneous with the target. In [2] it is shown that the performance of a CNN can drop dramatically without a proper domain adaptation strategy and target-adaptive solution is proposed.

To gain insight into this critical point we benefit from a simple test that gives an idea of the scale of the problem. In particular, we have considered several training-test mismatches by transferring temporally the learned models. The accuracy assessed in terms of correlation index (similar results are obtained for PSNR and SSIM) for all transfer combinations is shown in Tab. 5.12. The i-th row collects to the results obtained on all dates by the model trained on the i-th date. Surprisingly, given a target date, the best model does not necessarily lie on the matrix diagonal, as in three out of five cases a model transferred from a neighbouring date outperforms the model trained on the target date. More in general, with one exception, entry (sep-02, aug-03), diagonal-adjacent values are relatively high, while moving away from diagonal (toward cross-season transfer) the accuracy deteriorates progressively. In other words, this Table suggests that when weather conditions are such that no training data can be collected from the target, one can resort to some extent to models trained in the same period of the year as the spatio-temporal landscape dynamics are likely very similar. This means also that one can refer for training to acquisitions of previous years in similar periods. It is also worth to visually inspect some related estimates. In Fig. 5.20, for two sample target dates, we show the results obtained in normal conditions or by transferring the learning from

different dates, the best (same season) and the worst (cross-season) cases. Again it can be observed that models trained within the season of the target can work pretty well. On the contrary, although preserving spatial details, when crossing the season, over or under estimate phenomena can occur. In particular, if the model is trained in the rainy season (rich vegetation) and tested in the dry season (poor vegetation) we get over estimation, while in the opposite case we get under estimation.

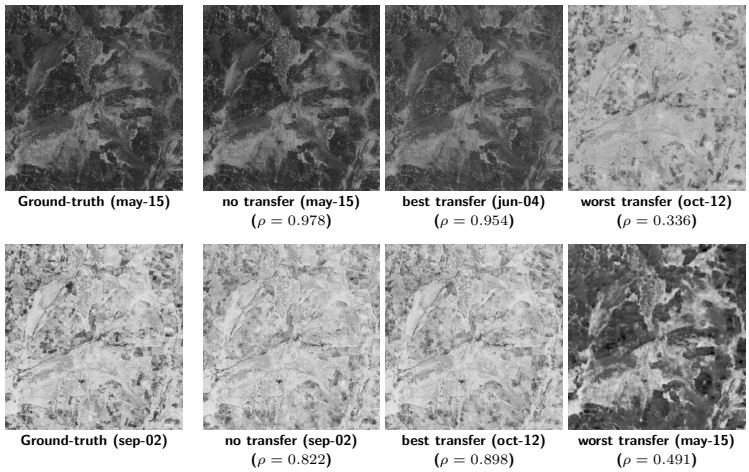


Figure 5.20. Temporal transfer learning tested on may-15 (top) and sep-02 (bottom). From left to right are the target F followed by estimates provided by model Optical-SAR+ trained on the target date (no transfer) and on two alternative dates (best and worst cases).

5.5 Deep Learning for Semantic Segmentation

So far, we have analyzed the potential of a CNN in estimating the multispectral vegetation index, the NDVI. This index and its time series are widely used in vegetation growth monitoring. As already mentioned, the limit may be, especially in tropical areas, the lack of this feature in the presence of clouds, and therefore, as we have analyzed, it becomes essential

to be able to reconstruct this feature starting from different data (SAR data) that do not undergo the influence of clouds. In this way, it is possible to have a sufficiently dense time series to determine information on the vegetative dynamics useful for farmers. Hence, as already described in the chapter on methods, we move forward a solution using the only SAR data for monitoring vegetated areas without going through the estimation of the NDVI index. The use of SAR data in the vegetation monitoring is widely known in the literature. We will evaluate in the following paragraphs the results obtained from SAR data alone, using deep neural networks, in classifying different land covers (three in the specific case) in a manner consistent with the multispectral data.

5.5.1 Proposed W-shape network versus state-of-the-art solutions

In order to build a dataset portraying rather different seasonal features we have chosen five date for training. Once left apart five 256×256 patches for testing (each from every date), 128×128 patches for training were uniformly sampled from all date in the remaining segments. In order to perform a complete evaluation we consider four full-reference numerical metrics commonly used for classification task: Accuracy, Recall, Precision and F1-score that is the harmonic average of the Precision and Recall. Overall, $12k$ patches were collected and randomly grouped in mini batches composed by 32 patches for the implementation of the ADAM-based training using the implementation of Tensorflow. Every epoch requires 100s to be completed. Furthermore other 1000 patches were also selected for the validation analysis, after the training phase.

In the numerical assessment three configurations were defined as input of the networks, and we considered the impact of these different input configurations . Both in terms of F1-score and Accuracy, the VH polarization (configuration 1) gives better results than the only VV polarization

Table 5.13. Average results in terms of the main metrics used in classification context

Models	Config.	Metrics			
		Accuracy	Precision	Recall	F1-score
FPN	1	0.8003	0.7366	0.6864	0.7085
	2	0.7384	0.7626	0.6578	0.6539
	3	0.8428	0.7470	0.7738	0.7556
Linknet	1	0.8729	0.7966	0.7841	0.7902
	2	0.7422	0.7686	0.6633	0.6591
	3	0.8965	0.8143	0.8287	0.8181
U-Net	1	0.8657	0.8110	0.7521	0.7772
	2	0.7385	0.7678	0.6578	0.6563
	3	**0.9083**	**0.8222**	**0.8441**	**0.8296**

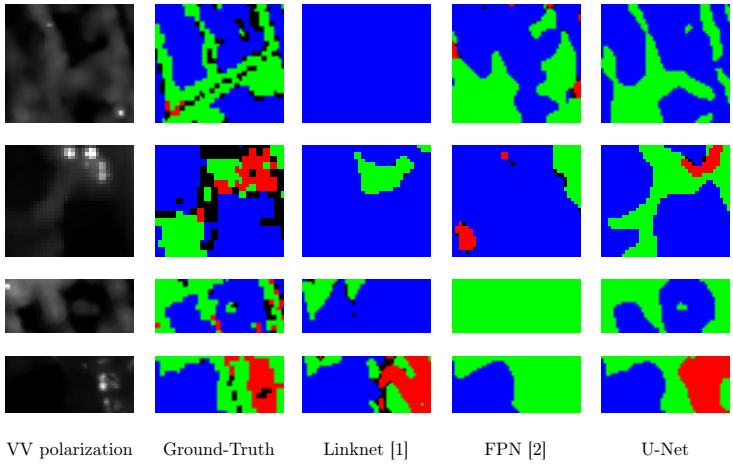


Figure 5.21. In the first and in the second columns, the S1 VV component and the Ground-Truth images, respectively. In the other columns, the maps obtained for the three analysed models.

(configuration 2) for all the considered architectures. Using the joint use of VV and VH (configuration 3) as input features, the results are better with respect to the previous ones, demostrating that the three considered

network structures benefit from the information derived by the combination of the VV and VH polarizations. In addition, observing the best performances given by the configuration 3, the goodness of the proposed U-Net appears (as highlighted in Tab. 5.13). This last aspect is more evident by a visual inspection of the Fig. 5.21, where the results of the configuration 3 are depicted. As we can see, in all the considered example scenarios the maps given by the U-Net are in good agreement with the provided Ground-Truth. In particular, the water (in blu) and the vegetation (in green) pixels are clearly recognized when the dual polarimetry is only considered. However, it sometimes failed for the bare soil class (in red). The not-classifiable pixels (in black) are completely lost in according to the fact that they have been limited during the training phase. This is a desired result, since the undefined pixels are truly associated to some of the presented classes.

5.5.2 Comparison of Pre-Trained and Trained U-Net

In Tab. 5.14 we show the results for the adopted U-Net using two design approaches and the three input configurations analyzed. In order to train the U-Net 120 s per epoch are required, i.e. 12000 s for the entire training. All the simulations are perfomed on Graphics card NVIDIA GeForce GTX 1050. As expected, the fine-tuned solution provides better results than the pre-trained one in all the configurations investigated and for all the metrics. The performance for the configuration 3 is very remarkable where an Accuracy and F1-score of 90.83% and 82.96%, respectively, are achieved. This is about 23% and 33%, respectively, better than the pre-trained strategy. Therefore, on one hand the presented results demonstrate the limited quality of this latter solution. On the other, further confirmation about the advantage of using joint information from different polarizations appears.

Table 5.14. Comparison between the U-Net with a pre-trained weights and U-Net trained on specific dataset.

