TESI DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI NAPOLI "FEDERICO II"

DIPARTIMENTO DI INGEGNERIA ELETTRONICA
E DELLE TECNOLOGIE DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN
INFORMATION TECHNOLOGIES AND
ELECTRICAL ENGINEERING

DEEP LEARNING BASED DATA-FUSION METHODS FOR REMOTE SENSING APPLICATIONS

ANTONIO MAZZA

Il Coordinatore del Corso di Dottorato	Il Tutore
Ch.mo Prof. Daniele RICCIO	Ch.mo Prof. Giuseppe SCARPA

A. A. 2020–2021

Contents

List of Figures

Introduction

In the last years, an increasing number of remote sensing sensors have been launched to orbit around the Earth, with a continuously growing production of massive data.

The importance of remote sensing images stems from the global coverage offer, often coupled with a relatively short revisit time, that provides multitemporal information to constantly monitor the Earth surface. Thanks to the increasing volume of freely-available data, a lot of remote sensing related applications have been made practically realizable at a large scale. Notable examples are food security, vegetation or ice monitoring, land-cover use and classification, and so on.

Among remote sensing products, optical and Synthetic Aperture Radar (SAR) images provide complementary information that is useful in many of these monitoring application. Besides, multitemporal information helps to detect expected or unexpected changes that can contribute to further characterize the observed phenomena of interest. Despite modern optical sensors provide rich spectral information about Earth's surface, at very high resolution, they are weather-sensitive, hence useless under cloudy conditions. SAR images are always available also in presence of clouds and are almost weather-insensitive, as well as daynight available. On the downside, SAR images, do not provide a rich spectral information in comparison to multispectral optical images and, also, as result of an active image formation process, the SAR geometry is rather complex to model, giving raise to phenomenon such as the speckle "noise" that make difficult the information extraction.

For the above reasons it is worth and challenging to fuse data provided by different sources and/or acquired at different times, in order to leverage on their diversity and complementarity to retrieve the target information. But this is not a simple task, especially in the case of the fusion of optical and SAR data, because of their deeply different imaging process.

Moreover, many classification tasks, for example those related to the veg-

etation, are often addressed by means of an alternative use of optical or SAR features. It is therefore worth seeking solutions that make a joint use of both sources. Another interesting "fusion" opportunity regards multi-resolution images. In fact, due to technological constraints, Earth observation optical satellites are often equipped with multiple resolutions imaging systems that provide different sets of spectral bands associated to different spatial resolutions. These systems are conceived to trade-off between spectral and spatial resolutions, offering worser spatial resolutions for the narrower spectral bands that are located in some given ranges of the spectrum that need to be densely sampled. Then, the multi-resolution data fusion aims to mix these different components to provide a spatial-spectral full-resolution datacube.

Nowadays, thanks to the availability of increased computational resources, coupled with the continuously growing bulk of data (sometimes unlimited, as in the case of the Sentinel missions) freely available to the community, and also thanks to the breakthrough advances in the machine learning domain [1], a new very promising data-driven paradigm has emerged in the very last years: Deep Learning (DL).

Indeed, in the last years, Deep Learning has been introduced in a lot of different tasks providing State-of-the-Art performances, especially in the Computer Vision domain, e.g. classification, segmentation, object detection and so forth, where the so-called Convolution Neural Networks (CNNs) became very popular.

Actually, a well-designed Deep Neural Network could learn complex tasks and perform complex non-linear functions that better describe the problem with an appropriate training phase provided that a "sufficiently rich" labeled dataset is made available for training.

In this thesis different typical remote sensing data-fusion problems are faced by means of suitably designed CNNs. These are the NDVI restoration, super-resolution, forest mapping and cloud detection, all tasks approached by mixing different input channels. The thesis outline is the following. Chapter 1 draws the background of the thesis, providing basic concepts about deep learning and convolutional neural networks, and introducing data-fusion for remote sensing.

In Chapter 2 the proposed framework for the restoration of the Normalized Difference Vegetation Index (NDVI) is presented, which applies when the spectral bands needed for its computation are not available, *e.g.*, because of cloudy conditions. In particular, different CNN-based fusion solutions of Sentinel-2 optical and/or Sentinel-1 SAR multitemporal series are presented.

The presentation of a framework that employs only SAR data to estimate the missing optical feature will conclude the Chapter.

In Chapter 3 a set of some state-of-the-art CNN architectures are modified and adapted to the task of forest classification with TanDEM-X SAR data and additional features such as interferometric coherence, incidence angle and the volume correlation coherence.

Next, in Chapter 4, a super-resolution CNN-based method is proposed. In particular, Sentinel-2 optical bands with 20-m spatial resolution are super-resolved up to 10-m by CNN-based fusion with the complementary 10-m subset of Sentinel-2 bands.

In the first part of Chapter 5 some preliminary results about cloud detection on Sentinel-2 bands are shown. Two alternative solutions are compared, the former based on a rough labeling of Sentinel-2 data, the latter relying on an accurate labeling available for a different sensor (Landsat-8). Follows a study on the impact of the training set choice for the despeckling of SAR image task with CNN. Indeed, for the this particular task, there is no 'clean' reference and the design of the training set has a great impact on the performance. Therefore, in order to have a 'clean' reference temporal multilooking and noise-simulated approaches are compared.

Finally, the thesis ends with concluding remarks and future perspectives.

Chapter 1

Background

In this chapter some background concepts are introduced. In particular, a brief introduction on the remote sensing sensors and a review of data-fusion method are given in Section 1.1. In Section 1.2 follows a short introduction to Deep Learning with focus on Convolutional Neural Networks.

1.1 Data-Fusion in Remote Sensing

Nowadays, Remote sensing for Earth Observation (EO) involves a plethora of different sensing systems that can be mounted on different flying systems, *e.g.*, satellites, aircrafts, drones. From the processing point of view, the most important distinction is between passive and active sensing systems. These two categories are often roughly referred to as optical and Synthetic Aperture Radar (SAR), respectively. Optical sensors can then be further divided in multispectral (MS) or hyperspectral, depending on the number of provided spectral bands, which is in the order of tens for MS images and of hundreds for hyperspectral ones. Additionally, optical images are frequently provided at multiple resolutions to optimize costs and benefits. On the SAR side, it is also possible to distinguish among different modalities. The Reader is referred to domain expert readings such as [2, 3] for further details. In addition to these main features, that are many parameters that further characterize any remote sensing system, for example spatial, spectral or radiometric resolutions, revisit time, sensor altitude and so on [4].

The main advantages of passive sensing are the possibility to provide very rich spectral information and a simple geometry relation with the ground geometry, that is a standard projective transformation. On the other hand, as it

leverages on sunlight reflection, it can only be used in daytime and under good weather conditions: pollution, water vapour, and clouds can severely affect the received signal. Besides, an active sensing system, such as SAR, can be considered almost insensitive to weather conditions and does not need any Sun radiation. Unfortunately, on the downside, SAR geometry is very complex, generating unwanted processes such as speckle "noise" and double bounds, and the spectral information cannot be exploited as extensively done with passive sensing systems. Among the different remote sensing systems, those boarded on satellites are of fundamental importance as thanks to them it can be ensured a global Earth monitoring with controlled revisit time, which can drop to just a few days in some cases. Thanks to these characteristics many phenomena, some of which require multitemporal analyses, can be observed through satellite systems. Notable examples are vegetation and glaciers monitoring [5], land-cover [6], canopy [7] and forest [8] monitoring, as well as oil-spill [9], ship [10] and change [11] detection , despeckling [12], pan-sharpening [13] and so forth. Due to the temporal, spectral or resolution diversity, as well as for the complementarity between optical and SAR systems highlighted above, many fusion techniques have been developed in past years both for information extraction [11, 14, 15] and image quality enhancement [13, 12].

According to the taxonomy given in [16] data fusion methods, *i.e.*, *processing dealing with data and information from multiple sources to achieve improved information for decision making*, can be grouped in three main categories:

- *pixel*-level: the pixel values of the sources to be fused are jointly processed [17, 15, 18, 19];
- *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 [20, 21, 22, 23, 24, 25, 26, 27];
- *decision*-level: the high-level information extracted independently from each source is combined to provide the final outcome, for example using fuzzy logic [28, 29], decision trees [30], Bayesian inference [31], Dempster-Shafer theory [32], and so forth.

In the context of remote sensing, with reference to the sources to be fused, fusion methods can be roughly gathered for the most part in the following categories:

- multi-*resolution*: concerns a single sensor with multiple resolution bands. One of the most frequent applications is pansharpening [17, 33, 34], although many other tasks can be solved under a multi-resolution paradigm, such as segmentation [35] or feature extraction [36], 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 [23], change detection [37, 11, 38], co-registration [39], and general-purpose tasks, like segmentation [15], despeckling [12], feature extraction [40, 41, 42], 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 represents also the most challenging case because of the several sources of mismatch (temporal, geometrical, spectral, radiometrical) among involved data. Like for other categories, a number of typical remote sensing problems can fit this paradigm, such as classification [20, 43, 44, 26, 8], coregistration [25], change detection [45], feature estimation [46, 47, 48, 6].

- *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 [49, 19]. Likewise, low-resolution temporally dense series can be fused with high-resolution but temporally sparse ones to simulate a temporal-spatial full-resolution sequence [50]. The monitoring of forests [31], soil moisture [51], environmental hazards [22], and other processes, can be also 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 [32, 21, 52].

In many cases, the data-fusion between too heterogeneous data could be too challenging, using traditional model-based approaches. For this reason, and in light of the recent advances in deep learning, it becomes really worth to explore the possibility to address such complex data-fusion tasks through learning, leveraging on the huge mass of data provided from remote sensing systems flying around the Earth.

1.2 Deep Learning

Machine Learning approaches, and in particular Deep Learning approaches refer to an automatic procedure that employs an Artificial Neural Network (ANN) that is trained to learn from a set of given examples. Among Deep learning approaches, Convolutional Neural Networks, have been massively used in computer vision and image processing in the last few years, since the publication of the breakthrough work of Krizhevsky *et al.* on image classification in 2012 [53]. Thanks to the CNNs capability to learn very complex non-linear relationships from huge labeled datasets with the help of commercial GPUs, unprecedented results have been obtained for many typical tasks such as super-resolution [54, 55], segmentation [56], denoising [57], object detection [58, 59], classification [60, 61, 27], and may others.

Recently, deep learning has started to significantly impact remote sensing applications as well, as testified by the recent survey of Zhu *et al.* [62]. Established techniques in remote sensing concern e.g. pansharpening [18, 63], vehicle detection [64] with optical images, crop classification [65, 66], anomaly detection with hyperspectral data [67], despeckling [68, 69], classification [70], or target recognition using SAR data [71, 10].

1.2.1 Convolutional neural networks

Convolutional Neural Network (CNN) is a general term that indicates a machine learning model, built up by interconnecting many different learnable or non-learnable processing units (layers), the most of which being convolutional ones. Neuronal weights sharing and locality, which make sense when working with images and videos, are the key characterizing elements that distinguish a convolutional layer from a fully connected (FC) layer that performs, instead, a linear combination of the inputs with learnable weights as well as a bias, so the number of weights grows with the amount of data that will be processed. Indeed, these distinctive features of CNNs allow a drastic reduction of the number of parameters to learn, and this is likely one of the reasons why the deep learning revolution has moved the first steps in the computer vision domain. Within the classification frame, the convolutional layers are usually employed in the early processing steps, in order to retain spatial layout and features localization. On the other hand, FC layers are normally applied after several processing units leading to abstract spatially unstructured features. Besides learnable layers, pooling and activation layers are non-learnable essential elements to build-up a deep learning classification model. Poolings that

normally interleave convolutional layer blocks aim to progressively "forget" the spatial structure reducing the spatial size, hence summarizing the image content with abstract features that could be obtained reducing the input spatial resolution reducing the search window to a pixel with the maximum or the average value within the window (MaxPooling or AveragePooling respectively). The input features will be reducted by the size of the search window. Roughly speaking, pooling helps to move from "where" to "what". On the other hand, non-linear point-wise activation layers, usually coupled with convolutional or FC layers, allow the overall network to mimic very complex non-linear functions, therefore expanding the network capacity.

In the last few years, CNNs have been successfully applied to many classical image processing problems, such as denoising [57], super-resolution [54], pansharpening [18, 34], segmentation [56], object detection [58, 59], change detection [37], classification [53, 60, 61, 27]. 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 correct training, CNNs require the availability of a large amount of data with ground-truth (examples). However, using large datasets has a cost in terms of complexity, and may lead to unreasonably long training times.

As already mentioned, a CNN is usually a chain[1] of different layers, like convolution, nonlinearities, pooling, deconvolution. For image processing tasks in which the desired output is an image at the same resolution of the input, the easiest solution employs only convolutional layers interleaved with nonlinear activations.

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,

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

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 pointwise activation function $g_l(\cdot)$, then, the overall layer output is given by

$$\mathbf{y}^{(l)} = g_l(\mathbf{z}^{(l)}) = g_l(\mathbf{w}^{(l)} * \mathbf{x}^{(l)} + \mathbf{b}^{(l)}) \triangleq f_l(\mathbf{x}^{(l)}, \Phi_l). \tag{1.1}$$

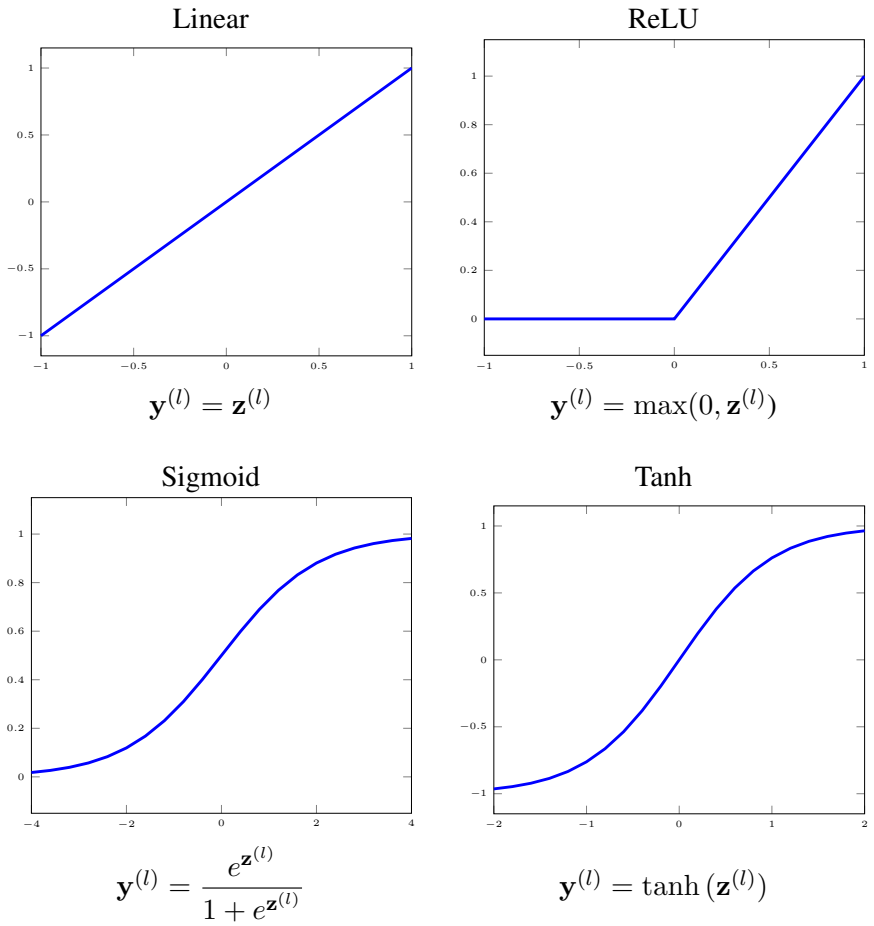


Figure 1.1: Some of the most used activation functions. The softmax cannot be plotted, since is not a function of a single fold from the previous layer

The activation function $g_l(\cdot)$ is used to add some non-linearities between the layers, since the convolution is a linear operator. Some examples of activa-

tion function are depicted in Figure 1.1. The choice of the activation function depends on the position of the layer. Indeed, Rectified Linear Unit [53] (ReLU) is usually employed after each hidden convolutional layer; while sigmoid and softmax[2] are usually employed before the output in classification tasks since the input is mapped a discrete probability distribution.

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

$$f(\mathbf{x}, \Phi) = f_L(f_{L-1}(\dots f_1(\mathbf{x}, \Phi_1), \dots, \Phi_{L-1}), \Phi_L), \tag{1.2}$$

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).

1.2.2 Learning

Once the architecture has been chosen, its parameters are learned by means of some optimization strategy, specifying the cost to be minimized over a properly selected training dataset. In order to learn the network parameters, a sufficiently large training set, say $\mathbf{T}$, of input-output examples $\mathbf{t}$ is needed:

$$\mathbf{T} \triangleq \{\mathbf{t}_1, \dots, \mathbf{t}_Q\}, \quad \mathbf{t} \triangleq (\mathbf{x}, \mathbf{y}^{\text{ref}})$$

Formally, the objective of the training phase is to find

$$\Phi = \arg\min_{\Phi} J(\mathbf{T}, \Phi) \triangleq \arg\min_{\Phi} \frac{1}{Q} \sum_{\mathbf{t} \in \mathbf{T}} L(\mathbf{t}, \Phi)$$

where $L(\mathbf{t}, \Phi)$ is a suitable loss function. Several losses can be found in the literature and the most commonly used are summarized in Table 1.1. The choice depends on the domain of the output, and affects the convergence properties of the networks [72].

For the minimization process an optimizer has to be chosen. For the sake of semplicity, in the following it is assumed the use of the Stochastic Gradient Descent (SGD) with momentum [73], but the following assumptions are valid also for other optimizers.

[2]Defined as: $\mathbf{y}_i^{(l)} = \dfrac{e^{\mathbf{z}_i^{(l)}}}{\sum_{j=1}^{J} e_j^{\mathbf{z}^{(l)}}}$, where $i = 1, \dots, J$ is the i-th output feature and J is the total number of output features. It maps the output values in order to have that the sum of all values of the feature maps in each pixel is 1.

L_n-norm	$L(\mathbf{t}, \Phi) = \|\|f(\mathbf{x}, \Phi) - \mathbf{y}^{\text{ref}}\|\|_n$
Cross-entropy	$L(\mathbf{t}, \Phi) = -\sum_i \mathbf{y}_i^{\text{ref}} \log(f(\mathbf{x}, \Phi)_i)$
Jaccard	$L(\mathbf{t}, \Phi) = \dfrac{f(\mathbf{x}, \Phi) \cdot \mathbf{y}^{\text{ref}}}{f(\mathbf{x}, \Phi) + \mathbf{y}^{\text{ref}} + f(\mathbf{x}, \Phi) \cdot \mathbf{y}^{\text{ref}}}$

Table 1.1: Some loss functions commonly used in the optimization process.

The training set is partitioned in batches of samples, $\mathbf{T} = \{\mathbf{B}_1, \ldots, \mathbf{B}_P\}$. At each iteration, a new batch is used to estimate the gradient and update parameters as

$$\nu^{(n+1)} \leftarrow \mu\nu^{(n)} + \alpha\nabla_\Phi J\left(\mathbf{B}_{j_n}, \Phi^{(n)}\right)$$
$$\Phi^{(n+1)} \leftarrow \Phi^{(n)} - \nu^{(n+1)}$$

A whole scan of the training set is called an *epoch*, and training a deep network may require from dozens of epochs, for simpler problems like handwritten character recognition [74], to thousands of epochs for complex classification tasks [53]. 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. During SGD optimization, the learning rate is fixed, but sometimes it could be useful to have a variable learning rate depending on the state of the training. Indeed, in the first phase of the training process, an high learning rate could be useful to rapidly move toward the minimum; while a smaller learning rate is useful when the solution is near by the optimum value. Therefore, it is useful to have the possibility to change the learning rate adaptively. To overcome this issue, a set of optimazer with an adaptive learning rate have been proposed. First of all, in the Adaptive Gradient Descent Algorithm (Adagrad) proposed in [75], the learning rate changes according with the cumulative gradient updated square values. Everytime the gradient is cumulated, the learning rate decreases accordingly and viceversa, but the drawback is that the learning rate could decrease faster and it may approach to zero. An upgraded version of Adagrad is RMSProp in which the cumulative sum of the square updates is replaced with a cumulative exponential sum. Finally, ADAM optimizer [76] combines the SGD method with momentum, to save the history of the gradients, with RMSProp, for learning rate adaptivity.

Chapter 2

NDVI regression

2.1 Introduction

Vegetation monitoring is critical for analyzing the characteristics of climate, soil, geology, and many other processes of interest. For this reason, many vegetation indexes have been defined, among which the NDVI is the most widely and frequently used [77]. Such an index, as well as many others, including those for water, bare soil and so on, is defined as combination of multispectral bands. Unfortunately, optical images are useless in cloudy conditions preventing a regular acquisition of NDVI time-series. This is a particularly critical problem in consideration of the stronger changes that affect vegetation during rainy seasons calling for finer-grain monitoring.

To address this shortcoming, several data fusion approaches have been proposed to fill the gaps caused by cloud occurrence in time-series of optical images. In particular, the reconstruction can be carried out both on spectral bands [78, 79] and derived features such as the NDVI [77, 80], by resorting to a complementary imaging system and/or by temporal interpolation. In [77, 78, 79] the moderate resolution – 500m – imaging spectroradiometer (MODIS) is used as complementary source to be combined with a 30m resolution Landsat optical series relying on the high revisit frequency (1 day) of MODIS. In [80] a pure temporal interpolation was considered for SPOT-4 and Landsat-8 images.

Recently, the launch of coupled optical/SAR Sentinel satellites, in the context of the Copernicus program, opens unprecedented opportunities for end users, both industrial and institutional, and poses new challenges to the remote sensing research community. The policy of free distribution of data allows large scale access to a very rich source of information. Besides this, the tech-

nical features of the Sentinel constellation make it a precious tool for a wide array of remote sensing applications. With revisit time ranging from two days to about a week, depending on the geographic location, spatial resolution from 10 to 60 meters, and wide coverage of the spectrum, from visible to short-wave infrared ($\sim 440 - 2200$ nm), Sentinel data may impact decisively on a number of 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 the coupled SAR and optical sensors, a key element for boosting the monitoring capability of the constellation. In fact, the information conveyed by the Sentinel-2 (S2) multi-resolution optical sensor depends on the spectral reflectivity of the target illuminated by sunlight, while 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 an accuracy that could not be achieved using only one of them. Of course, with this potential, comes the scientific challenge of how to exploit these complementary piece information in the most effective way.

In this chapter, the focus is on the estimation of the NDVI in critical weather conditions, fusing the information provided by temporal sequences of S1 and/or S2 images. In fact, the typical processing pipelines of many land monitoring applications rely, among other features, on the NDVI for a single date or a whole temporal series. Since, as already mentioned, the NDVI, as well as other spectral features, is unavailable under cloudy weather conditions, the commonly adopted solution consists in the interpolation between temporally adjacent images where the target feature is cloud-free. However, given the availability of weather-insensitive SAR data of the scene, it makes sense to pursue fusion-based solutions, exploiting SAR images that may be temporally very close to the target date, as it is well known that radar images can provide precious information on vegetation [81, 51, 82, 47]. Even if this holds true, however, it is by no means obvious how to exploit such dependency.

Some fusion techniques have been proposed for spatio-temporal NDVI super-resolution [50] or prediction [23], they use exclusively optical data. But, none of these papers attempts to directly estimate a pure multispectral feature, NDVI or the likes, from SAR data. Conversely, in different works, multi-sensor SAR-optical fusion at feature level, for the purpose of vegetation monitoring is considered [46, 21, 31, 47, 83, 26]. In [21] ALOS POLSAR and

Landsat time-series were combined at feature level for forest mapping and monitoring. The same problem was addressed in [31] through a decision-level approach. In [83] the fusion of single-date S1 and simulated S2 was presented for the purpose of classification. In [47], instead, RADARSAT-2 and Landsat-7/8 images were fused, by means of 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 [26] for the purpose of land cover mapping. To this end, the texture maps extracted from the SAR image were combined with several indices drawn from the optical bands.

In this chapter, to address this problem, the power of learning capability of deep learning methods is exploited. To this purpose a three-layer convolutional neural network (CNN) was designed and trained to account for both temporal and cross-sensor dependencies.

In Section 2.2 is introduced the dataset used in the following, then in Section 2.3 the fusion of optical and SAR images is performed aiming to the NDVI reconstruction [84] is proposed, and lastly, in Section 2.4 is proposed the regression on NDVI using only SAR images [85].

2.2 Dataset and Problem Statement

In the following, several solutions are employed in order to estimate a target optical feature at a given date from images acquired at adjacent dates, or even from the temporally closest SAR image. Such different solutions also reflect the different operating conditions found in practice. The main application is the reconstruction of a feature of interest in a target image which is available but 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 actually available.

Here, the focus is on the Normalized Difference Vegetation Index estimation, but it is straightforward to apply the same framework to other optical features.

With reference to 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:

$$\mathrm{NDVI} \triangleq \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red}} \in [-1, 1] \tag{2.1}$$

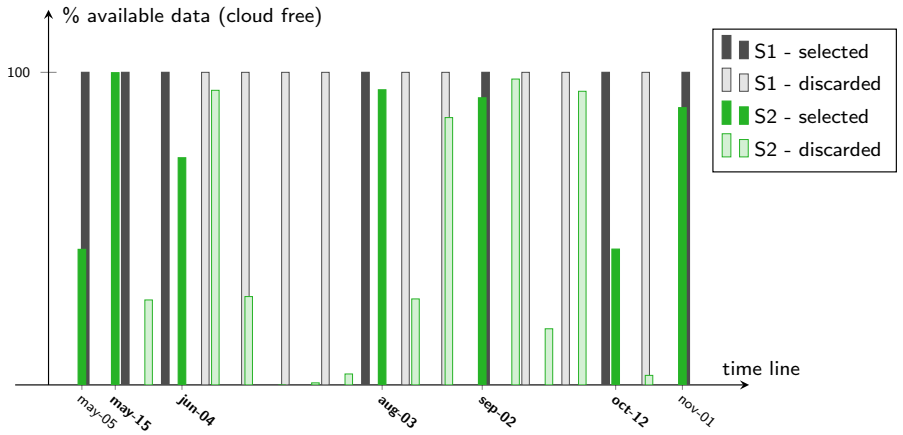


Figure 2.1: Available S1 (black) and S2 (green) images over the period of interest. The bar height indicates the fraction of usable data. Solid bars mark selected images, boldface date mark test images.

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 [80] 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.

Fig. 2.1 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

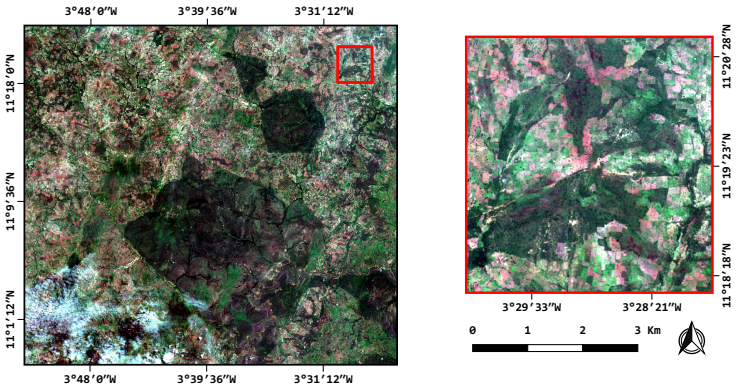


Figure 2.2: RGB representation of the 5253×4797 S2-Koumbia dataset (August 3rd, 2016), with a zoom on the area selected for testing.

insensitive to meteorological conditions.

For the purpose of training, validation and testing of the proposed methods, 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 have been considered. For the selected S2 images (solid bars in Fig. 2.1) the corresponding dates are indicated on the x-axis. The 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/or 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.

To allow temporal analyses, a test area, of size 470×450 is chosen, which is cloud-free in all the selected dates, and hence with available reference ground-truth for any possible optical feature. Fig. 2.2 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.

The dataset comprises also 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. All images have been calibrated (VH/VV intensities to sigma nought) 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 [86]. 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) [87] and preprocessed using the Multi-sensor Atmospheric Correction and Cloud Screening (MACCS) level 2A processor [88] 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, two more features are employed: 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.

2.3 Regression on NDVI fusing optical and SAR images

In this Section, several CNN-based algorithms are presented in order to estimate the NDVI through the fusion of optical and SAR Sentinel data, instead of using only SAR data [84].

With reference to a specific case study, temporal sequences of S1 SAR data and S2 optical data, covering the same time lapse, with the latter partially covered by clouds have been collected. Both temporal and cross-sensor (S1-S2) dependencies are used to obtain the most effective estimation protocol. From the experimental analysis, very interesting results emerge. On one hand, when

only optical data are used, CNN-based methods outperform consistently the conventional temporal interpolators. On the other hand, when also SAR data are considered, a further significant improvement of performance is observed, despite the very different nature of the involved signals. It is worth underlining that no peculiar property of the NDVI was exploited, and therefore these results have a wider significance, suggesting that other image features can be better estimated by cross-sensor CNN-based fusion.

2.3.1 Proposed prediction architectures

In the following developments, with reference to a given target S2 image acquired at time t, the symbols defined below will be used:

- 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.

The several models considered here differ in the composition of the input stack $\mathbf{x}$, while the output is always the NDVI at the target date, that is, $\mathbf{y} = F$. Apart from the input layer, the CNN architecture is always the same, depicted in Fig. 2.3, with hyper-parameters summarized in Tab. 2.1. A focus about the choice of this configuration is postponed to the end of this Subsection. This relatively shallow CNN is characterized by a rather small number of weights (as CNNs go), counted in Tab. 2.1, and hence can be trained with a small amount of data. Moreover, slightly different architectures have proven to achieve state-of-the-art performance in closely related applications, such as super-resolution [54] and data fusion [18, 34].

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 the compact interval [-1,1], as required by the NDVI semantics (Eq. 2.1), 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 simply take

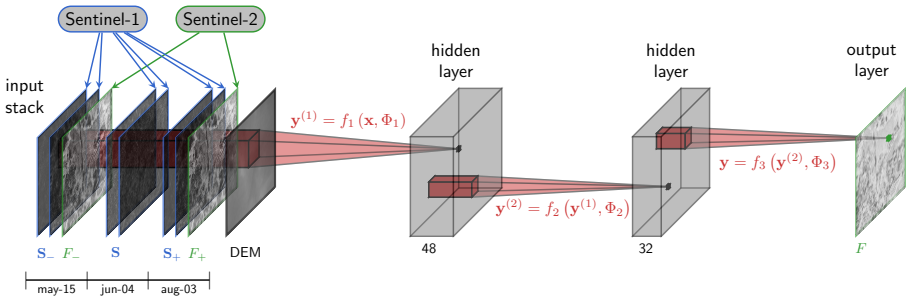


Figure 2.3: Proposed CNN architecture. The depicted input corresponds to the Optical-SAR+ case. Other cases use a reduced set of inputs.