		Metrics			
Config.	U-Net	Accuracy	Precision	Recall	F1-score
1	Pre-Trained	0.4591	0.6754	0.2551	0.1542
	Fine-Tuned	0.8657	0.8110	0.7521	0.7772
2	Pre-Trained	0.5524	0.4100	0.0700	0.1176
	Fine-Tuned	0.7385	0.7678	0.6578	0.6563
3	Pre-Trained	0.6729	0.7299	0.3321	0.4515
	Fine-Tuned	**0.9083**	**0.8222**	**0.8441**	**0.8296**

5.5.3 Experimental Results

In this section, we first report a brief description of the evaluation metrics used in the classification tasks and defined before, and the comparative methods (Section 5.5.4). Then, we provide numerical and visual results (Section 5.5.5).

5.5.4 Classification Metrics

To assess our model, we use a set of metrics derived from the confusion matrix. The confusion matrix allows us to give a complete evaluation of the performance, as shown in Fig. 5.22. In fact, all wrong and correct predicted values are reported in this Table. Thus, the main metrics used in classification evaluation are accuracy, precision, recall, and F1-Score. Ideally, the errors (counted in FP and FN) should be equal to zero, and consequently, accuracy, precision, recall and F1 score would be equal to 1. These metrics require reference and, likewise for the training data, we have produced the target using the indices-based technique from the Sentinel-2 L2A product. However, this indices-based classification is affected by errors, that constitute a limitations in training phase. Of course in order to further improve the performance of the deep learning approaches the target must be more accurate.

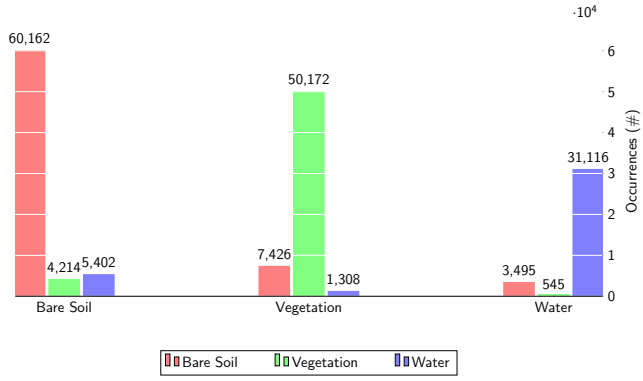


Figure 5.22. Bar plot of the Confusion Matrix.

5.5.5 Compared Methods

In this section, we compare the proposed architecture with others, commonly used in segmentation task: shallowNet [212], SegNet [213], LinkNet [214], U-Net [173], and FPN [215]. To understand the main differences between all these state-of-the-art(SoA) solutions, we start from the first part of the W-Net that corresponds to the U-Net solution [173]. But, in this comparative analysis the U-Net has the double of kernels in every convolutional layers. Then, the LinkNet has the same architecture of U-Net, but the concatenation layers are replaced by sum operator. In U-Net and LinkNet solutions, the disadvantages with respect to the proposed architecture are the lower number of convolutional levels, and the huge number of parameters. But the number of parameters in some cases could be an advantage, because it is possible to detect more complex link between input and output. Besides, the SegNet is characterized by the same encoder-decoder structure, but every block of the decoder path upsamples its input using the transferred pool indices from its corresponding block of the encoder path. This approach highly loses spatial information, with a limited number of parameters. The FPN and the shallowNet correspond

to a different basic architecture. In fact the shallowNet is a cascade of four convolutional layers interleaved by ReLu activation function, and the FPN is an architecture based on a parallel analysis of the input at different scales. This capability of the FPN allows to simultaneously obtain a segmentation at different scales. In Tab. 5.15 we show the number of learnable parameters and memory occupation of the state-of-the-art (SoA) considered architectures. Specifically, all architectures use the max pooling strategy, except for the shallowNet that is the lightest CNN with the lowest number of parameters. ShallowNet can be helpful in presence of small datasets, as in this case. However, in the specific context, the other solutions are trained starting from pre-trained weights, and so they obtained better results than the shallowNet. All SoA deep learning solutions are trained on all input stack configurations, and are adapted on the different dimensions of the input stacks. Some SoA competitors have a lot of parameters, and this amount of learnable parameters allow to obtain appreciable results, and give an idea of the goodness of these deep learning methods, confirmed by the results obtained by the proposed W-Net. In fact our method improves the performance in terms of all metrics, in particular in the whole configuration with the three dates in the input stack.

Table 5.15. General information about the SoA architectures.

Models	**# Parameters**	**Time per Epoch [s]**	**Memory**
ShallowNet	45.6k	32.0	191k
SegNet	1.8M	63.2	7.17M
FPN	6.9M	711.0	26.8M
LinkNet	4.1M	428.8	30.9M
U-Net	8M	209.2	30.9M
Proposed	1.2M	89.6	4.75M

5.5.6 Numerical and Visual Results

In order to build a dataset with different rice growing phase we collected twelve dates for training, one for each month. This dataset is more general than the data used in our previous work [143], and gives more information about the site under investigation. In fact, in the above-mentioned work only 5 dates were used to train SoA architectures. Even in this case we only left apart ten 128×128 patches for testing (each around the lake of the Natural Park, and from different dates), 128×128 patches for training were selected from all dates in the remaining segments. Overall, $1k$ patches were collected and randomly grouped in mini batches composed by 32 patches for the implementation of the Adam-based training. Furthermore other 1000 patches were also selected for the validation analysis. The training phase was performed for just 5 epochs. The average assessments over the test images are gathered in Tab. 5.16. This table provides objective evaluation of the results, and in fact we can see that the proposed solution outperforms all compared SoA solutions on all indicators. Instead, in Tab. 5.17 we report the same analysis in terms of all metrics, but in the bottom of the Table (last four rows) it is possible to understand the importance of the dual polarization and multi-temporal approach in input stack. In fact, the third (dual polarization) configuration gives better results in terms of accuracy and F1-score than the single ones. On the other hand, the dual polarization configuration (III in Tab. 5.17) shows a highly worse precision compared to the configuration I, the VH polarization configuration.

Observing the definition (5.2.2), it is clear that this technique overestimates the false positive. Further, in this bottom part of the Tab. 5.17, it is also possible to underline the importance of the multi-temporal configuration. In fact, the fourth configuration outperforms all others. In order to perform a complete numerical evaluation we consider the confusion matrix (Fig. 5.22) related to the prediction of the proposed method. From this Figure, we deeply observe the capability of the W-Net to distinguish

Table 5.16. Numerical Results for all architectures in the whole configuration, that means three dates and dual polarisation.

Methods	Metrics			
	Accuracy	**Precision**	**Recall**	**F1**
ShallowNet	0.8271	0.7743	0.7651	0.7639
SegNet	0.8240	0.7691	0.7610	0.7596
FPN	0.8418	0.8107	0.7746	0.7801
LinkNet	0.8310	0.8083	0.7623	0.7667
U-Net	0.8846	0.8567	0.8318	0.8405
Proposed	**0.9121**	**0.8860**	**0.8682**	**0.8762**

Table 5.17. Comparison between single polarization, single date and multi date in terms of all metrics.

Methods	Configuration	Metrics				Time per Epoch
		Accuracy	**Precision**	**Recall**	**F1**	
U-Net	III	0.7938	0.7503	0.7002	0.6812	**177.6**
U-Net	IV	**0.8846**	**0.8567**	**0.8318**	**0.8405**	209.2
Proposed	I	0.7162	0.7306	0.6091	0.5832	70.2
Proposed	II	0.7263	0.6865	0.6280	0.5806	**63.6**
Proposed	III	0.735	0.6563	0.6299	0.6213	81.0
Proposed	IV	**0.9121**	**0.8860**	**0.8682**	**0.8762**	89.6

considered classes, and in particular a good capability to separate vegetation from water. Infact we notice that the pixels of the vegetation class correspond to the water value in 2.22% of cases, and the water pixels to the vegetation ones in 1.50%. Instead, according to Section 4.8, the bare soil class is more difficult to classify, but the shown results for this class seem to be as accurate as the ones for the other classes. In fact, the pixels of the bare soil class do not correspond to the right value, in 13.11% of cases. Moreover, the vegetation pixels are wrongly recognized as bare soil in 12.61% of cases, and the water pixels in 9.94%. Let us now look at some samples from the test images. Samples in Fig. 5.23 show some details used for test, associated with different rice growing conditions. From this visual inspection we compare some results of the SoA approaches, and the proposed one. The visual analysis allows us to understand the results of the

proposed methods. In fact, in the first row we can see that our proposed W-Net correctly consider a large part of the pixels with respect to the others, that wrongly detect the red pixels. In all these shown details, as in the second and in the third rows, we can see that the small lines are totally lost, but the errors are lower than in other solutions. Moreover, as we can see in Fig. 5.23, the problem is also related to not accurate reference.