Table 2.1: 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	$32\times(5\times5)$
# 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 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, but this additional transformation does not participate in the error back propagation, hence should be considered external to the network. Through preliminary experiments, it has found this latter solution more effective than the former, for this task, and therefore the train of the CNN is performed considering a linear activation in the last layer, $g_3(\mathbf{z}^{(3)}) = \mathbf{z}^{(3)}$.

The different solutions considered here, differ for the available input data and the required response time.

Concerning data, the estimation could be based on optical-only, SAR-only, and optical+SAR data. When using SAR images, the DEM is included since may convey relevant information on them.

Instead, the DEM is useless, and hence neglected, when only optical data are used. All 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 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 by means of both scattering properties and spectral signatures. The analysis of the *SAR-only* case will allow to understand if significant dependencies exist between the NDVI and SAR images, and if a reasonable quality can be achieved even when only this source is used for estimation.

  To this aim, the temporal dependencies is not considered in this case, trying to estimate a S2 feature from the closest S1 image only.

- The *optical-SAR* fusion is the case of highest interest. Given the most 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.

Turning to response time, except for the SAR-only case, will be 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 whenever the data must be used right away for the application of interest. This is the case, for example, of early warning systems for food security. It will be included here also the case in which the closest SAR image becomes available after time t, since the maximum delay is at most 5 days. Hereinafter, it will be referred to this "nearly" causal case as Causal for short.

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

Table 2.2: 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

– 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. 2.2 summarizes all these different solutions.

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 the relatively small networks used here ($\sim 7 \cdot 10^4$ weights to learn in the worst case – see Tab. 2.1), a set of 19000 patches is sufficient for accurate training, as already observed for other generative tasks like super-resolution [54] or pansharpening [18] addressed with CNNs of similar size. Therefore, with the patch extraction process used, this number requires an overall cloud-free area of about 1000×1000 pixels, namely, about 4% of the 5253×4797 target scene (Fig. 2.2). 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, for the chosen date). However, if the scene is almost completely covered by clouds

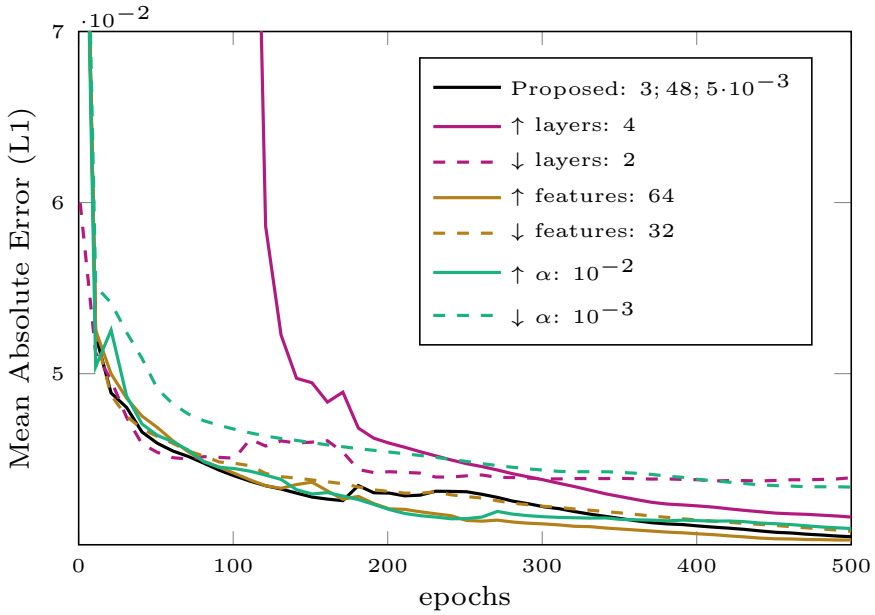


Figure 2.4: Loss functions for the validation dataset of August 3th. The proposed Optical-SAR model (with 3 layers, 48 features in the 1st layer, and $\alpha = 5{\cdot}10^{-3}$) is compared to several variants obtained by changing one hyper-parameter at time.

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. 2.3.3. 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. In particular, it has been found exprimentally optimal values for these parameters which are $\alpha = 0.5 \cdot 10^{-3}$ and $\mu = 0.9$.

To conclude this section in Fig. 2.4 some preliminary results are presented showing the evolution of the loss computed on the validation dataset during the training process for a sample proposed architecture and for some deviations

Table 2.3: 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

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, in Fig. 2.4 are gathered 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. 2.4 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.

Besides accuracy, complexity is also affected by architectural choices. For the same variants compared in Fig. 2.4, the average training time is reported in Tab. 2.3, 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 2.4: 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**

2.3.2 Experimental results

In order to assess the accuracy of the proposed solutions, two reference methods are considered 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 [80] 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 [89]. This is a practical solution used by analysts [80] 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, a further causal interpolator is proposed 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, as a further reference the affine regressors are considered, 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

Table 2.5: 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**

the following alternatives are considered for comparison:

$$\widehat{F} = \begin{cases} F_- & \text{Interpolator/C} \\ \frac{\Delta_+}{\Delta_-+\Delta_+}F_- + \frac{\Delta_-}{\Delta_-+\Delta_+}F_+ & \text{Interpolator ([80])} \\ a_-F_- + b & \text{Regressor/C} \\ a_-F_- + a_+F_+ + b & \text{Regressor} \end{cases}$$

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

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

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 2.4, 2.5 and 2.6, respectively, for all proposed and reference methods and for all dates.

The target dates are shown in the first row, while the second row gives the temporal gaps (days) between the target and the previous and next dates used for prediction, respectively. The following two lines show results for fully cross-sensor, that is, SAR-only estimation, while in the rest of the table all causal (top) and non-causal (bottom) models are grouped together, 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 Figg. 2.5 and 2.6.

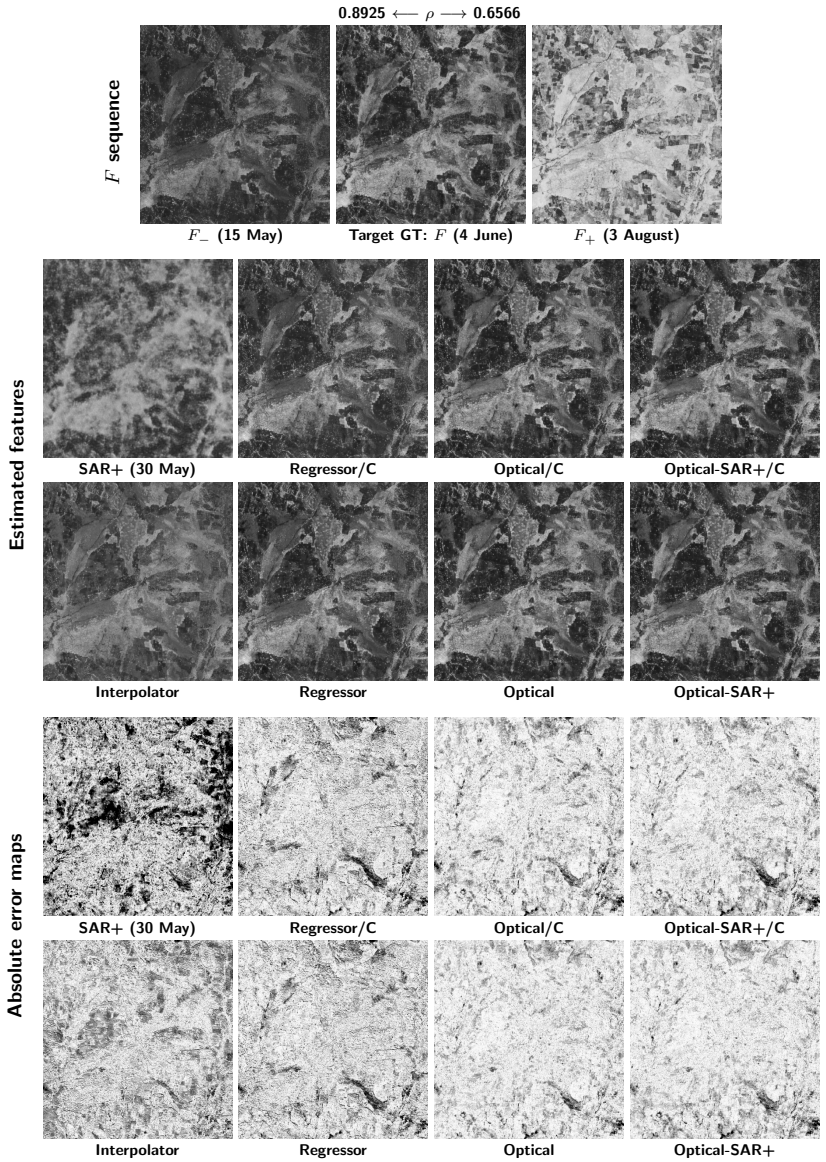


Figure 2.5: Sample results for the jun-04 target date. Top row: previous, target, and next NDVI maps of the crop selected for testing. Second/third rows: NDVI maps estimated by causal/non-causal methods. Last two rows: corresponding absolute error images.

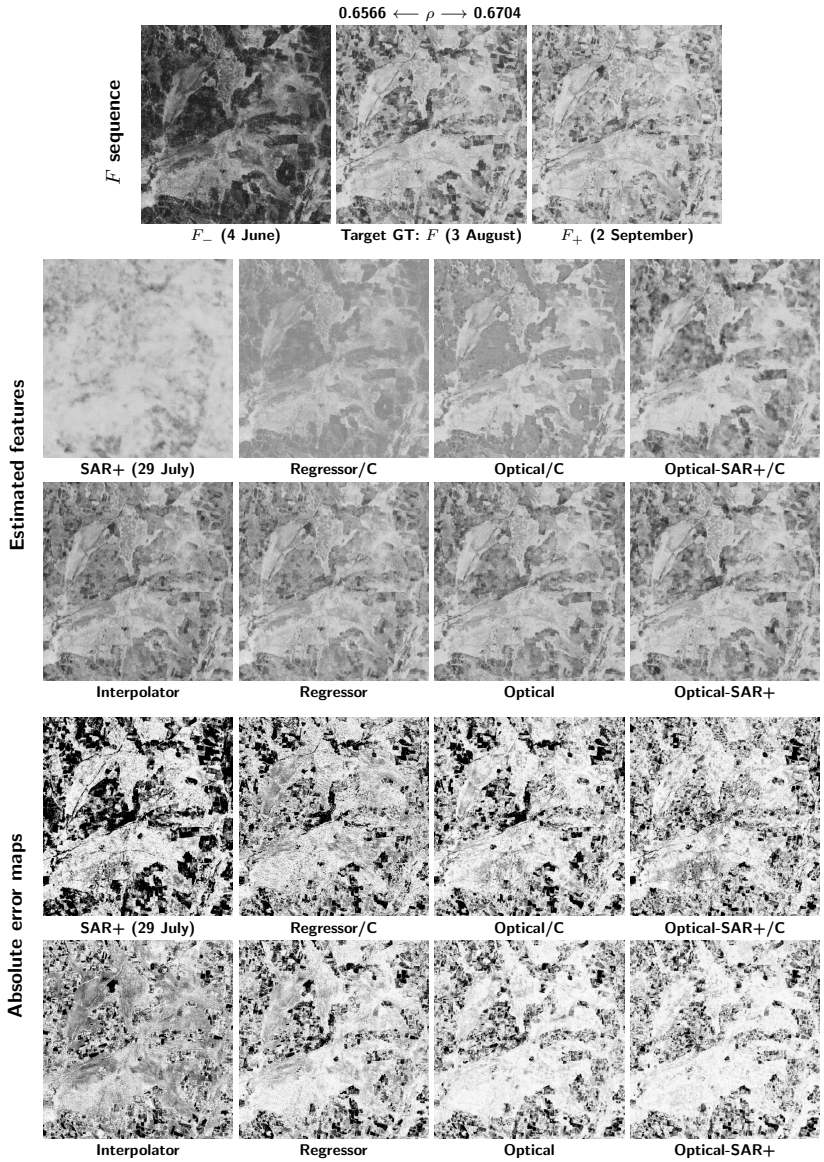


Figure 2.6: Sample results for the aug-03 target date. Top row: previous, target, and next NDVI maps of the crop selected for testing. Second/third rows: NDVI maps estimated by causal/non-causal methods. Last two rows: corresponding absolute error images.

Table 2.6: 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**

2.3.3 Discussion

In this section follows a discussion about the accuracy of the proposed methods both objectively, through the numerical results gathered in Tabb. 2.4-2.6, and subjectively by visually inspecting Figg. 2.5 and 2.6. Then in the final part of the section follows a discussion about critical conditions when training data cannot be retrieved from the target.

Observing Tab. 2.4, focusing on the ρ, and, in particular, on the last column with average values, which accounts well for the main trends, 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, it can be observed, 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 it can be observed 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 a much better baseline is considered, 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 to the PSNR and SSIM data, shown in Tab. 2.5 and Tab.2.6. 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 two target dates have been considered, 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

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

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.2.5 and Fig.2.6, 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.

To conclude this discussion a focus on the learning related issues must be done. 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). To what extent we a machine learning model trained elsewhere could be used? This is a key problem in machine learning, and is very relevant for a number of remote sensing applications, such as coregistration [90] or pansharpening [34]. In [90] it has been underlined the importance of selecting training data which are homogeneous with the target. In [34] it is shown that the performance of a CNN can drop dramatically without a proper domain adaptation strategy and target-adaptive solution is proposed.

Table 2.7: 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 gain insight into this critical point a simple test that gives an idea of the scale of the problem is considered. In particular, several training-test mismatches are considered 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. 2.7.

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. 2.7, for two sample target dates, are showed 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) an over estimation can be observed,

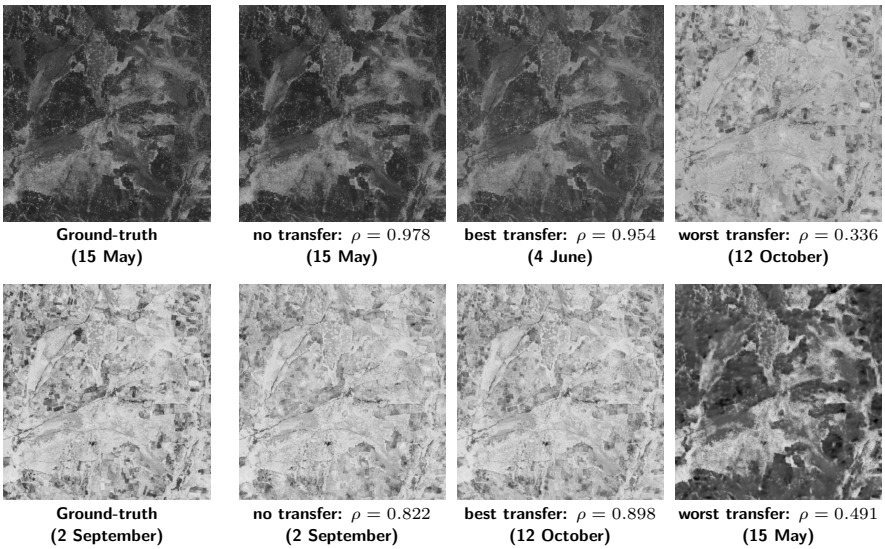


Figure 2.7: Temporal transfer learning tested on may-15 (top) and sep-02 (bottom). From left to right are the target F followed by estimates provided by model Optical-SAR+ trained on the target date (no transfer) and on two alternative dates (best and worst cases).

while in the opposite case an under estimation is evident.

These results suggest that in presence of abrupt changes (e.g., due to forest fires), the reconstruction could perform quite poorly. Indeed, these phenomena are often localized and cover only a small part of the dataset both in spatial and temporal domains. Further studies are needed to improve the robustness of the proposed method with respect to such particular cases.

2.4 Regression on NDVI with SAR

In the previous Section several fusion settings have been taken in consideration (single vs multiple dates and/or sensors) and for each case a dedicated convolutional neural network (CNN) was designed and trained [84]. Major attention was put in particular on schemes for (causal or not) temporal prediction with eventual inclusion of SAR components as complement to available cloud-free optical features temporally close to the target cloudy date. Minor care was put on the full cross-sensor case where only SAR images feed the

prediction network, although it is a very challenging and important case, since there can be such unfortunate conditions where the closest available Sentinel-2 images of the target are too far apart for a reliable prediction. On the other hand, several works dealing with the fusion of SAR and optical data for the purpose of vegetation monitoring [91, 21, 47, 83, 92] indicate the potential of SAR data for recovering missing optical features. Moreover, as already mentioned, Sentinel-1 sattelites provide double polarized (VV-VH)[3] SAR images that are useful since that vegetation is mainly characterized by a vertical geometry, therefore it backscatters signals transmitted in vertical polarization, while it tends to be transparent to horizontally polarized ones. For example, cover crops such as wheat or corn present high VV response, so as grasslands and seedlings give higher VH response [81]. Motivated by the above considerations, keeping on the track of Section2.3 [84], in the following, are proposed several CNN models with different input settings that include only SAR. These models have been trained, all including a DEM D as additional input band as it provides an additional gain when used in combination with the SAR [84]. In particular the CNN models considered in this case take up to three temporally adjacent SAR images.

2.4.1 Only SAR Method

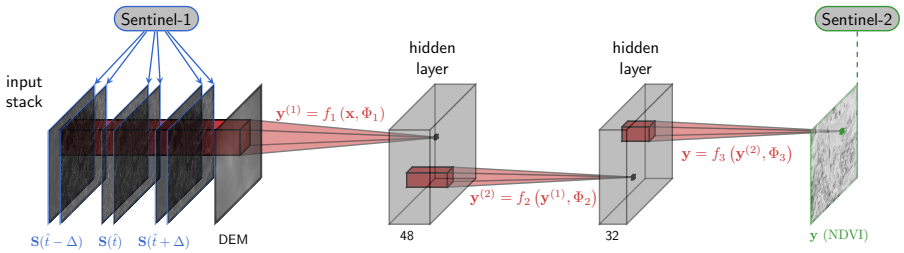


Figure 2.8: CNN architecture of the proposed method.

Said t the time instant of the target NDVI, in the simplest case introduced in the previous Section, only the closest VV-VH SAR signal $\mathbf{S}(\hat{t}) = (S^{VV}, S^{VH})$ is considered as input: $\mathbf{x} = (\mathbf{S}(\hat{t}), D)$.[4]

[3] VV and VH refer to the vertical and horizontal backscattered components, respectively, when a vertically polarized signal is transmitted.

[4] The temporal misalignement $|t - \hat{t}|$ is function of the geographic location of the target. In this case it is smaller than 5 days in the worst case.

	ConvLayer 1	$g_1(\cdot)$	ConvLayer 2	$g_2(\cdot)$	ConvLayer 3
Shape	48×$b_{\mathbf{x}}$×9×9	ReLU	32×48×5×5	ReLU	1×32×5×5
Learn. rate	$5 * 10^{-3}$		$5 * 10^{-3}$		$5 * 10^{-3}$
Momentum	0.9		0.9		0.9

Table 2.8: Hyper-parameters of the CNN architecture.
Shape = # features × # channels × 2D support. $b_{\mathbf{x}} \in \{3, 5, 7\}$

input x	method	may-5	may-15	jun-4	aug-3	sep-2	oct-12	nov-1
$\mathbf{S}(\hat{t})$, D	**L. Reg.**	0.650	0.690	0.554	0.340	0.354	0.480	0.632
	L. Reg.*	0.774	0.782	0.713	0.548	0.500	0.647	0.779
	CNN	0.830	0.832	0.791	0.578	0.541	0.668	0.805
	CNN*	**0.859**	**0.856**	**0.825**	**0.617**	**0.573**	**0.706**	**0.835**
$\mathbf{S}(\hat{t})$, $\mathbf{S}(\hat{t}-\Delta)$, D	**L. Reg.**	0.659	0.736	0.685	0.492	0.408	0.511	0.670
	L. Reg.*	0.760	0.797	0.775	0.612	0.531	0.644	0.777
	CNN	0.825	0.855	0.854	0.657	0.600	0.686	0.811
	CNN*	**0.854**	**0.869**	**0.868**	**0.680**	**0.625**	**0.728**	**0.841**
$\mathbf{S}(\hat{t})$, $\mathbf{S}(\hat{t}-\Delta)$, $\mathbf{S}(\hat{t}+\Delta)$, D	**L. Reg.**	0.718	0.748	0.713	0.525	0.420	0.598	0.714
	L. Reg.*	0.789	0.797	0.786	0.630	0.536	0.693	0.797
	CNN	0.853	0.860	0.869	0.708	0.601	0.746	0.834
	CNN*	**0.872**	**0.875**	**0.878**	**0.723**	**0.629**	**0.768**	**0.857**

Table 2.9: Correlation coefficient, $\rho \in [-1, 1]$. (*) marks solutions for despeckled SAR images.

Moving from this baseline solution here other settings are considered including one or both the nearest SAR acquisitions $\mathbf{S}(\hat{t} \pm \Delta)$, where Δ depends on the geographic position (12 days in this experiments). In addition, the contribution of despeckled SAR images as input on the train of the network is also considered.

As in Section 2.3, all solutions differ in the input layer, sharing the same shallow architecture. In Fig.2.8 is depicted the network architecture used in the following, while in Tab.2.8 the hyper-parameters, learning rates and momentum used for the implementation of the stochastic gradient descent (SGD) algorithm for training. Also in this case, has been adopted the L_1 norm between the estimated and the target NDVI as loss.

input x	method	may-5	may-15	jun-4	aug-3	sep-2	oct-12	nov-1	mean
$\mathbf{S}(\hat{t})$, D	**L. Reg.**	20.80	21.46	17.50	16.65	18.29	18.92	18.65	18.90
	L. Reg.*	22.18	22.38	18.99	17.64	18.98	20.04	20.39	20.09
	CNN	23.37	23.91	20.08	17.60	18.69	20.61	21.50	20.83
	CNN*	**24.32**	**24.57**	**20.76**	**17.94**	**19.28**	**20.97**	**22.04**	**21.41**
$\mathbf{S}(\hat{t})$, $\mathbf{S}(\hat{t}-\Delta)$, D	**L. Reg.**	20.53	21.91	18.30	17.01	18.53	19.58	19.07	19.28
	L. Reg.*	21.47	22.51	19.74	18.15	19.18	20.00	20.40	20.21
	CNN	22.99	24.71	21.01	17.86	19.59	21.44	21.42	21.29
	CNN*	**23.73**	**24.94**	**21.55**	**18.47**	**19.73**	**21.82**	**21.93**	**21.74**
$\mathbf{S}(\hat{t})$, $\mathbf{S}(\hat{t}-\Delta)$, $\mathbf{S}(\hat{t}+\Delta)$, D	**L. Reg.**	21.38	22.11	18.96	17.57	18.61	19.65	19.50	19.68
	L. Reg.*	22.56	22.66	19.92	18.34	19.21	20.45	20.66	20.54
	CNN	24.01	24.99	21.76	19.01	19.30	21.59	22.03	21.81
	CNN*	**24.74**	**25.04**	**22.23**	**19.17**	**19.79**	**21.84**	**22.67**	**22.21**

Table 2.10: Peak signal-to-noise ratio (PSNR) [dB] (*) marks solutions for despeckled SAR images.

input x	method	may-5	may-15	jun-4	aug-3	sep-2	oct-12	nov-1	mean
$\mathbf{S}(\hat{t})$, D	**L. Reg.**	0.4182	0.4739	0.2800	0.2833	0.3111	0.3412	0.4108	0.3598
	L. Reg.*	0.4891	0.5460	0.3835	0.4298	0.4330	0.4622	0.5531	0.4709
	CNN	0.5510	0.5726	0.4640	0.3707	0.3718	0.3674	0.4490	0.4495
	CNN*	**0.6514**	**0.6584**	**0.5968**	**0.4536**	**0.4563**	**0.4829**	**0.5817**	**0.5544**
$\mathbf{S}(\hat{t})$, $\mathbf{S}(\hat{t}-\Delta)$, D	**L. Reg.**	0.4107	0.5089	0.3280	0.3211	0.3341	0.3922	0.4478	0.3918
	L. Reg.*	0.4847	0.5537	0.4174	0.4492	0.4416	0.4553	0.5623	0.4806
	CNN	0.5594	0.6189	0.5508	0.3784	0.4120	0.4332	0.4809	0.4905
	CNN*	**0.6592**	**0.6763**	**0.6335**	**0.4837**	**0.4832**	**0.5238**	**0.6006**	**0.5800**
$\mathbf{S}(\hat{t})$, $\mathbf{S}(\hat{t}-\Delta)$, $\mathbf{S}(\hat{t}+\Delta)$, D	**L. Reg.**	0.4519	0.5191	0.3647	0.3723	0.3462	0.3987	0.4822	0.4193
	L. Reg.*	0.5040	0.5586	0.4250	0.4566	0.4423	0.4633	0.5687	0.4884
	CNN	0.6098	0.6418	0.6029	0.4545	0.4146	0.4470	0.5225	0.5276
	CNN*	**0.6793**	**0.6870**	**0.6522**	**0.5013**	**0.4875**	**0.5332**	**0.6176**	**0.5940**

Table 2.11: Structural similarity measure (SSIM) [-1,1] (*) marks solutions for despeckled SAR images.

2.4.2 Experimental results

In accordance with Section2.3 the area under study is located in the province of Tuy, Burkina Faso, around the commune of Koumbia. The 5253×4797 pixels scene is monitored on the same seven sample dates between May 5th and November 1st 2016, since it is the period that corresponds to a regular agricultural season in the region, but also with the rainy season. The same area of 470×450 pixels was reserved for test, while small tiles for training were uniformly sampled from the remaining data.

As comparative approach, here, for each input configuration a linear regressor estimated over 10^5 cloud-free points from the training segments has set-up. The results were numerically assessed with three commonly used indicators: correlation coefficient (ρ), peak signal-to-noise ratio (PSNR), and structural similarity measure (SSIM) gathered in Tables 2.9, 2.10 and 2.11.

First of all, observing Tabb. 2.9 - 2.11 that there can be large variations from one date to another, suggesting a variable correlation degree between NDVI and SAR along the seasons, which are likely due to the evolution of the vegetation. Second, it is registered a clear and consistent performance gain over the linear regression approach. Third, spatial regularized (despeckled) SAR data allow a better reconstruction of the NDVI. This is particularly true for the regressor, but not for the CNN-based approach especially when more dates are considered. This is a confirm that the network is able to implicitly regularize the input bands when these have not been despeckled in advance, thanks to the embedded spatio-temporal integrations.

The experimental analysis is then completed showing some sample results in Figg. 2.9- 2.10. The top row shows the $\mathbf{S}(\hat{t})$ component always put in the input $\mathbf{x}$ and the ground-truth of $\mathbf{y}$ (right). The next rows collect the results obtained with different methods using only $\mathbf{S}(\hat{t})$ (left), including also the preceding date (middle), and using all three dates (right). The second row gathers the results obtained with the regressor over despeckled data. Those provided by the proposed method without or with despeckling are gathered in the third and fourth rows, respectively. The visual inspection reveals some residual blur likely due to the intrinsic differences between SAR and optical imaging systems.

2.5 Conclusions

In this chapter a CNN-based data fusion approach to estimate a widespread spectral feature for vegetation monitoring, the NDVI, from multitemporal Optical and/or SAR images has been described ([84, 85]. Very promising results highlight the strong relationship between SAR and NDVI which can be captured through a deep learning approach such as the proposed solutions.

Although the introduced method refers to a specific feature and has been tested on coupled time-series of Sentinel-1 and Sentinel-2, it is rather general and can be readily extended to other features and practical real-world applications. The encouraging results obtained suggest further investigation on these topics, in particular focusing on deeper architecture and different learning strategies.

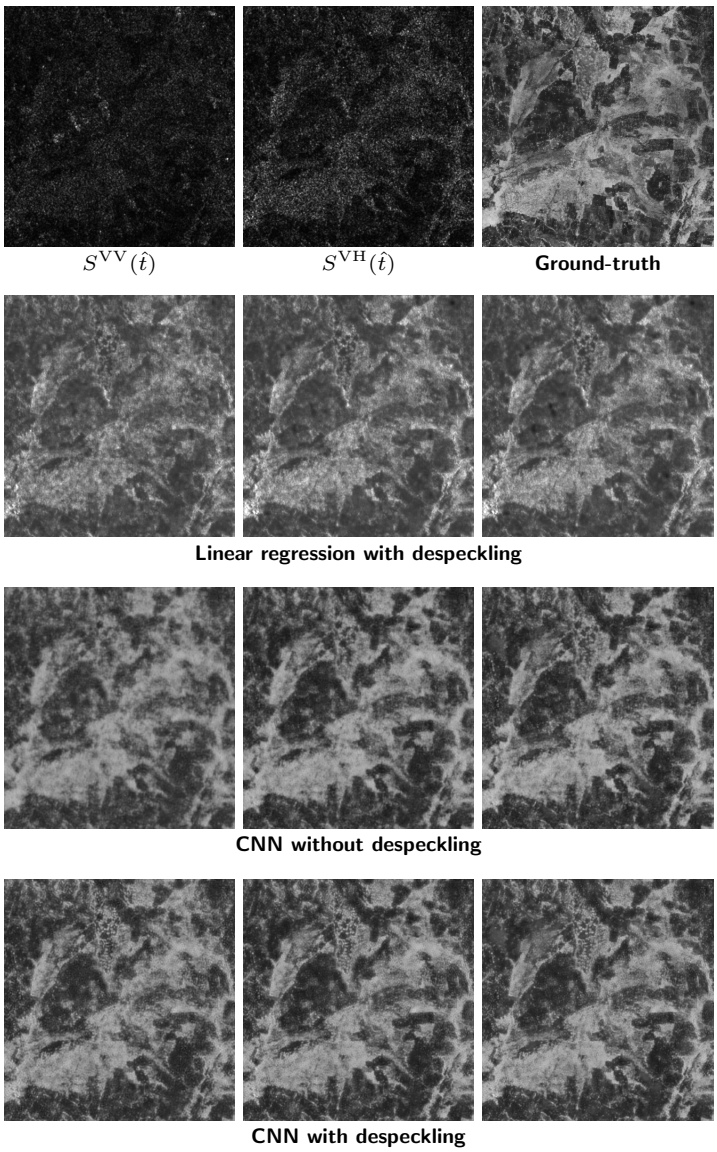


Figure 2.9: Results obtained on the test image of June 4th. Sample SAR bands and target $\mathbf{y}$ on the top row. NDVI estimations with three compared methods in the next three rows. From left to right one, two and three adjacent SAR acquisitions are considered in input, respectively.