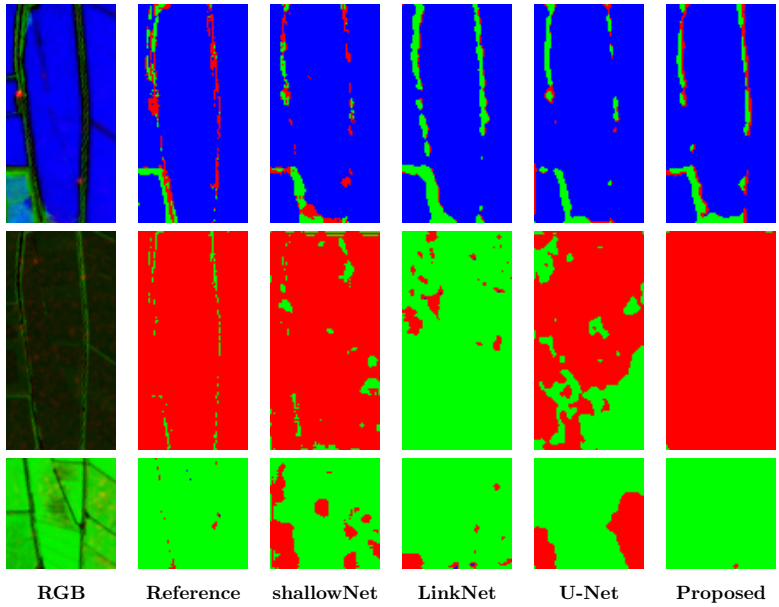


Figure 5.23. Zoomed details of segmentation results for a subset of SoA approaches, and our proposed method (W-Net). In the first column, a false colour images (R: VH, G: NDVI, B: MNDWI) is shown; in the other columns, the segmentation maps obtained with the methods under comparison are depicted. In all the segmentation maps green, red and blue pixels represent Vegetation, Bare Soil and Water, respectively.

5.5.7 Discussion

In this section, we explore in more details some results underlining the importance of using multi-date input stack and a lighter architecture to

ensure good performances and preserve the time consuming aspect.

5.5.8 Single Date and Multi Date

In Tab. 5.16 we have seen that the proposed architecture outperforms all others, and in Tab. 5.17 we underline the improvement in multi date configuration. In particular, we focus on U-Net and the proposed W-Net. In multi-temporal configuration we can assume that the convolutions of the considered architectures mitigate the effect of the multiplicative speckle noise. Furthermore, each considered competitor has a consistent improvement in multi date configuration with respect to the single date ones. This gain in terms of all metrics is really remarked in W-Net, and this is because in W-Net we have much more convolution layers, and a lower number of parameters than in the U-Net. Thus, we assume that more levels of convolution in W-Net than in U-Net correspond to a better multi-temporal despeckling effect, as confirmed in [216]. Taking into account that in the Sentinel-1 pre-processing we did not consider the speckle filter, we obtain a good segmentation maps with a reduction in terms of computational complexity. This is underlined by the W-Net that increases its performances more than the U-Net in multi date configuration without increasing too much the training times. In fact the W-Net solution highly improves the performance (+0.2 for accuracy, +0.14 for precision, +0.28 for recall, and +0.29 for F1-score) accepting an increase in training time of just 8.6 s per epoch, instead the U-Net has a lower increase in performance with a considerable increase in training times (+21.6 s per epoch). Since all the simulations are perfomed on Graphics card NVIDIA GeForce GTX 1050 and the training phase is carried out for just 5 epochs, the total training time is equal to 448 s (just over 7 min) and 1046 s (just over 17 min), respectively. For further validation, we consider the visual inspection (in Fig. 5.24), and in particular we can notice that in the first row of Fig. 5.24 the multi-temporal configuration is able to identify the

linear paths visible in target image and in the false colour RGB. Especially in the W-Net these bare soil linear paths are not present in the single date configuration. Furthermore, in both networks with the multi-date configuration we are able to better classify these bare soil pixels that in the single date U-Net solution are incorrectly recognized as vegetation (V) and in the W-Net as water (B). In the second row of Fig. 5.24 we can see that in particular the W-Net improves the classification reducing the misclassified pixels, instead the multi-data U-Net in this specific case obtains a worse classification than the single-date one. Thus, the use of deeper input stacks improves the accuracy of segmentation maps. Furthermore, in order to provide a complete comparison and analysis about the performances, we repeated the training phase 20 times for these two architectures using different initial random weights. From these simulations, we concluded that the proposed method achieves greater robustness and repeatability than the U-Net. In fact, the standard deviations of Accuracy, F1, Precision and Recall turn out to be 0.009, 0.013, 0.011, and 0.012, respectively, for the W-Net, and 0.081, 0.061, 0.053, 0.076, respectively, for the U-Net.

5.5.9 Computation Time, Number of Parameters and Memory Occupation

In this section, the results are compared in terms of architecture complexity, and in terms of computation time. The proposed W-Net has the better trade-off between computational complexity, memory occupation, and performances. In fact this proposed architecture is the second for memory occupation, time, and is largely better than all other networks in terms of all metrics used in this paper to evaluate the goodness of the obtained segmentation maps. The shallowNet has the lowest number of parameters, and consequently the lowest time per epoch, and memory occupation, however its performances are worse than the more deeper solutions. On one hand, the proposed architecture is extremely deep, but is

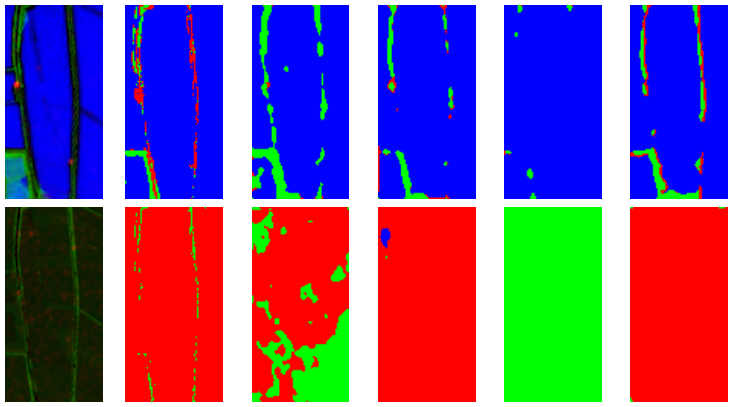


Figure 5.24. Zoomed details of segmentation results. In the first column, a false colour images (R: VH, G: NDVI, B: MNDWI) is shown; in the other columns, the segmentation maps obtained with the methods under comparison are depicted. In all the segmentation maps green, red and blue pixels represent Vegetation, Bare Soil and Water, respectively. With (Config. III) we consider the dual polarization configuration both for the U-Net and the proposed W-Net.

composed of a small number of parameters. This feature gives noticeable advantages. On the other hand this became the main disadvantage of the proposed method in presence of greater datasets. In fact greater datasets require more parameters to avoid the overfitting problem.

5.6 Limitations and Future Perspectives

The main limitation in the case of super-resolution is related to spectral distortion reduction [217, 218]. The levels reached are certainly sufficient but still improvable, especially if a semantic segmentation is to be associated downstream of the super-resolution of these bands. Especially in a context like this, a further improvement in terms of spectral distortion can be substantial. One way to improve this aspect would be to introduce

an additional Loss term based precisely on the angle formed between two different spectral bands. The Loss could be developed like this following definition in [219]:

$$\mathscr{L}_{\text{bands}} = 1 - \frac{1}{Q}\sum_{k=1}^{Q}\frac{y^{(i)}\hat{y}^{(i)}}{||y^{i}||_2||\hat{y}^{(i)}||_2} \tag{5.1}$$

where $y^{(i)}$ is the i-th column of y that represents the spectral information of the i-th pixel of the target y and $\hat{y}^{(i)}$ is the i-th column of $\hat{y}$ that denotes the spectral information of the output $\hat{y}$. The final Loss would then consist of 4 terms, as shown below:

$$\mathscr{L} = \lambda_1\mathscr{L}_{\text{Spec}} + \lambda_2\mathscr{L}_{\text{Struct}} + \lambda_3\mathscr{L}_{\text{Reg}} + \lambda_4\mathscr{L}_{\text{bands}} \tag{5.2}$$

At that point, it remains to test how to weigh the different Losses, as previously done for the three terms.

Although we are considering a multi-temporal solution in the segmentation case, we do not consider previous successful segmentation, but only the dynamics on an input composed of three dates. The information linked to a previous segmentation could be used to help the final segmentation, in the sense that one could move towards a differential solution, that is to consider a refresh of the segmentation only in the points in which the difference between the instant and the instant t + 1 is more significant than a certain threshold, it too could be trainable using a neural network [220, 221]. The flowchart of this possible future approach could be the one represented in Fig. 5.25, which highlights how there is a significant part of the two images at different times that has not changed which is therefore not analyzed in the step relating to the classification of the image at time t + 1.

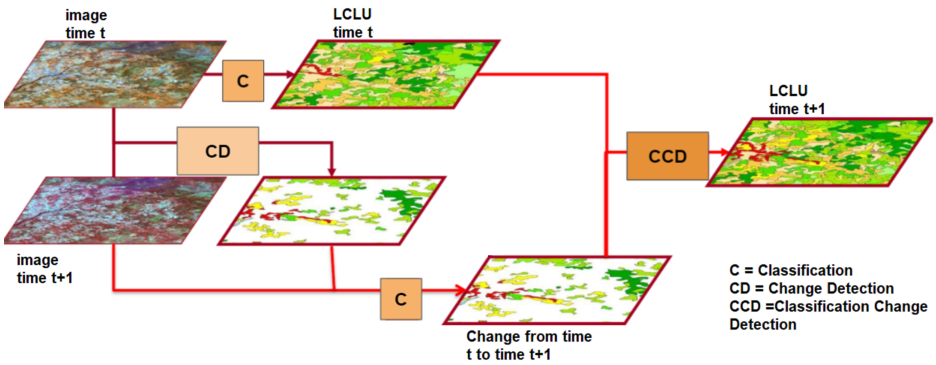


Figure 5.25. General Flowchart for change detection in semantic segmentation task.

This page intentionally left blank.

Chapter 6

Conclusions

In this thesis, a deep learning framework is designed, implemented and validated for selected remote sensing applications. Remote sensing in general and Sentinel constellations, in particular, are becoming increasingly important in monitoring land use, forests, or in creating hydrological models. The free availability of Sentinel data encourages their widespread usage. Furthermore, free access to data allows for a continuous collection of data that requires processing approaches to handle a considerable amount of data. In this thesis, I developed deep learning algorithms to increase the spatial resolution of remote sensing images (with the use of pansharpening or super-resolution algorithms) and guarantee continuous monitoring (with an appropriate combination of data from the different sensors). Deep learning requires a considerable amount of data and high computing power. These requirements are often not available for many research groups, so deep learning solutions are set aside. This thesis work provided an analysis on these aspects and provided answers to the following three questions:

- Are we able to increase spatial resolution? Moreover, are we able to improve the classic and state-of-the-art methods?

- Are we able to merge information from different satellites? So thickening the time series?
- Are we able to realize the previous objectives with shallow neural networks?

The answer to the first question was given in a series of works that have shown solutions based on shallow CNNs. CNN, even if not particularly deep, have proved to be particularly effective in these works. In fact, in the super-resolution of Sentinel-2 bands, I gave a comparable result, if not even superior, to deep nets proposed in [44]. This result was obtained by taking into account two essential aspects of a network's training phase: the pre-processing of Sentinel 2 data and the choice of the appropriate loss function. It was shown that high-pass filtering accelerates the convergence of the network. In fact, with few epochs and a limited dataset, the performances were compared with a more complex and deeper CNN-based state-of-the-art technique that used a particularly numerous training dataset [44]. Training on a large dataset corresponds to the main limitation. In fact, it cannot be done on any GPU. A method based on a large dataset and on a deep neural network requires excessive resources and time, and this aspect is primarily analyzed in this work, in which the effectiveness of a smaller network in the computational savings was also shown. However, it is possible to use the pre-trained or fine-tuned versions of the deeper solutions, but the problem with these two possibilities is the time in obtaining the desired result. Indeed, for a large-scale analysis on complete Sentinel-2 tiles, for example, the deeper solution cannot be obtained with not too high computing resources, as shown in our results. Thus, the computational resources saving and the excellent performances obtained by these methods are the main reasons why this part of the thesis work can find an interest in the remote sensing community. In particular, because the S2 bands at 20-m are widely used in many applications,

a better resolution with low resource usage can interest different research groups facing remote sensing perspectives in the most disparate fields. After answering the first and third questions and discussing some conclusions related to the spatial resolution of the Sentinel-2 sensor, we faced the second question, and in particular on the use of Sentinel-1 data to integrate the Sentinel-2 data. Firstly, the possibility of estimate a specific Sentinel 2 feature was evaluated starting from Sentinel 1 and Sentinel 2 data fusion: the NDVI estimation problem. This approach allowed us to eventually use this result to thicken a time series of NDVIs used for land use classifications. Furthermore, it encouraged subsequent works, in which starting from Sentinel-1 data, it was possible to directly estimate some classes (i.e. vegetation, bare soil and water) useful for wetland and rice field monitoring. The considered classes were obtained from simple approaches based on thresholding the multispectral Sentinel-2 indices. This classification is obtained with less effort from a computational point of view since the preprocessing of Sentinel-2 is generally faster. This threshold-based product from Sentinel-2 data was then used as output in appropriate deep neural network training (with few learnable parameters) that uses the Sentinel-1 data as input. The encouraging results of this work allowed us to state that it is possible to merge information from different satellites with the help of shallow neural networks with consequent resources' saving. In this case, the main limitation is the spatial resolution of the considered satellites, and of course, the coregistration between the datasets can be an important contribution for the improvement of future works. Further, in the future, we will extend these approaches to other case studies and new datasets so that both spatial and temporal resolution can be improved. For example, on the one hand, many works in the literature use PlanetScope data (provided at 3-m) to increase the spatial resolution of Sentinel-2. On the other hand, other sensors, such as Landsat-8 (with a worse resolution than Sentinel-2), can be used to increase temporal resolution. However,

data such as Landsat-8, S2 and PlanetScope all suffer from the same problem, i.e. it cannot be seen below the clouds, so the best way to tackle time series condensation is to exploit data from SAR satellites, such as Sentinel-1 and Cosmo-Skymed. Obtaining a suitable data fusion without data compatibility problems is necessary to make the necessary corrections to adapt to the different datasets. In all these steps, some companies' computing power, such as Google Earth Engine, regarding data collection, or Kaggle/Google Colab for collecting datasets and use of shared resources would allow the acceleration of the training of models based on CNNs. The tools provided by Google Earth Engine, Kaggle or Google Colab can undoubtedly be of great help and are already being used extensively to train deeper and more data-hungry neural networks. However, as demonstrated in our works, appropriate processing and the proper knowledge and mastery of data are always required since, without it, there is the risk of wasting resources that could, instead, be better used. The problem of the deep learning model is precisely that of being data-dependent. Therefore, in-depth knowledge of the physical phenomena involved in image creation and appropriate data processing allows the neural network to make the most of its potential to achieve more complex objectives. Thus, future developments can be sure of using different satellites with higher resolutions in super-resolution task and the introduction of a more specific loss that considers the spectral information of the bands. Furthermore, we can investigate the more accurate ground truth from a ground inspection or use other satellites with higher spatial resolution in the integration task. However, we also have to consider different input datasets in both these cases because S1 provides a limited spatial resolution. Instead, a further improvement can be introduced by a differential approach in which we only consider the modifications to previous stable and accurate segmentation.

Bibliography

[1] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[2] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *ArXiv e-prints*, 2017. arXiv: 1709.06054 [cs.CV].

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114.

[4] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[5] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[6] Y. Bengio, *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[7] P. E. Utgoff and D. J. Stracuzzi, "Many-layered learning," *Neural computation*, vol. 14, no. 10, pp. 2497–2529, 2002.

[8] Y. Bengio, Y. LeCun, *et al.*, "Scaling learning algorithms towards ai," *Large-scale kernel machines*, vol. 34, no. 5, pp. 1–41, 2007.

[9] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," Citeseer, 2014.

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

[11] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, Y. Cun, *et al.*, "Learning convolutional feature hierarchies for visual recognition," *Advances in neural information processing systems*, vol. 23, pp. 1090–1098, 2010.

[12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`.

[14] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *ICML*, 2011.

[15] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[16] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, "Advances in optimizing recurrent networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2013, pp. 8624–8628.

[17] I. Sutskever, *Training recurrent neural networks*. University of Toronto Toronto, Canada, 2013.

[18] I. Kodrasi and H. Bourlard, "Single-channel late reverberation power spectral density estimation using denoising autoencoders.," in *INTERSPEECH*, 2018, pp. 1319–1323.

[19] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural networks*, vol. 2, no. 6, pp. 459–473, 1989.

[20] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning*, vol. 42, no. 1, pp. 177–196, 2001.

[21] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[22] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA., vol. 1, 1967, pp. 281–297.

[23] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe, "Global optimization of a neural network-hidden markov model hybrid," *IEEE transactions on Neural Networks*, vol. 3, no. 2, pp. 252–259, 1992.

[24] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[25] I. Goodfellow, A. Courville, and Y. Bengio, "Large-scale feature learning with spike-and-slab sparse coding," *arXiv preprint arXiv:1206.6407*, 2012.

[26] J. Chen and X. Liu, "Transfer learning with one-class data," *Pattern Recognition Letters*, vol. 37, pp. 32–40, 2014.

[27] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," pp. 391–407, 2016.

[28] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," pp. 2980–2988, 2017.

[30] W. Guo, W. Yang, H. Zhang, and G. Hua, "Geospatial Object Detection in High Resolution Satellite Images Based on Multi-Scale Convolutional Neural Network," *Remote Sensing*, vol. 10, no. 1, p. 131, 2018.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," pp. 2672–2680, 2014.