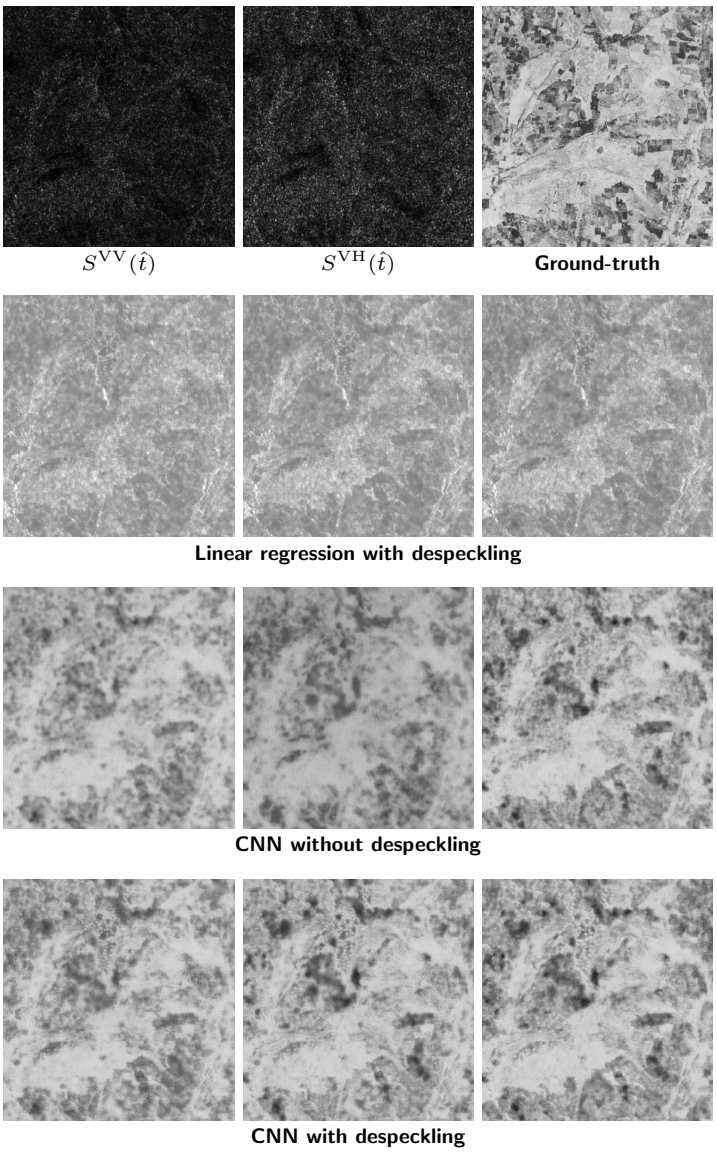


Figure 2.10: Results obtained on the test image of August 3rd. Sample SAR bands and target $\mathbf{y}$ on the top row. NDVI estimations with three compared methods in the next three rows. From left to right one, two and three adjacent SAR acquisitions are considered in input, respectively.

Chapter 3

Forest monitoring

In this chapter is faced the problem of forest mapping from TanDEM-X data by means of Convolutional Neural Networks (CNNs). The study aims to highlight the relevance of domain-related features for the extraction of the information of interest, thanks to their joint nonlinear processing through CNN. In particular, the focus is on the main InSAR features as the backscatter, coherence, and volume decorrelation, as well as the acquisition geometry through the local incidence angle by using different state-of-the-art CNN architectures. In particular, three state-of-the-art CNN architectures, such as ResNet, DenseNet, and U-Net are compared [93].

3.1 Forest monitoring

Forests are of paramount importance for the Earth's ecosystem, since they play a fundamental role in reducing the concentration of carbon dioxide in the atmosphere and regulating global warming. The study of deforestation and development of global forest coverage and biomass is necessary to assess how forests impact the ecosystem. In this framework, remote sensing represents a powerful tool for a regular monitoring at a global scale of vegetated areas. A successful example is the product provided in [94], which maps World's forest coverage and its evolution between the years 2000 and 2010, by exploiting multi-spectral data provided by the Landsat optical spaceborne mission. Other notable examples include the fusion of multispectral and Lidar data [95] or the use of hyperspectral images [96]. However, as well known, passive imaging systems are useless under cloudy conditions, whereas synthetic aperture radar (SAR) systems, providing a continuous large-scale coverage

ranging from mid- to very high-resolution, can operate effectively regardless of weather and daylight conditions. This feature is particularly important for tropical zones which are characterized by heavy rain seasons. As originally proposed in [97], SAR backscatter data from the ALOS PALSAR mission have been fruitfully applied to global forest mapping in [98]. On the other hand, SAR interferometry (InSAR) provides yet more descriptive parameters, such as the interferometric coherence, that can better explain the nature of the observed target [99, 100].

Among InSAR systems, the German TanDEM-X SAR mission provides single-pass interferometric data at X band. The simultaneity of the bistatic acquisition pair guarantees high correlation between the master and slave images, enabling for high resolution interferometric measurements with an unprecedented quality. The constellation comprises two twin satellites flying in a bistatic close-orbit configuration, which allows for a flexible selection of the acquisition geometries and, in particular, of the interferometric baselines [101]. The main goal of the mission was the generation of a global consistent high-resolution digital elevation model (DEM) with unprecedented accuracy, which has been successfully completed in 2016 [102]. Besides the nominal DEM product, for each bistatic interferometric TanDEM-X acquisition, additional quantities can be computed as by-pass products. Indeed, the bistatic acquisition is not affected by temporal decorrelation, allowing for an accurate isolation of volume scattering phenomena from the interferometric coherence. This feature was exploited in [103, 104], where the authors presented a framework for the development of a global TanDEM-X forest/non-forest map [105] as described more in details in Section 3.2.1.

Motivated by the works in [103, 104], in this chapter is proposed the use of CNNs for high-resolution forest mapping using TanDEM-X data, aiming at proving the effectiveness of deep learning for the generation of high-quality products. In particular, the contribution is two-fold: i) finding the CNN model that better fits to the problem at hand, and ii) assessing the impact on the prediction due to handcrafted SAR features used as additional input. Three modular architectures where built according to three state-of-the-art approaches: ResNet [106], DenseNet [107], and U-Net [108], respectively. For each architecture different input combinations are tested, ranging from the single-band SAR image to a 4-band stack that encloses three additional features: the incidence angle, the interferometric coherence, and the volume decorrelation contribution, which carry relevant information on the nature of the illuminated target.

The Chapter is organized as follows. Section 3.2 provides a brief summary of the baseline reference method and introduces basic concepts about CNNs. Then, the proposed methods are described in Section 3.3, while the used datasets and experimental results are presented and discussed in Section 3.4. Finally, conclusions are drawn in Section 3.5.

3.2 Background Concepts

This section provides background concepts that will introduce the reader to the context of this chapter from both applicative and methodological points of view. In particular, Section 3.2.1 deals with the definition of some SAR features which can be associated to TanDEM-X data that have been demonstrated to be very effective for forests mapping [109, 104].

Then, Section 3.2.2 recalls general concepts about the CNNs used here, contextualized to the cases of classification and segmentation.

3.2.1 Baseline algorithms for forest mapping using TanDEM-X

The framework presented in the following is born from the experience matured within the TanDEM-X Forest/Non-Forest Map project, developed at the Microwaves and Radar Institute at the German Aerospace Center (DLR), within the framework of the TanDEM-X mission [103]. Its goal was the generation of a global forest/non-forest classification mosaic from TanDEM-X bistatic InSAR data, acquired for the generation of the global DEM between 2011 and 2015, in stripmap single polarization (HH) mode. The derived product has been made available in May 2019 and can be downloaded free of charge for scientific use [105].

Several products, systematically provided by the TanDEM-X system, can be exploited for classification purposes, such as the calibrated amplitude, the bistatic coherence, and the digital elevation model (DEM). As an example, Figure 3.1 shows a sample image set. Together with the absolutely calibrated backscatter image β^0, several features of interest for the present work are shown, that are the local incidence angle θ_{i}, the interferometric coherence γ_{Tot}, and the volume correlation coefficient γ_{Vol}. These features have been proven to be effective for forest classification in several works [103, 104] and are easy to compute. Baseline solutions considered in the following, in fact, use volume correlation coefficient and local incidence angle. For these reasons these features are included as additional input layers.

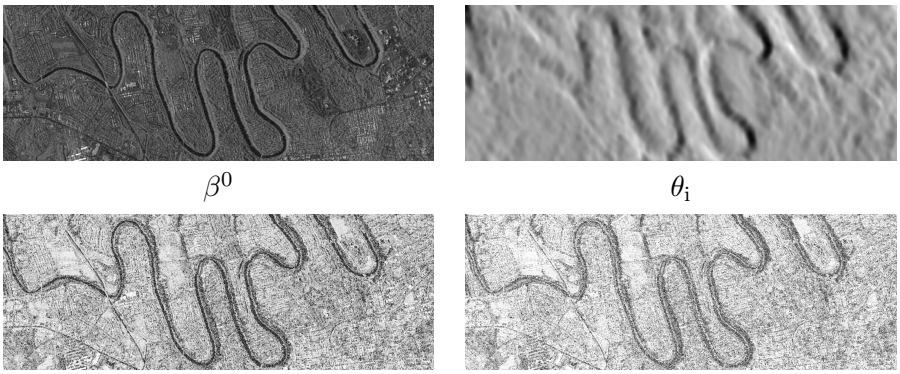


Figure 3.1: Sample data. From the top to the bottom image: absolutely calibrated backscatter β^0, local incidence angle θ_{i}, interferometic coherence γ_{Tot}, and volume correlation coefficient γ_{Vol}.

In the following a suitable definition of them with a related description of the meaning is provided.

The interferometric coherence γ_{Tot} represents the main indicator for assessing the quality of an interferogram and is defined as the normalized cross-correlation coefficient between the interferometric images pair

$$\gamma_{\mathrm{Tot}} = \frac{|E[xy^*]|}{\sqrt{E[|x|^2]E[|y|^2]}}, \tag{3.1}$$

where $E[\cdot]$ is the statistical expectation, $*$ the complex conjugate operator, and $|\cdot|$ the absolute value. x and y represent the master and slave image, respectively. γ_{Tot} varies between 0 and 1 and it is an image itself being computed locally at each pixel location using a sliding window averaging.

The interferometric coherence is affected by several decorrelation factors which can be singularly interpreted and computed. In particular, as shown in [101], γ_{Tot} can be factorized as

$$\gamma_{\mathrm{Tot}} = \gamma_{\mathrm{SNR}}\gamma_{\mathrm{amb}}\gamma_{\mathrm{quant}}\gamma_{\mathrm{az}}\gamma_{\mathrm{rg}}\gamma_{\mathrm{Vol}}\gamma_{\mathrm{Temp}}, \tag{3.2}$$

where the different factors take into account decorrelations due to limited signal-to-noise ratio (γ_{SNR}), range and azimuth ambiguities (γ_{amb}), quantization noise (γ_{quant}), relative shift of the Doppler spectra (γ_{az}), baseline differ-

ences (γ_{rg}), volume scattering (γ_{Vol}), and temporal changes (γ_{Temp}).

The factor γ_{Vol}, also called volume correlation factor, is severely affected by the presence of multiple scattering from volumes, that are easily penetrated by the incident electromagnetic waves. The received signal consists therefore of the coherent superposition of multiple reflections. This term is a reliable indicator of the presence of vegetation on ground and can be extrapolated form the interferometric coherence as

$$\gamma_{\mathrm{Vol}} = \frac{\gamma_{\mathrm{Tot}}}{\gamma_{\mathrm{SNR}}\gamma_{\mathrm{amb}}\gamma_{\mathrm{quant}}\gamma_{\mathrm{az}}\gamma_{\mathrm{rg}}\gamma_{\mathrm{Temp}}}. \tag{3.3}$$

In this specific case, all factors at the denominator but γ_{Temp} have been estimated as described in [103]. γ_{Temp} is equal to one, since are considered TanDEM-X bistatic acquisitions, which are not affected by temporal decorrelation.

Being γ_{Vol} in turn a very sensitive indicator of the presence of vegetation on ground, it was therefore selected in [103] as main feature for forest mapping with TanDEM-X data at global scale. Moreover, it also has to be remarked that γ_{Vol} is strongly influenced by the acquisition geometry, and in particular by the height of ambiguity h_{amb}. This latter figure represents the elevation difference corresponding to a complete 2π cycle in the interferogram and, for the bistatic systems, is defined as

$$h_{\mathrm{amb}} = \frac{\lambda r \sin\theta_{\mathrm{i}}}{B_{\perp}}, \tag{3.4}$$

where λ is the wavelength, r the slant range distance, and $B_{\perp}$ the baseline perpendicular to the line of sight. As it has been demonstrated in [109], the lower the height of ambiguity, the higher the volume decorrelation contribution and, hence, the lower the γ_{Vol}. For this reason, in order to discriminate between forested and non-forested areas, a supervised geometry-dependent fuzzy clustering classification approach, which takes into account the geometric acquisition configuration for the definition of the cluster centers, was proposed in [103] and applied to each acquired TanDEM-X scene for the generation of the global product. Additionally, a certain variability of the interferometric coherence at X band was observed among different forest types, mainly due to changes in forest structure, density, and tree height. This aspect led to an adjustment of the algorithm settings and, in particular, to the derivation of different sets of cluster centers, depending on the specific type of forest.

In order to limit the computational burden, the global TanDEM-X data set of quicklook images with 50×50 m^2 ground resolution was used for the

generation of the global forest/non-forest map. Full-resolution results were obtained on a subset of 12×12 m^2 resolution TanDEM-X images using an enhanced version of [103] proposed in [104], aimed to preserve both global classification accuracy and local precision thanks to the introduction of non-local filtering. This latter work represents the starting point of the following dissertation and will be therefore referred to as Baseline. The same work [104] also shows the forest prediction results masking water and built-up regions using external ground-truths. This was motivated by the sensitivity of the volume correlation factor to these two classes. This solution make sense as in many real-world practical applications one may rely on the availability of urban and water maps. For these reasons we decided to keep also this variant in our experiments, which will be referred to as Baseline+.

3.2.2 Convolutional Neural Networks

As already mentioned, Convolutional Neural Networks (CNNs) are spreading in several tasks. One of them is, of course, classification. Many state-of-the-art CNN models for classification, *e.g.* AlexNet [53], VGG [110], GoogLeNet [111], extract hierarchically-related feature layers of decreasing scale, usually shown as a pyramidal stack whose head is the K-vector that returns the class membership scores associated with the image (as whole) to be classified. This vector is normally provided as discrete probability distribution by simply using a softmax activation layer in output, which is defined as

$$\hat{z}_i = g(\mathbf{s})_i = \frac{e^{s_i}}{\sum_{j=1}^{K} e^{s_j}} \qquad \text{for} \quad i = 1, \ldots, K, \tag{3.5}$$

being $\mathbf{s} = (s_1, \ldots, s_K)$ the unnormalized score vector entering the softmax layer and $\hat{z}_i$ the i-th class membership probability singled out.

In order to move from image-wise to pixel wise classification (the latter is also referred to as semantic segmentation), it is necessarily to provide spatially localized features toward the output layer. That is to say, now we need to know "what" and "where". A first notable attempt to do this was proposed in [56] by converting the FC stages of classification nets, such as [53, 110, 111], in convolutional ones obtaining Fully-Convolutional Networks (FCN) for semantic segmentation. Another approach is to resort to a encoder-decoder paradigm where an image classifier plays as encoder, while a coupled decoder aims to restore the spatial (classified) layout by means of upscaling layers and scale-wise skip connections. Examples of this kind are the U-Net architecture

for segmentation or the feature pyramid network (FPN) for object detection, proposed in [108] and [112], respectively.

On the other hand, depending on the target task, a suitable loss function to be optimized with a training process needs to be defined according to our expectations. Moreover, as already mentioned, the loss must be differentiable and, more in general, have a shape that speeds up the gradient descent optimization process. In this problem, which is a particular case of semantic segmentation where only two classes are considered (forest/non-forest), the output is just a single probability map resulting from a pixel-wise softmax (that reduces to a sigmoid in the binary case) activation layer. The softmax activation is typically associated to a cross-entropy loss function [113], as the gradient of their combination has good convergence properties. In the binary case, the cross-entropy loss for the i-th input-output training example $\mathbf{t} = (\mathbf{x}, \mathbf{z})$, generalized to the case of pixel-wise classification, is defined as

$$L_{\mathbf{t}}^{\mathrm{bce}} = -\frac{1}{N}\sum_{n=1}^{N}\left[z_n \log(\hat{z}_n) + (1 - z_n)\log(1 - \hat{z}_n)\right], \tag{3.6}$$

being $\mathbf{x} \in \mathbb{R}^N$ the N-pixel input image, $\mathbf{z} = \{z_n\} \in \{0, 1\}^N$ the corresponding binary ground-truth map, and $\hat{\mathbf{z}} = \{\hat{z}_n\} = f_\Phi(\mathbf{x}) \in [0, 1]^N$ the probability map predicted by the network f_Φ having parameters Φ. The target loss to be minimized over Φ is the average of the sample loss over the whole training dataset:

$$L^{\mathrm{bce}} = \mathrm{E}_{\mathbf{t}\sim\mathrm{train}}\left[L_{\mathbf{t}}^{\mathrm{bce}}\right]. \tag{3.7}$$

The cross-entropy loss works pretty well for classification tasks where the predictors is asked to take a global decision about the input image. On the contrary, when dealing with pixel-wise prediction, although it still gives a rapid loss decay, it does not necessary correspond to satisfactory results. This is primarily due to the underlying assumption of independence among predictions in different locations encoded in the loss. Infact, according to Eq. 3.6, each pixel location contributes to the loss through the sum, independently from any other pixel. For segmentation tasks this assumption is too strong as, said in simple words, neighboring pixels are likely to belong to the same segment, therefore this should be reflected in the loss. On the basis of this consideration, a more suited option is the Jaccard similarity loss [114] which is defined as

$$L_{\mathbf{t}}^{\mathrm{J}} = 1 - \frac{\sum_{n=1}^{N} z_n \hat{z}_n}{\sum_{n=1}^{N} \left[(z_n + \hat{z}_n) - z_n \hat{z}_n \right]}, \tag{3.8}$$

which is the complement of the intersection over union (IoU) defined for binary masks generalized to probability masks.

3.3 Proposed models

In light of the great success of deep learning to solve vision problems, three different CNN models to extract forest maps from TanDEM-X data and/or related products are proposed and compared. In particular, the proposed models refer to three different network topologies commonly referred to as residual network (ResNet) [106], dense network (DenseNet) [107] and U-shaped network (U-Net) [108]. ResNet models were conceived origianally to speed-up the training process by forcing convolutional modules to process in a "differential" manner thanks to skip-connections. By following a similar idea, DenseNet models also achieve fast convergence rates thanks to the "feature reuse" concept. On the other hand, the U-Net topology allows to preserve spatial details and is therefore often used for segmentation purposes. For each approach several input stacking options re considered in order to assess weather and which TanDEM-X side products can boost the network accuracy on the given task. In particular, up to four input bands were selected among the following:

- β^0, absolutely calibrated backscatter image;
- θ_i, local incidence angle;
- γ_{Tot}, interferometric coherence;
- γ_{Vol}, volumetric decorrelation.

It is also worth notice that, although CNNs are able to learn features end-to-end, the injection of suitably defined hand-crafted features can be beneficial for the network performance (see [18]) as eventually confirmed also by the experiments.

For all models the same loss is used for the minimization through the training process which is a combination of the cross-entropy (3.6) and the Jaccard (3.8) losses:

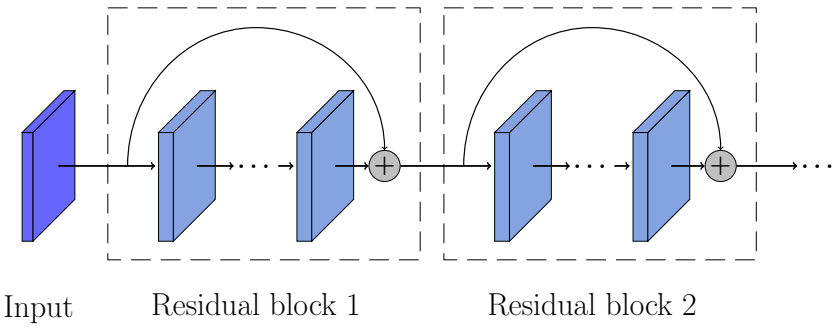


Figure 3.2: ResNet module.

$$L = \mathrm{E}_{\mathbf{t}\sim\mathrm{train}}\left[L_{\mathbf{t}}^{\mathrm{bce}} + L_{\mathbf{t}}^{\mathrm{J}}\right].$$

This choice conjugates the nice convergence properties of the cross-entropy with the good spatial characteristics of the Jaccard loss as discussed in the above section. The network output, which is in all cases a probability map, will be thresholded at 0.5 to provide the final forest map.

3.3.1 TDX-Res

A well-known bottleneck in deep learning is the computational time for training. The ResNet approach [106] is a notable solution recently proposed to mitigate this problem, which has proved also to be effective to limit overfitting. The idea is to concatenate *residual* blocks as done for example in Fig.3.2.

A residual block is nothing but a sequence of convolutional layers, inclusive of related nonlinear activations, $f(x)$, put in parallel to a identity function, or *skip connection*, yielding the overall block function

$$y = g_{\Phi}(x) = f_{\Phi}(x) + x,$$

being Φ the learnable parameters of the block. In other words, the convolutional branch works as a differential, or residual, operator, $f(x) = y - x$. In some problems, such as pansharpening, this modeling has a very nice explicit interpretation, since the desired output is already partially contained in the input, and the convolutional layers are therefore asked to just recover the missing high frequency content [115]. More in general, in [106] it has been shown that

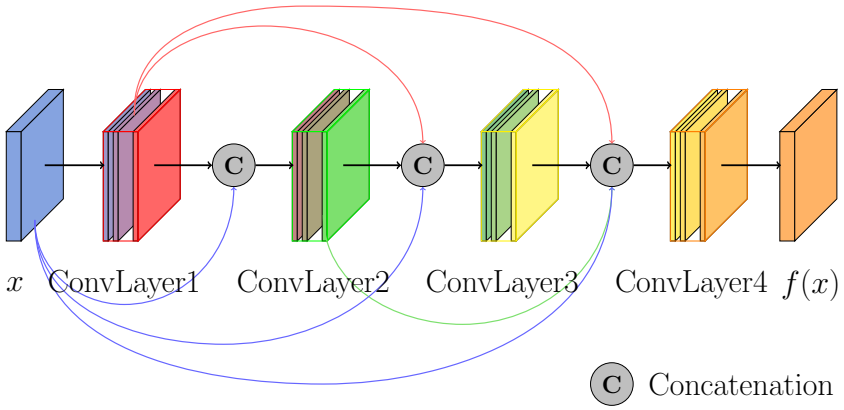


Figure 3.3: Example of DenseNet model.

by replacing unidirectional network blocks with residual schemes (just additional skip connections) consistently speeds-up the training process. This has been verified with respect to several state-of-the-art models, such as VGG-16 [110], GoogLeNet [111], BN-InceptionNet [116].

By following this rationale it has been designed a 7-layer residual network, hereinafter referred to as TDX-Res (TanDEM-X ResNet), whose hyperparameters are gathered in Tab. 3.1. All but the last layer are coupled with a ReLU activation [53] and are singularly residual. The 64 feature bands provided by the 6th layer are then transformed in a single-band, the probability map, by means of a 1×1 convolution coupled with a sigmoid activation.

3.3.2 TDX-Dense

In essence the ResNet approach creates short paths from early layers to later ones, and this is done to contrast the so-called "vanishing gradient" problem. As information about input or gradient passes through many layers, it can vanish and "wash out" by the time it reaches the end (or beginning) of the network, preventing the network from the minimization of the loss during training. On the basis of this same consideration it has been proposed also the DenseNet approach [107], that is an architecture that distills this insight into a simple connectivity pattern: to ensure maximum information flow between layers in the network, all layers (with matching feature-map sizes) are directly connected with each other. To preserve the feed-forward nature, each layer obtains additional inputs from all preceding layers and passes on its own feature-maps

to all subsequent layers. This principle is summarized in Fig. 3.3. The ℓ-th layer has ℓ inputs, consisting of the feature-maps of all preceding convolutional blocks. Its own feature are passed on to all $L - \ell$ subsequent layers. This introduces $L(L+1)/2$ connections in an L-layer network, instead of just L, as in traditional architectures. A key distinguishing characteristic of the DenseNet approach with respect to ResNet is that features are never combined through summation, as they are simply concatenated. Said k the number of feature maps produced by each layer, it is easy to verify that the l-th layer has $b + k(\ell - 1)$ input feature maps, being b the number of channels in the input layer. For this reason the hyperparameter k is referred to as *growth rate* of the network.

The proposed DenseNet model for forest segmentation over TanDEM-X data, named TDX-Dense, is a relatively shallow architecture with only 7 layers, and a growth rate of 64. In Tab. 3.2 are summarized the main hyperparameters of the network. Both TDX-Res and TDX-Dense use batch normalization on the input layer to regularize the network behaviour [116].

3.3.3 TDX-U

The third model for segmentation belongs to the U-Net family originally proposed for medical images [108]. The original idea is to use an encoder-decoder paradigm in order to inheritate well-established CNN classification models pre-trained on huge datasets, *e.g.*, ImageNet, which can play the role of encoder. As the head of the feature pyramid generated by the encoder which summarizes the image content does not carry spatial information, progressively lost flowing through convolutional (spreading) and pooling (subsampling) layers, a "mirror" decoding section is properly linked to the encoder in order to recover the image spatial layout enriched with the class information, that is the semantic segmentation (see Fig. 3.4). Symmetrically disposed with respect to pooling layers are upsampling layers. Moreover, in addition to the main feature path (U-shaped trajectory), information flows through scale-wise shortpaths that brings encoder features directly to the corresponding decoding stages, working at the same resolution, where they are concatenated with the mainstream features coming from the lower levels.

While the encoder can be any imported pre-trained net, fine-tuned if needed, the decoder is typically trained from scratch. In this case, due to the very specific characteristics of the input images, it is decided to train from scratch the whole proposed network on the given dataset avoiding any transfer learning. Fig. 3.4 refers to this specific implementation, referred to as TDX-U

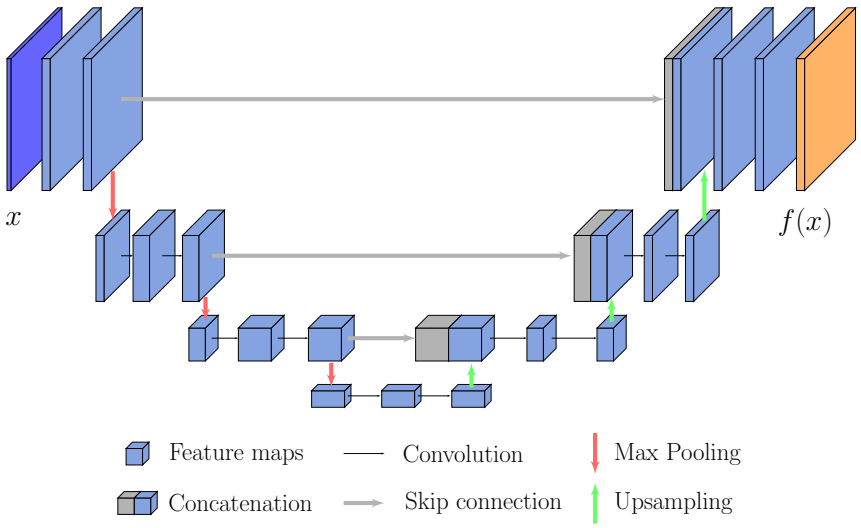


Figure 3.4: Proposed U-Net structure for forest segmentation from TanDEM-X data.

for short in the following, which works on four scale levels. At each level two chained convolutional layers are located on both the encoder and the decoder sides, with exception of the network head where an additional 1×1 convolution is used to map 64 features in a single probability channel. Additional information about network hyperparameters are summarized in Tab. 3.3.

3.4 Experimental results

In the following are discussed the experimental results obtained with the proposed CNN models in comparison with some reference solutions. Firstly the dataset and training details are described in Sec 3.4.1, then a summary of involved methods and accuracy measures in Sec. 3.4.2 is presented. Finally, the numerical accuracy assessment of the compared methods in Sec. 3.4.3 is shown, and a subjective comparison through the visual inspection of some sample results in Sec. 3.4.4.

Layer	Kernel shape
Batch Normalization	-
Conv + ReLU	$64\times b\times(3\times3)$
Conv + ReLU	$64\times64\times(3\times3)$
Conv + ReLU	$64\times64\times(3\times3)$
Conv + ReLU	$64\times64\times(3\times3)$
Conv + ReLU	$64\times64\times(3\times3)$
Conv + ReLU	$64\times64\times(3\times3)$
Conv + Sigmoid	$1\times64\times(1\times1)$

Table 3.1: TDX-Res hyper-parameters.
Shape: #kernels $\times$ #input channels $\times$ kernel support.

Layer	Kernel shape
Batch Normalization	-
Conv + ReLU	$64\times b\times(3\times3)$
Conv + ReLU	$64\times(64+b)\times(3\times3)$
Conv + ReLU	$64\times(128+b)\times(3\times3)$
Conv + ReLU	$64\times(192+b)\times(3\times3)$
Conv + ReLU	$64\times(256+b)\times(3\times3)$
Conv + ReLU	$64\times(320+b)\times(3\times3)$
Conv + Sigmoid	$1\times64\times(1\times1)$

Table 3.2: TDX-Dense hyper-parameters.
Shape: #kernels $\times$ #input channels $\times$ kernel support.