[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.

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

[34] R. Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," pp. 91–99, 2015.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[37] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, no. 7, p. 594, 2016.

[38] Y. Rao, L. He, and J. Zhu, "A residual convolutional neural network for pan-shaprening," pp. 1–4, 2017.

[39] G. Vivone, L. Alparone, J. Chanussot, M. D. Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[40] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[41] G. Scarpa, M. Gargiulo, A. Mazza, and R. Gaetano, "A CNN-based fusion method for feature extraction from sentinel data," *Remote Sensing*, vol. 10, no. 2, 2018, ISSN: 2072-4292. DOI: 10.3390/rs10020236. [Online]. Available: http://www.mdpi.com/2072-4292/10/2/236.

[42] N. Brodu, "Super-resolving multiresolution images with band-independent geometry of multispectral pixels," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4610–4617, Aug. 2017.

[43] C. Lanaras, J. Bioucas-Dias, E. Baltsavias, and K. Schindler, "Super-resolution of multispectral multiresolution images from a single sen-

sor," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jul. 2017.

[44] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 305–319, 2018.

[45] M. Gargiulo, A. Mazza, R. Gaetano, G. Ruello, and G. Scarpa, "A CNN-Based Fusion Method for Super-Resolution of Sentinel-2 data," *IGARSS*, 2018.

[46] M. Gargiulo, "Advances on cnn-based super-resolution of sentinel-2 images," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2019, pp. 3165–3168.

[47] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "Pannet: A deep network architecture for pan-sharpening," pp. 5449–5457, 2017.

[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," pp. 448–456, 2015.

[49] M. Gargiulo, "Advances on cnn-based super-resolution of sentinel-2 images," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019.

[50] M. Drusch et al., "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote Sensing of Environment*, vol. 120, no. Supplement C, pp. 25–36, 2012, The Sentinel Missions - New Opportunities for Science, ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2011.11.026. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425712000636.

[51] M. Mura, F. Bottalico, F. Giannetti, R. Bertani, R. Giannini, M. Mancini, S. Orlandini, D. Travaglini, and G. Chirici, "Exploiting the capabilities of the sentinel-2 multi spectral instrument for predicting growing stock volume in forest ecosystems," *International Journal of Applied Earth Observation and Geoinformation*, vol. 66, pp. 126–134, 2018.

[52] J. A. A. Castillo, A. A. Apan, T. N. Maraseni, and S. G. Salmo, "Estimation and mapping of above-ground biomass of mangrove forests and their replacement land uses in the philippines using sentinel imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 134, pp. 70–85, 2017.

[53] J. G. P. W. Clevers, L. Kooistra, and M. M. M. Van den Brande, "Using sentinel-2 data for retrieving lai and leaf and canopy chlorophyll content of a potato crop," *Remote Sensing*, vol. 9, no. 5, p. 405, 2017.

[54] C. Delloye, M. Weiss, and P. Defourny, "Retrieval of the canopy chlorophyll content from sentinel-2 spectral bands to estimate nitrogen uptake in intensive winter wheat cropping systems," *Remote Sensing of Environment*, vol. 216, pp. 245–261, 2018.

[55] F. Paul, S. H. Winsvold, A. Kääb, T. Nagler, and G. Schwaizer, "Glacier remote sensing using sentinel-2. part ii: Mapping glacier extents and surface facies, and comparison to landsat 8," *Remote Sensing*, vol. 8, no. 7, p. 575, 2016.

[56] K. Toming, T. Kutser, A. Laas, M. Sepp, B. Paavel, and T. Nõges, "First experiences in mapping lake water quality parameters with sentinel-2 msi imagery," *Remote Sensing*, vol. 8, no. 8, p. 640, 2016.

[57] M. Immitzer, F. Vuolo, and C. Atzberger, "First experience with sentinel-2 data for crop and tree species classifications in central europe," *Remote Sensing*, vol. 8, no. 3, p. 166, 2016.

[58] M. Pesaresi, C. Corbane, A. Julea, A. J. Florczyk, V. Syrris, and P. Soille, "Assessment of the added-value of sentinel-2 for detecting built-up areas," *Remote Sensing*, vol. 8, no. 4, p. 299, 2016.

[59] L. Cicala, C. Angelino, N. Fiscante, and S. Ullo, "Landsat-8 and Sentinel-2 for fire monitoring at a local scale: A case study on Vesuvius," in *2018 IEEE International Conference on Environmental Engineering (EE)*, IEEE, 2018, pp. 1–6.

[60] D. Tzelidi, S. Stagakis, Z. Mitraka, and N. Chrysoulakis, "Detailed urban surface characterization using spectra from enhanced spatial resolution sentinel-2 imagery and a hierarchical multiple endmember spectral mixture analysis approach," *Journal of Applied Remote Sensing*, vol. 13, no. 1, p. 016 514, 2019.

[61] M. Zhang, W. Su, Y. Fu, D. Zhu, J.-H. Xue, J. Huang, W. Wang, J. Wu, and C. Yao, "Super-resolution enhancement of sentinel-2 image for retrieving lai and chlorophyll content of summer corn," *European Journal of Agronomy*, vol. 111, p. 125 938, 2019.

[62] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[63] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral superresolution by coupled spectral unmixing," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 3586–3594.

[64] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[65] E. Ibarrola-Ulzurrun, L. Drumetz, J. Marcello, C. Gonzalo-Martın, and J. Chanussot, "Hyperspectral classification through unmixing abundance maps addressing spectral variability," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4775–4788, Jul. 2019.

[66] L. Wald, "Data fusion: Definitions and architectures–fusion of images of different spatial resolutions," *Les Presses de l'École des Mines*, 2002.

[67] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive pca approach and contourlets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1323–1335, May 2008.

[68] T.-M. Tu, S.-C. Su, H.-C. Shyu, and P. S. Huang, "A new look at ihs-like image fusion methods," *Information Fusion*, vol. 2, no. 3, pp. 177–186, 2001, ISSN: 1566-2535. DOI: http://dx.doi.org/10.1016/S1566-2535(01)00036-7. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253501000367.

[69] P. Chavez and J. Anderson, "Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic," *Photogrammetric Engineering and Remote Sensing*, vol. 57, no. 3, pp. 295–303, 1991.

[70] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: the ARSIS concept and its implementation," *Photogrammetric engineering and remote sensing*, vol. 66, no. 1, pp. 49–61, 2000.

[71] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian data fusion for adaptable image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1847–1857, Jun. 2008.

[72] A. Garzelli, "Pansharpening of multispectral images based on nonlocal parameter optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2096–2107, Apr. 2015, ISSN: 0196-2892. DOI: 10.1109/TGRS.2014.2354471.

[73] F. Palsson, J. Sveinsson, and M. Ulfarsson, "A new pansharpening algorithm based on total variation," *Geoscience and Remote Sensing Letters, IEEE*, vol. 11, no. 1, pp. 318–322, Jan. 2014, ISSN: 1545-598X.

[74] Y. Du, Y. Zhang, F. Ling, Q. Wang, W. Li, and X. Li, "Water bodies' mapping from sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the swir band," *Remote Sensing*, vol. 8, no. 4, p. 354, 2016.

[75] Q. Wang, W. Shi, Z. Li, and P. M. Atkinson, "Fusion of sentinel-2 images," *Remote Sensing of Environment*, vol. 187, pp. 241–252, 2016, ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2016.10.030. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425716304023.

[76] A. D. Vaiopoulos and K. Karantzalos, "Pansharpening on the narrow vnir and swir spectral bands of sentinel-2," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLI-B7, pp. 723–730, 2016. DOI: 10.5194/isprs-archives-XLI-B7-723-2016. [Online]. Available: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLI-B7/723/2016/.

[77] H. Park, J. Choi, N. Park, and S. Choi, "Sharpening the vnir and swir bands of sentinel-2a imagery through modified selected and synthesized band schemes," *Remote Sensing*, vol. 9, no. 10, p. 1080, 2017.

[78] M. Gašparović and T. Jogun, "The effect of fusing sentinel-2 bands on land-cover classification," *International Journal of Remote Sensing*, vol. 39, no. 3, pp. 822–841, 2018.

[79] C. Paris, J. Bioucas-Dias, and L. Bruzzone, "A hierarchical approach to superresolution of multispectral images with different spatial resolutions," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul. 2017, pp. 2589–2592.

[80] C. Pohl and J. L. V. Genderen, "Review article multisensor image fusion in remote sensing: Concepts, methods and applications," *International Journal of Remote Sensing*, vol. 19, no. 5, pp. 823–854, 1998.

[81] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.

[82] R. Gaetano, D. Amitrano, G. Masi, G. Poggi, G. Ruello, L. Verdoliva, and G. Scarpa, "Exploration of multitemporal COSMO-skymed data via interactive tree-structured MRF segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 7, pp. 2763–2775, 2014.

[83] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, no. 7, p. 594, 2016, ISSN: 2072-4292. DOI: 10.3390/rs8070594. [Online]. Available: http://www.mdpi.com/2072-4292/8/7/594.

[84] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, May 2017, ISSN: 1545-598X. DOI: 10.1109/LGRS.2017.2668299.

[85] R. Gaetano, G. Moser, G. Poggi, G. Scarpa, and S. B. Serpico, "Region-based classification of multisensor optical-sar images," in *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, vol. 4, Jul. 2008, pp. IV - 81-IV –84.

[86] J. Reiche, C. M. Souza, D. H. Hoekman, J. Verbesselt, H. Persaud, and M. Herold, "Feature level fusion of multi-temporal alos palsar and landsat data for mapping and monitoring of tropical deforestation and forest degradation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 5, pp. 2159–2173, Oct. 2013, ISSN: 1939-1404. DOI: 10.1109/JSTARS.2013.2245101.

[87] A. Errico, C. V. Angelino, L. Cicala, G. Persechino, C. Ferrara, M. Lega, A. Vallario, C. Parente, G. Masi, R. Gaetano, G. Scarpa, D. Amitrano, G. Ruello, L. Verdoliva, and G. Poggi, "Detection of environmental hazards through the feature-based fusion of optical and sar data: A case study in southern italy," *International Journal of Remote Sensing*, vol. 36, no. 13, pp. 3345–3367, 2015.

[88] M. Das and S. K. Ghosh, "Deep-step: A deep learning approach for spatiotemporal prediction of remote sensing data," *IEEE Geosci. Remote Sensing Lett.*, vol. 13, no. 12, pp. 1984–1988, 2016. DOI: 10.1109/LGRS.2016.2619984. [Online]. Available: https://doi.org/10.1109/LGRS.2016.2619984.

[89] C. Sukawattanavijit, J. Chen, and H. Zhang, "Ga-svm algorithm for improving land-cover classification using sar and optical remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 3, pp. 284–288, Mar. 2017, ISSN: 1545-598X. DOI: 10.1109/LGRS.2016.2628406.

[90] W. Ma, Z. Wen, Y. Wu, L. Jiao, M. Gong, Y. Zheng, and L. Liu, "Remote sensing image registration with modified sift and enhanced

feature matching," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 1, pp. 3–7, Jan. 2017, ISSN: 1545-598X. DOI: 10.1109/LGRS.2016.2600858.

[91] N. Clerici, C. A. V. Calderón, and J. M. Posada, "Fusion of sentinel-1a and sentinel-2a data for land cover mapping: A case study in the lower magdalena region, colombia," *Journal of Maps*, vol. 13, no. 2, pp. 718–726, 2017.

[92] F. Jahan and M. Awrangjeb, "Pixel-Based Land Cover Classification by Fusing Hyperspectral and LIDAR Data," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 711–718, 2017.

[93] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Decision fusion for the classification of urban remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 10, pp. 2828–2838, Oct. 2006.

[94] C. Márquez, M. I. López, I. Ruisánchez, and M. P. Callao, "Ft-raman and nir spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud," *Talanta*, vol. 161, pp. 80–86, 2016.

[95] B. Waske and S. van der Linden, "Classifying multilevel imagery from sar and optical sensors by decision fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1457–1466, May 2008, ISSN: 0196-2892. DOI: 10.1109/TGRS.2008.916089.

[96] J. Reiche, S. de Bruin, D. Hoekman, J. Verbesselt, and M. Herold, "A bayesian approach to combine landsat and alos palsar time series for near real-time deforestation detection," *Remote Sensing*, vol. 7, no. 5, pp. 4973–4996, 2015.

[97] P. Du, S. Liu, J. Xia, and Y. Zhao, "Information fusion techniques for change detection from multi-temporal remote sensing images," *Information Fusion*, vol. 14, no. 1, pp. 19–27, 2013.

[98] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Cnn-based pansharpening of multi-resolution remote-sensing images," in *Joint Urban Remote Sensing Event 2017*, Dubai, Jun. 2017.

[99] R. Gaetano, G. Masi, G. Poggi, L. Verdoliva, and S. G., "Marker controlled watershed based segmentation of multi-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 1987–3004, Jun. 2015.

[100] A. Ding, Q. Zhang, X. Zhou, and B. Dai, "Automatic recognition of landslide based on CNN and texture change detection," in *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, Nov. 2016, pp. 444–448.

[101] M. Zanetti and L. Bruzzone, "A theoretical framework for change detection based on a compound multiclass statistical model of the difference image," *IEEE Transactions on Geoscience and Remote Sensing*, 2017, ISSN: 0196-2892. DOI: 10.1109/TGRS.2017.2759663.

[102] W. Liu, J. Yang, J. Zhao, and L. Yang, "A novel method of unsupervised change detection using multi-temporal polsar images," *Remote Sensing*, vol. 9, no. 11, p. 1135, 2017.

[103] Y. Han, F. Bovolo, and L. Bruzzone, "Segmentation-based fine registration of very high resolution multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2884–2897, May 2017.

[104] G. Chierchia, M. E. Gheche, G. Scarpa, and L. Verdoliva, "Multitemporal sar image despeckling based on block-matching and collaborative filtering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5467–5480, Oct. 2017.

[105] S. Maity, C. Patnaik, M. Chakraborty, and S. Panigrahy, "Analysis of temporal backscattering of cotton crops using a semiempirical model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 3, pp. 577–587, Mar. 2004.

[106] T. Manninen, P. Stenberg, M. Rautiainen, and P. Voipio, "Leaf area index estimation of boreal and subarctic forests using vv/hh envisat/asar data of various swaths," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 3899–3909, Jul. 2013.

[107] E. F. Borges, E. E. Sano, and E. Medrado, "Radiometric quality and performance of timesat for smoothing moderate resolution imaging spectroradiometer enhanced vegetation index time series from western bahia state, brazil," *Journal of Applied Remote Sensing*, vol. 8, no. 1, pp. 083 580–083 580, 2014.

[108] H. Zhang, H. Lin, and Y. Li, "Impacts of feature normalization on optical and sar data fusion for land use/land cover classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1061–1065, May 2015, ISSN: 1545-598X. DOI: 10.1109/LGRS.2014.2377722.

[109] Q. Man, P. Dong, and H. Guo, "Pixel-and feature-level fusion of hyperspectral and lidar data for urban land-use classification," *International Journal of Remote Sensing*, vol. 36, no. 6, pp. 1618–1644, 2015.

[110] M. Lu, B. Chen, X. Liao, T. Yue, H. Yue, S. Ren, X. Li, Z. Nie, and B. Xu, "Forest types classification based on multi-source data fusion," *Remote Sensing*, vol. 9, no. 11, p. 1153, 2017.

[111] S. K. Pal, T. J. Majumdar, and A. K. Bhattacharya, "ERS-2 SAR and IRS-1C LISS III data fusion: A PCA approach to improve remote sensing based geological interpretation," *ISPRS Journal of*

Photogrammetry and Remote Sensing, vol. 61, pp. 281–297, 2007. DOI: 10.1016/j.isprsjprs.2006.10.001.

[112] J. D. Bolten, V. Lakshmi, and E. G. Njoku, "Soil moisture retrieval using the passive/active l- and s-band radar/radiometer," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 12, pp. 2792–2801, Dec. 2003.

[113] N. N. Baghdadi, M. E. Hajj, M. Zribi, and I. Fayad, "Coupling sar c-band and optical data for soil moisture and leaf area index retrieval over irrigated grasslands," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 3, pp. 1229–1243, Mar. 2016, ISSN: 1939-1404. DOI: 10.1109/JSTARS.2015.2464698.

[114] E. Santi, S. Paloscia, S. Pettinato, D. Entekhabi, S. H. Alemohammad, and A. G. Konings, "Integration of passive and active microwave data from smap, amsr2 and sentinel-1 for soil moisture monitoring," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul. 2016, pp. 5252–5255. DOI: 10.1109/IGARSS.2016.7730368.

[115] P. Addabbo, M. Focareta, S. Marcuccio, C. Votto, and S. L. Ullo, "Land cover classification and monitoring through multisensor image and data combination," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul. 2016, pp. 902–905.

[116] J. Jelének, V. Kopačková, L. Koucká, and J. Mišurec, "Testing a modified pca-based sharpening approach for image fusion," *Remote Sensing*, vol. 8, no. 10, p. 794, 2016, ISSN: 2072-4292. DOI: 10.3390/rs8100794. [Online]. Available: http://www.mdpi.com/2072-4292/8/10/794.

[117] M. Bisquert, G. Bordogna, M. Boschetti, P. Poncelet, and M. Teisseire, "Soft fusion of heterogeneous image time series," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Springer, 2014, pp. 67–76.