Layer	Kernel shape
Batch Normalization	-
Conv + ReLU	$64\times b\times(3\times3)$
Conv + ReLU	$64\times64\times(3\times3)$
2×2 Max Pooling	-
Conv + ReLU	$128\times64\times(3\times3)$
Conv + ReLU	$128\times128\times(3\times3)$
2×2 Max Pooling	-
Conv + ReLU	$256\times128\times(3\times3)$
Conv + ReLU	$256\times256\times(3\times3)$
2×2 Max Pooling	-
Conv + ReLU	$512\times256\times(3\times3)$
Conv + ReLU	$512\times512\times(3\times3)$
2×2 Upsampling	-
Conv + ReLU	$256\times512\times(3\times3)$
Conv + ReLU	$256\times256\times(3\times3)$
2×2 Upsampling	-
Conv + ReLU	$128\times256\times(3\times3)$
Conv + ReLU	$128\times128\times(3\times3)$
2×2 Upsampling	-
Conv + ReLU	$64\times128\times(3\times3)$
Conv + ReLU	$64\times64\times(3\times3)$
Conv + Sigmoid	$1\times64\times(1\times1)$

Table 3.3: TDX-U hyper-parameters.
Shape: #kernels $\times$ #input channels $\times$ kernel support.

3.4.1 The Pennsylvania Data Set and training details

The dataset of bistatic TanDEM-X images used in the following was acquired over the state of Pennsylvania, USA, during the first year of the mission and belongs to the global dataset of nominal acquisitions used for the generation of the TanDEM-X DEM. It consists of ten image tiles of about 9200×6700 pixels on average, nine of which are used for training or validation, while the remaining one is reserved for tests. Training and validation sets are created as follows. 18.000 randomly chosen 128×128 patches are grouped in 32-dimensional training mini-batches. 2.000 more patches with the same size were used for validation instead. The training was carried out running the Adam optimizer [76] with an initial learning rate of 10^{-4} for 20 epochs. Moreover, the test set composed of five 1400×1800 samples extracted from the additional tile was reserved for the accuracy assessment of the compared solutions. The five samples were selected with different characteristics in terms of class content (*e.g.*, forest, water, urban, bare soil,...) for a more comprehensive evaluation of the generalization properties of the method.

The region of interest is largely covered by temperate forests (about 60%), mainly characterized by the presence of deciduous trees and birch. The remaining lightly vegetated areas can be associated to shrubs, bushes, and wildflowers. Moreover, Pennsylvania is characterized by the presence of a dominant southwest-to-northeast oriented barrier ridge of high-relief terrain, namely the Appalachian Mountains. The reason for the choice of such an area of interest is the availability of a high-resolution reference forest/non-forest map, derived from lidar and optical data [117]. This data set was generated by a joint collaboration between the University of Maryland and the University of Vermont, and released later in 2015. Input data, acquired between 2006 and 2008, were combined together to generate a forest/non-forest binary layer for vegetation higher than 2 m and with ground resolution of 1×1 m^2.

3.4.2 Methods and metrics

The proposed solutions described in detail in Sec. 3.3 are cast in three groups, TDX-Res, TDX-Dense, and TDX-U, corresponding to three state-of-the-art CNN building approaches, ResNet, DenseNet, and U-Net, respectively, particularized to the problem of forest mapping from TanDEM-X data. In order to show the discriminative power of domain-raleted SAR features, different input configurations are tested for each model category, by selecting up to 4 input channels selected among SAR amplitude β^0, incidence angle θ_i, interfer-

ometric coherence γ_{Tot}, and volumetric decorrelation γ_{Vol}. All networks were trained from scratch.

As reference solutions for a comparative evaluation, in addition to the two versions of the baseline method [104] briefly described in Sec. 3.2.1, namely Baseline and Baseline+, Random Forest classifiers with different input configurations as well as done for the proposed methods are taken in account. The learning procedure of each Random Forest classifier has been conducted by the means of 10^7 samples, in order to have almost the same size of the dataset used for the training phase of the Deep Learning models. The accuracy evaluation was based on classical and widespread measures used for detection such as the true/false positives/negatives rates:

[TP] True positives: rate of pixels correctly classified as forest.

[TN] True negatives: rate of pixels correctly classified as non forest.

[FP] False positives: rate of pixels wrongly classified as forest.

[FN] False negatives: rate of pixels wrongly classified as non forest.

Based on these measurements several indicators can be computed to simplify the interpretation of the assessed methods. In particular, Precision, Recall, F_1-score and Accuracy are provided, which are defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3.9}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3.10}$$

$$F_1 = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{3.11}$$

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3.12}$$

Precision and Recall are usually shown together as they represent the trade-off between the need of catching the target class whenever it occurs ($\mathrm{FN} = 0$, $\mathrm{Recall} = 1$) and that of reducing false alarms ($\mathrm{FP} = 0$, $\mathrm{Precision} = 1$). Ideally, both these measures should be maximized to unity. A compact representation of both is given by their harmonic average, the F_1-score. In case one only cares about the global rate of correctly classified (as forest or non forest, in our case) pixels this is provided by the last indicator (Accuracy).

Model	β^0	θ_i	γ_{Tot}	γ_{Vol}	Recall	Prec.	F_1-score	Acc.
TDX-Res	×				80.09%	61.79%	69.76%	69.67%
TDX-Res	×		×		84.44%	79.97%	82.15%	83.96%
TDX-Res		×	×		78.69%	75.32%	76.97%	79.43%
TDX-Res		×		×	91.29%	64.32%	75.47%	74.08%
TDX-Res	×	×			**84.06%**	65.67%	73.74%	73.84%
TDX-Dense	×	×			79.12%	67.87%	73.06%	74.51%
TDX-U	×	×			80.48%	**84.57%**	**82.48%**	**85.06%**
TDX-Res [93]	×	×	×		**85.04%**	81.38%	83.17%	84.97%
TDX-Dense [93]	×	×	×		83.99%	83.76%	83.88%	85.89%
TDX-U	×	×	×		84.46%	**88.19%**	**86.29%**	**88.27%**
TDX-Res	×	×		×	86.97%	79.42%	83.02%	84.46%
TDX-Dense	×	×		×	**91.94%**	74.10%	82.06%	82.44%
TDX-U	×	×		×	82.94%	**89.25%**	**85.98%**	**88.18%**
TDX-Res	×	×	×	×	**89.80%**	75.59%	82.08%	82.88%
TDX-Dense	×	×	×	×	86.40%	80.83%	83.52%	85.11%
TDX-U	×	×	×	×	80.98%	**90.97%**	**85.68%**	**88.18%**

Table 3.4: Forest detection accuracy assessment on the test dataset for different proposed CNN models. Input bands are marked in columns 2-5.

3.4.3 Numerical assessment

For ease of presentation the numerical results are grouped in two tables. The former (Tab. 3.4) gathers a meaningful set of proposed CNN models and is useful to understand several design choices. The latter (Tab. 3.5) compares some selected proposed models with reference methods. An overview of the performances of all compared methods is then depicted on the Precision-Recall plane of Fig. 3.5.

In Tab. 3.4, the models are grouped depending on the input layer setting which is specified in columns 2-5. A more extensive evaluation with respect to the input configuration is presented for the TDX-Res case only, without loss of generality.

It can be observed that each input band brings its own contribution to improve the CNN discrimination capability, with the exception of the pair (γ_{Tot}, γ_{Vol}) that seems highly correlated according to the numerical results. Infact,

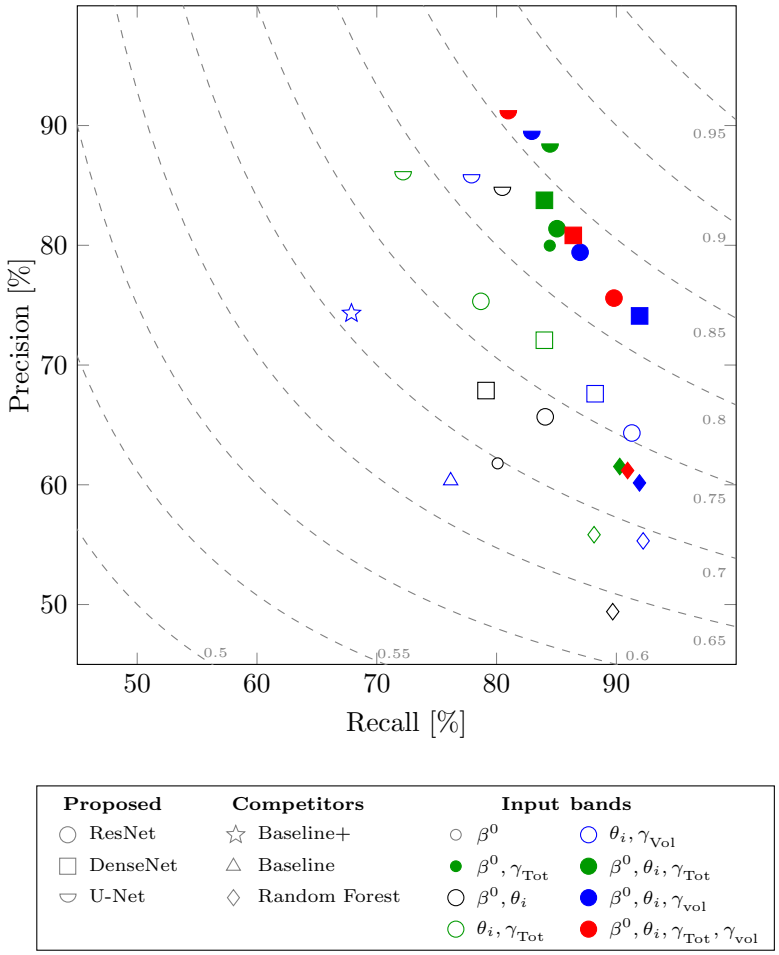


Figure 3.5: Precision-Recall comparison. Dashed lines show F_1-score level curves.

Method	β^0	θ_i	γ_{Tot}	γ_{Vol}	Recall	Prec.	F_1-score	Acc.
Baseline [104]		×		×	76.17%	60.34%	67.34%	67.72%
Baseline+ [104]		×		×	68.23%	74.32%	71.14%	75.82%
Random forest		×		×	**92.24%**	55.32%	69.16%	64.06%
TDX-U		×		×	77.91%	**85.62%**	**81.58%**	**84.63%**
Random forest	×	×			**89.70%**	49.40%	63.71%	55.37%
TDX-U	×	×			80.48%	**84.57%**	**82.48%**	**85.06%**
Random forest	×	×	×		**90.28%**	61.53%	73.18%	71.09%
TDX-U	×	×	×		84.46%	**88.19%**	**86.29%**	**88.27%**
Random forest	×	×		×	**91.93%**	60.16%	72.72%	69.88%
TDX-U	×	×		×	82.94%	**89.25%**	**85.98%**	**88.18%**
Random forest	×	×	×	×	**90.94%**	61.19%	73.16%	70.85%
TDX-U	×	×	×	×	80.98%	**90.97%**	**85.68%**	**88.18%**

Table 3.5: Numerical comparison with reference methods.

Accuracy moves from 69.67% using just β^0 to 73.84% including θ_i, jumping over the 80% barrier including one or two more input channels. Besides, the simultaneous inclusion of γ_{Tot} and γ_{Vol} can be even slightly detrimental for accuracy. Indeed they are rarely used simultaneously in the literature because of their direct relationship (Eq.3.3). In this case, for all CNN architectures these two parameters look nearly equivalent and the best option is to use just one of them together with β^0 and θ_i, if available. Comparing the different architectural options it results that both TDX-Res and TDX-Dense overestimate the forest class (maximize Recall) while TDX-U is more conservative maximizing the Precision metric. However, the latter clearly provides the best trade-off between Precision and Recall, performing consistently better than the formers in terms of both F_1-score and Accuracy. The above considerations can be easily recognized observing the Precision-Recall plane in Fig. 3.5. The incidence angle θ_i, in fact, is particularly effective when used in conjunction with the SAR signal β^0 (see the gain between small to large black circles, associated to TDX-Res), but it also boosts the accuracy when other features such as γ_{Tot} are enclosed (small green circle *vs* filled green circle). The above considerations about Precision-Recall trade-offs can also be immediately verified on the same scatter plot in Fig. 3.5, with TDX-U located on the upper-triangular image section contrarily to TDX-Res/Dense that lie on the lower-triangular part.

In Tab. 3.5 the best architecture, TDX-U, are compared with the refer-

ence methods, differentiating the analysis with respect to the input configuration. Since the baseline methods apply to the pair $(\theta_i, \gamma_{\rm Vol})$ of TanDEM-X by-products, for a fair comparison, a CNN model is trained with this input configuration. The proposed network outperforms the baseline methods with a large margin using the same input, with a further gain including other input channels. For other input settings random forest classifiers are considered, registering large gains in this case, as well.

3.4.4 Visual comparison

Besides numerical evaluation it is worth to analyse some sample results by visual inspection. Therefore, in the following the described solution is compared with the baseline methods, assuming the same input configuration which is the pair $(\theta_i, \gamma_{\rm Vol})$. Three samples representative of different contexts are shown in Fig. 3.6. The forested areas are highlighted with a green mask that overlays the backscattered SAR signal β^0. The first column shows the ground-truth, then the Baseline and Baseline+ map predictions are in the middle columns, and the best model using the same input is in the last column. Observe preliminarly that Baseline predicts as forest also water (middle sample) and built-up areas (top and bottom samples). As already underlined above $\gamma_{\rm Vol}$ does not allow to discriminate forest from these two classes, and for this reason in [104] was also proposed the masked version of Baseline, that is Baseline+ here, whose predictions are shown in the third column. The proposed solution is clearly consistent with the ground-truth and with Baseline+, although it does not make use of any external mask. In consideration of the low separability between forest and built-up or water from $\gamma_{\rm Vol}$ this is a quite surprising achievement. The comparison with the random forest solutions does not add much information to what has been already seen with the numerical results because of the large numerical gap registered also with other input configuration.

In the following, focusing for simplicity on TDX-U without loss of generality, three input settings for the proposed model on the previous running samples are compared in fig. 3.7 All three configurations include the incidence angle channel θ_i, which is concatenated with $\gamma_{\rm Vol}$, β^0, or both and are shown in the second, third and fourth columns, respectively. The combined use of SAR amplitude and volumetric decorrelation provides uniformly better results. Similar results are obtained if $\gamma_{\rm Vol}$ is replaced with $\gamma_{\rm Tot}$, or simply add the latter to the input. In general, the use of β^0 seem to improve the accuracy on fine details, likely because of the coarser resolution of the other input features (see Fig. 3.1).

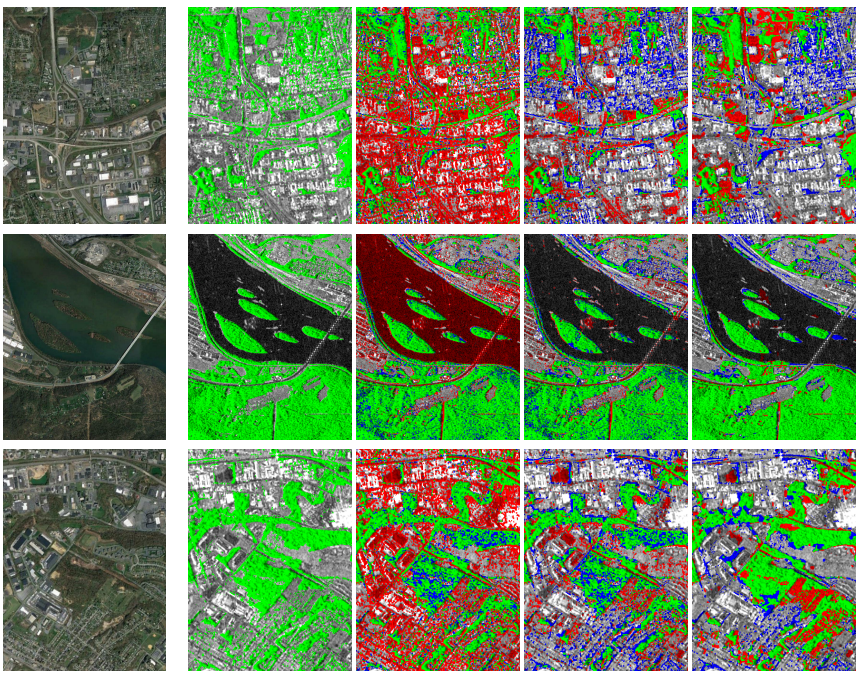


Figure 3.6: Forest mapping comparison among Baseline, Baseline+ and TDX-U, using $(\theta_i, \gamma_{\text{Vol}})$ in input. Correctly classified forest pixels (TP) are shown in green; non-forest pixels erroneously classified as forest (FP) are in red; the blue indicates missed forest pixels (FN).