[118] M. S. Moran, D. C. Hymer, J. Qi, and E. E. Sano, "Soil moisture evaluation using multi-temporal synthetic aperture radar (sar) in semiarid rangeland," *Agricultural and Forest Meteorology*, vol. 105, no. 1, pp. 69–80, 2000, ISSN: 0168-1923. DOI: http://dx.doi.org/10.1016/S0168-1923(00)00189-1. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168192300001891.

[119] Q. Wang, G. A. Blackburn, A. O. Onojeghuo, J. Dash, L. Zhou, Y. Zhang, and P. M. Atkinson, "Fusion of landsat 8 oli and sentinel-2 msi data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3885–3899, Jul. 2017.

[120] H. Xu, "Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery," *International journal of remote sensing*, vol. 27, no. 14, pp. 3025–3033, 2006.

[121] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 639–643, 2017.

[122] J. K. L. J. Kim and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016, pp. 1646–1654.

[123] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018, ISSN: 0196-2892. DOI: 10.1109/TGRS.2018.2817393.

[124] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. 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*, vol. 11, pp. 978–989, Mar. 2018. DOI: 10.1109/JSTARS.2018.2794888. arXiv: 1712.09809 [cs.CV].

[125] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolution: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sensing*, pp. 691–699, 1997.

[126] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "Pannet: A deep network architecture for pan-sharpening," in *ICCV*, Oct. 2017. DOI: 10.1109/ICCV.2017.193.

[127] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[128] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, Jun. 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965.

[129] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *Proceedings of European Conference on Computer Vision*, 2014.

[130] E. Maltezos, N. Doulamis, A. Doulamis, and C. Ioannidis, "Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds," *Journal of Applied Remote Sensing*, vol. 11, no. 4, p. 042 620, 2017.

[131] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, "Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5585–5599, Oct. 2017, ISSN: 0196-2892. DOI: 10.1109/TGRS.2017.2710079.

[132] K. Fotiadou, G. Tsagkatakis, and P. Tsakalides, "Deep convolutional neural networks for the classification of snapshot mosaic hyperspectral imagery," *Electronic Imaging*, vol. 2017, no. 17, pp. 185–190, 2017.

[133] L. Bottou, "Stochastic gradient learning in neural networks," *Proceedings of Neuro-Nımes*, vol. 91, no. 8, p. 12, 1991.

[134] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[135] D. C. Cireşan, L. M. Gambardella, A. Giusti, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *In NIPS*, 2012, pp. 2852–2860.

[136] M. Gargiulo, A. Mazza, R. Gaetano, G. Ruello, and G. Scarpa, "A CNN-based fusion method for super-resolution of Sentinel-2 data," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2018, pp. 4713–4716.

[137] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456.

[138] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[139] J. Johnson, A. Alahi, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016. arXiv: 1603.08155. [Online]. Available: http://arxiv.org/abs/1603.08155.

[140] Y. Jiang, X. Ding, D. Zeng, Y. Huang, and J. Paisley, "Pan-sharpening with a hyper-laplacian penalty," in *ICCV*, Dec. 2015.

[141] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[142] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[143] M. Gargiulo, D. A. Dell'Aglio, A. Iodice, D. Riccio, and G. Ruello, "Semantic segmentation using deep learning: A case of study in albufera park, valencia," in *2019 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*, IEEE, 2019, pp. 134–138.

[144] G. Scarpa, M. Gargiulo, A. Mazza, and R. Gaetano, "A cnn-based fusion method for feature extraction from sentinel data," *Remote Sensing*, vol. 10, no. 2, p. 236, 2018.

[145] M. Gargiulo, D. A. Dell'Aglio, A. Iodice, D. Riccio, and G. Ruello, "Integration of sentinel-1 and sentinel-2 data for land cover mapping using w-net," *Sensors*, vol. 20, no. 10, p. 2969, 2020.

[146] L. R. Beck, B. M. Lobitz, and B. L. Wood, "Remote sensing and human health: New sensors and new opportunities.," *Emerging infectious diseases*, vol. 6, no. 3, p. 217, 2000.

[147] H. Gao, C. Birkett, and D. P. Lettenmaier, "Global monitoring of large reservoir storage from satellite remote sensing," *Water Resources Research*, vol. 48, no. 9, 2012.

[148] R. Brakenridge and E. Anderson, "Modis-based flood detection, mapping and measurement: The potential for operational hydrological applications," in *Transboundary floods: reducing risks through flood management*, Springer, 2006, pp. 1–12.

[149] D.-H. Yoon, W.-H. Nam, H.-J. Lee, E.-M. Hong, T. Kim, D.-E. Kim, A.-K. Shin, and M. D. Svoboda, "Application of evaporative stress index (esi) for satellite-based agricultural drought monitoring in south korea," *Journal of The Korean Society of Agricultural Engineers*, vol. 60, no. 6, pp. 121–131, 2018.

[150] J.-C. Kim and H.-S. Jung, "Application of landsat tm/etm+ images to snow variations detection by volcanic activities at southern volcanic zone, chile," *Korean Journal of Remote Sensing*, vol. 33, no. 3, pp. 287–299, 2017.

[151] T. Murakami, S. Ogawa, N. Ishitsuka, K. Kumagai, and G. Saito, "Crop discrimination with multitemporal spot/hrv data in the saga plains, japan," *International journal of remote sensing*, vol. 22, no. 7, pp. 1335–1348, 2001.

[152] T. N. Carlson and D. A. Ripley, "On the relation between ndvi, fractional vegetation cover, and leaf area index," *Remote sensing of Environment*, vol. 62, no. 3, pp. 241–252, 1997.

[153] F. Maselli, M. Chiesi, and M. Pieri, "A new method to enhance the spatial features of multitemporal ndvi image series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4967–4979, 2019.

[154] C. Manzo, A. Mei, G. Fontinovo, A. Allegrini, and C. Bassani, "Integrated remote sensing for multi-temporal analysis of anthropic activities in the south-east of mt. vesuvius national park," *Journal of African Earth Sciences*, vol. 122, pp. 63–78, 2016.

[155] S. K. McFeeters, "The use of the normalized difference water index (ndwi) in the delineation of open water features," *International journal of remote sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.

[156] K. Rokni, A. Ahmad, K. Solaimani, and S. Hazini, "A new approach for detection of surface water changes based on principal component analysis of multitemporal normalized difference water index," *Journal of Coastal Research*, vol. 32, no. 2, pp. 443–451, 2016.

[157] S. t. Wu and S. A. Sader, "Multipolarization SAR data for surface feature delineation and forest vegetation characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-25, no. 1, pp. 67–76, 1987.

[158] A. R. H. E. E. Sano L. G. Ferreira, "Synthetic aperture radar (l band) and optical vegetation indices for discriminating the Brazilian Savanna physiognomies: A comparative analysis," *Earth Interactions*, vol. 9, no. 15, pp. 1–15, 2005.

[159] T. Nagler, H. Rott, M. Hetzenecker, J. Wuite, and P. Potin, "The sentinel-1 mission: New opportunities for ice sheet observations, remote sens., 7, 9371–9389," 2015.

[160] H. Bazzi, N. Baghdadi, M. El Hajj, M. Zribi, D. H. T. Minh, E. Ndikumana, D. Courault, and H. Belhouchette, "Mapping paddy rice using sentinel-1 sar time series in camargue, france," *Remote Sensing*, vol. 11, no. 7, p. 887, 2019.

[161] D. Amitrano, G. Di Martino, A. Iodice, D. Riccio, and G. Ruello, "Unsupervised rapid flood mapping using sentinel-1 grd sar

images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3290–3299, 2018.

[162] S. Abdikan, F. B. Sanli, M. Ustuner, and F. Calò, "Land cover mapping using sentinel-1 sar data," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 41, p. 757, 2016.

[163] J. Haas and Y. Ban, "Sentinel-1a sar and sentinel-2a msi data fusion for urban ecosystem service mapping," *Remote Sensing Applications: Society and Environment*, vol. 8, no. Supplement C, pp. 41–53, 2017, ISSN: 2352-9385. DOI: https://doi.org/10.1016/j.rsase.2017.07.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352938517300125.

[164] F. Lang, J. Yang, S. Yan, and F. Qin, "Superpixel segmentation of polarimetric synthetic aperture radar (sar) images based on generalized mean shift," *Remote Sensing*, vol. 10, no. 10, p. 1592, 2018.

[165] D. Stutz, A. Hermans, and B. Leibe, "Superpixels: An evaluation of the state-of-the-art," *Computer Vision and Image Understanding*, vol. 166, pp. 1–27, 2018.

[166] M. Ciecholewski, "River channel segmentation in polarimetric sar images: Watershed transform combined with average contrast maximisation," *Expert Systems with Applications*, vol. 82, pp. 196–215, 2017.

[167] J. Cousty, G. Bertrand, L. Najman, and M. Couprie, "Watershed cuts: Thinnings, shortest path forests, and topological watersheds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 925–939, 2009.

[168] A. M. Braga, R. C. Marques, F. A. Rodrigues, and F. N. Medeiros, "A median regularized level set for hierarchical segmentation of sar

images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 7, pp. 1171–1175, 2017.

[169] R. Jin, J. Yin, W. Zhou, and J. Yang, "Level set segmentation algorithm for high-resolution polarimetric sar images based on a heterogeneous clutter model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 10, pp. 4565–4579, 2017.

[170] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric sar image classification using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1935–1939, 2016.

[171] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017.

[172] A. Stoian, V. Poulain, J. Inglada, V. Poughon, and D. Derksen, "Land cover maps production with high resolution satellite image time series and convolutional neural networks: Adaptations and limits for operational systems," *Remote Sensing*, vol. 11, no. 17, p. 1986, 2019.

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

[174] X. Zhao, Y. Yuan, M. Song, Y. Ding, F. Lin, D. Liang, and D. Zhang, "Use of unmanned aerial vehicle imagery and deep learning unet to extract rice lodging," *Sensors*, vol. 19, no. 18, p. 3859, 2019.

[175] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *International symposium on visual computing*, Springer, 2016, pp. 234–244.

[176] J. Deng, R. Socher, L. Fei-Fei, W. Dong, K. Li, and L.-J. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, vol. 00, Jun. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5206848/.

[177] P. D'Odorico, A. Gonsamo, A. Damm, and M. E. Schaepman, "Experimental evaluation of sentinel-2 spectral response functions for ndvi time-series continuity," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 3, pp. 1336–1348, 2013.

[178] L. Gao, W. Song, J. Dai, and Y. Chen, "Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network," *Remote Sensing*, vol. 11, no. 5, p. 552, 2019.

[179] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 4062–4076, 2019.

[180] X. Xia and B. Kulis, "W-net: A deep model for fully unsupervised image segmentation," *arXiv preprint arXiv:1711.08506*, 2017.

[181] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," *arXiv preprint arXiv:1605.07648*, 2016.

[182] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thir-*

teenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.

[183] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas," in *GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, May 2003, pp. 90–94.

[184] A. Verhegghen, H. Eva, G. Ceccherini, F. Achard, V. Gond, S. Gourlet-Fleury, and P. Cerutti, "The potential of Sentinel satellites for burnt area mapping and monitoring in the Congo Basin forests," *Remote Sensing*, vol. 8, no. 12, p. 986, 2016.

[185] J. Pereira, E. Chuvieco, A. Beudoin, and N. Desbois, "Remote sensing of burned areas: A review. A review of remote sensing methods for the study of large wildland fires," *Departamento de Geografía, Universidad de Alcalá*, pp. 127–184, Jan. 1997.

[186] Y. Kant and K. V. S. Badarinath, "Studies on land surface temperature over heterogeneous areas using AVHRR data," *International Journal of Remote Sensing*, vol. 21, no. 8, pp. 1749–1756, 2000. DOI: 10.1080/014311600210029. eprint: https://doi.org/10.1080/014311600210029. [Online]. Available: https://doi.org/10.1080/014311600210029.

[187] L. Mascolo, M. Sarti, F. Nunziata, and M. Migliaccio, "Vesuvius national park monitoring by COSMO-SkyMed PingPong data analysis," in *ESA Special Publication*, vol. 713, 2013.

[188] G. Bovio, M. Marchetti, L. Tonarelli, M. Salis, G. Vacchiano, R. Lovreglio, M. Elia, P. Fiorucci, and D. Ascoli, "Gli incendi boschivi stanno cambiando: Cambiamo le strategie per governarli," *Foresta - Rivista di Selvicoltura ed Ecologia Forestale*, no. 4, pp. 202–205, 2017. DOI: 10.3832/efor2537-014. eprint: http://foresta.

sisef.org/pdf/?id=efor2537-014. [Online]. Available: http://foresta.sisef.org/contents/?id=efor2537-014.

[189] H. Huang, D. Roy, L. Boschetti, H. Zhang, L. Yan, S. Kumar, J. Gomez-Dans, and J. Li, "Separability Analysis of Sentinel-2A Multi-Spectral Instrument (MSI) Data for Burned Area Discrimination," *Remote Sensing*, vol. 8, Nov. 2016. DOI: 10.3390/rs8100873.

[190] W. Schroeder, P. Oliva, L. Giglio, B. Quayle, E. Lorenz, and F. Morelli, "Active fire detection using Landsat-8/OLI data," *Remote Sensing of Environment*, vol. 185, Sep. 2015. DOI: 10.1016/j.rse.2015.08.032.

[191] A. Barducci, D. Guzzi, P. Marcoionni, and I. Pippi, "Infrared detection of active fires and burnt areas: Theory and observations," *Infrared physics & technology*, vol. 43, no. 3-5, pp. 119–125, 2002.

[192] L. Giglio, I. Csiszar, Á. Restás, J. T. Morisette, W. Schroeder, D. Morton, and C. O. Justice, "Active fire detection and characterization with the advanced spaceborne thermal emission and reflection radiometer (aster)," *Remote sensing of environment*, vol. 112, no. 6, pp. 3055–3063, 2008.

[193] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[194] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[195] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2565–2586, 2015.

[196] C. Yang and J. H. Everitt, "Using spectral distance, spectral angle and plant abundance derived from hyperspectral imagery to characterize crop yield variation," *Precision agriculture*, vol. 13, no. 1, pp. 62–75, 2012.

[197] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm," 1992.

[198] P. Jagalingam and A. V. Hegde, "A review of quality metrics for fused image," *Aquatic Procedia*, vol. 4, pp. 133–142, 2015.

[199] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3012–3021, 2007.

[200] Z. Wang and A. Bovik, "A universal image quality index," *Signal Processing Letters, IEEE*, vol. 9, no. 3, pp. 81–84, Mar. 2002, ISSN: 1070-9908. DOI: 10.1109/97.995823.

[201] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogrammetric Engineering & Remote Sensing*, vol. 74, no. 2, pp. 193–200, 2008.

[202] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques," *Remote Sensing of Environment*, vol. 22, no. 3, pp. 343–365, 1987.

[203] M. M. Khan, J. Chanussot, L. Condat, and A. Montanvert, "Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 1, pp. 98–102, 2008.

[204] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 1, pp. 295–309, 2011.

[205] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and Pan imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 5, pp. 591–596, 2006.

[206] J. Inglada, M. Arias, B. Tardy, O. Hagolle, S. Valero, D. Morin, G. Dedieu, G. Sepulcre, S. Bontemps, P. Defourny, and B. Koetz, "Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery," *Remote Sensing*, vol. 7, no. 9, pp. 12 356–12 379, 2015. [Online]. Available: `http://www.mdpi.com/2072-4292/7/9/12356`.

[207] ESA, *ESA Sentinel Application Platform (SNAP) software*, `http://step.esa.int/main/toolboxes/snap`, (accessed on 13 December 2017).

[208] *THEIA home page*, `http://www.theia-land.fr`, (accessed on 13 December 2017).

[209] O. Hagolle, M. Huc, D. Villa Pascual, and G. Dedieu, "A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of formosat-2, landsat, venμs and sentinel-2 images," *Remote Sensing*, vol. 7, no. 3, pp. 2668–2691, 2015. [Online]. Available: `http://www.mdpi.com/2072-4292/7/3/2668`.

[210] *Orfeo Toolbox: Temporal gap-filling*, `http://tully.ups-tlse.fr/jordi/temporalgapfilling`, (accessed on 13 December 2017).

[211] H. Zhang and B. Huang, "Support vector regression-based downscaling for intercalibration of multiresolution satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 3, pp. 1114–1123, Mar. 2013.

[212] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[213] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[214] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, 2017, pp. 1–4.

[215] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399–6408.

[216] A. Mazza, M. Gargiulo, G. Scarpa, and R. Gaetano, "Estimating the ndvi from sar by convolutional neural networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2018, pp. 1954–1957.

[217] J. Li, C. Wu, R. Song, W. Xie, C. Ge, B. Li, and Y. Li, "Hybrid 2-d-3-d deep residual attentional network with structure tensor constraints for spectral super-resolution of rgb images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[218] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Joint camera spectral response selection and hyperspectral image recovery,"

IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[219] W. Chen, X. Zheng, and X. Lu, "Hyperspectral image super-resolution with self-supervised spectral-spatial residual network," *Remote Sensing*, vol. 13, no. 7, p. 1260, 2021.

[220] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International journal of Remote sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[221] S. Talukdar, P. Singha, S. Mahato, S. Pal, Y.-A. Liou, A. Rahman, *et al.*, "Land-use land-cover classification by machine learning classifiers for satellite observations—a review," *Remote Sensing*, vol. 12, no. 7, p. 1135, 2020.