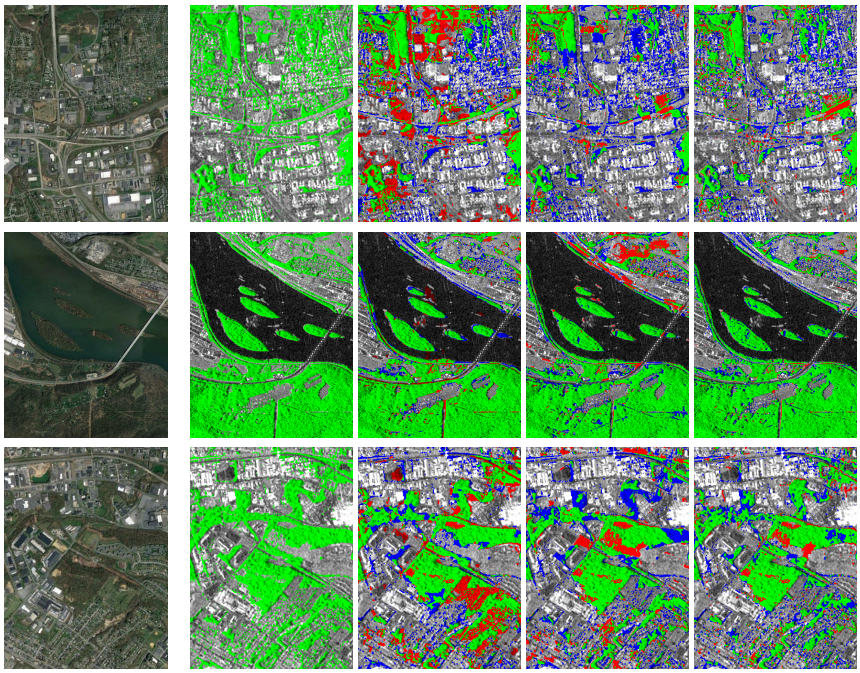


Figure 3.7: Detection results provided by TDX-U using different input settings. Correctly classified forest pixels (TP) are shown in green; non-forest pixels erroneously classified as forest (FP) are in red; the blue indicates missed forest pixels (FN).

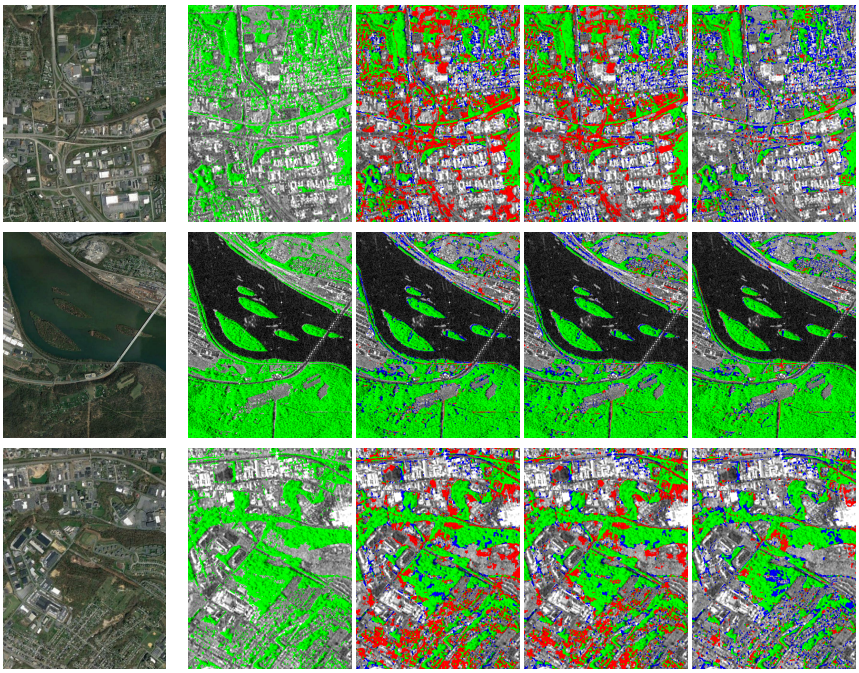


Figure 3.8: Segmentation results provided by TDX-Res, TDX-Dense and TDX-U under the best input configuration: $(\beta^0, \theta_i, \gamma_{\text{Tot}})$.Correctly classified forest pixels (TP) are shown in green; non-forest pixels erroneously classified as forest (FP) are in red; the blue indicates missed forest pixels (FN).

Finally, in Fig. 3.8 the three proposed models are compared in the best input configuration according to the numerical results of Tab. 3.4, that is $(\beta^0, \theta_i, \gamma_{\text{Tot}})$. The U-Net apporach clearly outperforms the other two in all cases, coherently with the numerical results reported in Tab. 3.4.

3.5 Conclusions

In this chapter is explored the use of Convolutional Neural Networks for the purpose of forest mapping from TanDEM-X products. Regardless of the employed CNN building strategy, results demonstrate that CNNs can effectively

fuse input data with heterogeneous dynamics, such as the SAR backscatter, the interferometric coherence, and the incidence angle. This is likely the most distinguishing feature of the CNN approach compared to traditional methods [103, 104], which only extract one key feature from the InSAR signal to exclusively exploit it for the classification. CNN have shown better performance with respect to the pixel-wise Random Forest algorithm too. This is probably caused by the ability of CNNs to account for the textural content of the signal that provides additional discriminative information. Among the considered architectural solutions, U-Net clearly provides the best performance in terms of accuracy and F1-score. This study open new research scenarios showing how to effectively extract information of interest from data acquired by means of a mission designed mainly for Digital Elevation Model retrieval.

A future research will aim to extend the proposed approach to diverse scenarios. In order to do this a key enabling factor will be the collection of a wider and richer dataset for training, validation and test, that is representative of much diverse climate conditions (boreal, temperate, tropical, and so forth) and anthropological contexts (rural, industrial, urban, and so on). The use of additional input features from TanDEM-X or other information sources will also be explored as they may compensate to some extent the lack of referenced data. In addition to the diversity with respect to climate and cover types, it will be also worth to explore to what extent the proposed approach can generalize in resolution, in order to enable wide scale applicability of the method using lower resolution data.

Chapter 4

Super Resolution of Sentinel-2 bands

In light of its free availability, world-wide coverage, revisit frequency and, not least, its above remarked wide applicability, several research teams have proposed solutions to super-resolve Sentinel-2 images, rising 20 m and/or 60 m bands up to 10 m resolution. In particular, the 10 and 20 m 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), while lower resolution 60 m bands can be used for monitoring of water vapor, aerosol corrections, pollution monitoring, cirrus clouds estimation and so forth [118, 119]. Specifically, beyond land-cover classification, S2 images can be useful in such diverse applications as the prediction of growing stock volume in forest ecosystems [120], the estimation of the Leaf Area Index (LAI) [121, 122], the retrieval of canopy chlorophyll content [123], the mapping of the extent of glaciers [5], the water quality monitoring [124], the classification of crop or tree species [125], and the built-up areas detection [126], fire detection [127], urban mapping [128], and vegetation monitoring [129].

According to the taxonomy suggested by Lanaras et al. [119] 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 these category, it is also worth mentioning the matrix factorization approaches (e.g., [130, 131]), 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 by means of unmixing oriented methodologies [96, 132]. 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 [133, 134], the multi-resolution analysis approaches [135, 136], or other energy minimization methods [137, 138, 139] belong to this category. A recent survey on pansharpening can be found in [140]. 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 [36, 141, 142, 143, 144].

The second group refers to methods that face the super-resolution as an inverse problem under the hypothesis of known imaging model. The ill-posedness is therefore addressed by means of additional regularization constraints encoded in a Bayesian or a variational framework. Brodu's super-resolution method [145] 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. [146] adopted an observation model with per-band point spread functions that accounts for convolutional blur, downsampling, and noise. The regularization consists of two parts, a dimensionality reduction that implies 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. [147] 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 approaches, and notably deep learning (DL) ones, 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 modeling (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 [96, 132], since in both cases the general aim is to derive the different spectral responses covered by

a single, spatially coarse observation. However, more specifically, 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 the retrieval of the corresponding spectral responses without the need to improve their spatial localization. In pansharpening, the focus is mainly on spatial resolution enhancement while preserving at most the spectral properties of the sources, and 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 [135, 136], instead of using radiative transfer models associated to 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 first notable example of DL applied to the super-resolution of remote sensing images is the pansharpening convolutional neural network (PNN) proposed by Masi et al. [18], which has been recently upgraded [115] with the introduction of a residual learning block and a fine-tuning stage for target adaptivity and cross-sensor usage. Another residual network for pansharpening (PanNet) is proposed in [148]. However, none of these methods can be applied to S2 images without some architectural network adaptation and retraining. Examples of convolutional networks conceived for Sentinel-2 are instead proposed in [119]. In Lanaras et al. [119], a very large training dataset has been collected 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 vision tasks such as object recognition or instance segmentation,thanks to the knowledge hidden in huge and shared training databases, deep learning has shown really impressive results compared to model-based approaches. Data sharing has represented a key enabling factor in these cases allowing researchers to compete with each other or reproduce others' models. In this chapter two solution to the super-resolution of Sentinel-2 bands are described. In particular, in Section4.1 , the 20 m SWIR band of Sentinel-2 is super-resolved at 10 m for the specific application of the water mapping [149]; while in Section 4.2 a general solution to super-resolve all 20 m bands is presented [55].

4.1 Super-Resolution on the SWIR band of Sentinel-2

Sentinel-2 satellites provide global acquisitions of relatively fine spatial resolution multispectral images with a high revisit frequency, whose objective is to supply data for services such as risk management (floods, forest fires, subsidence, landslide), land monitoring, food security/early warning systems, water management, soil protection and so forth [118]. Unfortunately, due to a balance between technological constraints and the objectives of the mission, only four out of thirteen bandsare provided at the highest resolution of 10 meters. The remaining bands are given at 20 or 60 meters. One such bands is for example the SWIR, provided at 20 meters, which has proven to be very useful for water detection [150].

Motivated by the above considerations, in the following it is described a super-resolution method for the SWIR band of Sentinel-2, for the purpose of water monitoring at fine-scale through the Modified Normalized Difference Water Index (MNDWI). The basic approach to address a super-resolution problem is to use a bicubic or a more general polynomial interpolation. State-of-the-art solutions resort instead to the use of deep learning methods such as CNNs [54, 151]. However, these methods do not take into account any other source of data but the objective image. In the problem at hand are available companion bands which are coregistered with the target and carry important information, in particular spatial details. Therefore it would be more effective to resort to pansharpening-like methods meant to fuse a low-resolution multispectral (MS) image with a high-resolution single panchromatic (PAN) band, to rise the resolution of the MS to that of the PAN. In our case the SWIR band would play the role of the MS while one or more higher resolution companion bands would replace the single PAN. By following this paradigm several pansharpening methods, based on both component substitution [133, 134] and multiresolution analysis [135, 136], were adapted to the Sentinel-2/SWIR case and compared in [36]. Somehow related to this case is the fusion of multispectral and hyperspectral images for which a deep learning approach has been already proposed in [19].

Here, according with this line of research, it is described a CNN-based approach similar to [18, 34] which have proved to be very successful to pansharpen very high resolution data like Ikonos, GeoEye, or WorldView. In particular, three CNN models corresponding to three input combinations are presented. In the simplest case (M1) the network is feeded only with the objective band ρ_{11}, without high-resolution guiding bands (pure super-resolution). Then also higher resolution bands are included ,moving to the pansharpening-like

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

Figure 4.1: Example of super-resolution of ρ_{11} (SWIR band).

case, considering the limit cases when only the most correlated band (near infra-red, NIR) is enclosed (M2) opposed to the case when all 10-m resolution bands are used (M5).

A preview of the proposed solution compared with a simple bicubic interpolation is given in Fig. 4.1. On the left is the "guide" band ρ_8, in the middle is the bicubic interpolation of ρ_{11}, and on the right is the proposed upsampling with model M2. Numerical results discussed further show that the quality of the method compares favourably against different pansharpening-like alternatives according to several indicators. In addition, the proposal is also tested from the user's perpective by detecting water basins through the MNDWI computed at 10m resolution using the upsampled SWIR component.

The rest of the Section is organized as follows. Section 4.1.1 describes the proposed method in more detail, while Section 4.1.3 summarizes experimental results.

4.1.1 Proposed CNN-based method

Convolutional neural networks have been successfully applied to many image processing problems, like super-resolution [54] and pansharpening [18], 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. On the downside the availability of a large amount of "labelled" data is required for training.

In the following is used a relatively shallow architecture which is a cascade of $L = 3$ convolutional layers interleaved by Rectified Linear Unit (ReLU) activations that ensure fast convergence of the training process [53].

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}, ρ_8	3×3 (48)	3×3 (32)	3×3 (1), $\widehat{\rho}_{11}$	7×7
M5	ρ_{11}, ρ_8 ρ_2, ρ_3, ρ_4	3×3 (48)	3×3 (32)	3×3 (1), $\widehat{\rho}_{11}$	7×7

Table 4.1: Hyper-parameters of the proposed networks.

Let $\mathbf{x} \triangleq (\rho_{11}, \rho^{\mathrm{HR}_1}, \ldots, \rho^{\mathrm{HR}_B})$ be the input to the network[1], and $\mathbf{y} \triangleq \widehat{\rho}_{11}$ be the network output that is the sharpened SWIR band. The network hyper-parameters are summarized in Tab.4.1. 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 guide, while M5 uses all available high-resolution bands. In the last column it is reported the scope of the overall network function readily obtained as comulative convolutional spread, as the non-linear ReLU is a punctual operator which does not increase the scope. These hyper-parameters were selected among several alternative configurations as optimal choice in terms of complexity and accuracy. It is worth notice that the overall scope is relatively small compared to that of the CNN pansharpening method [18], which is 17×17. This should not surprise as [18] is conceived for a super-resolution ratio which is double, therefore requiring in principle major efforts to work "equally" well.

4.1.2 Learning

In order to train the network's parameters Φ a sufficiently large number of input-output examples and the choice of a suitable cost function to minimize on them are required to run any learning algorithm, like for example the Stochastic Gradient Descent (SGD) adopted in [34]. In the specific case of pansharpening, in [18] it has been proposed to generate examples for training through

[1] In CNN-based super-resolution or pansharpening the most commonly used solution consists of a preliminarly upsample of the lower resolution components in input with a standard ideal interpolator, *e.g.* bicubic, to align the input stack. In the following, the notation ρ_{11} refers to this interpolated band which actually feeds the net.

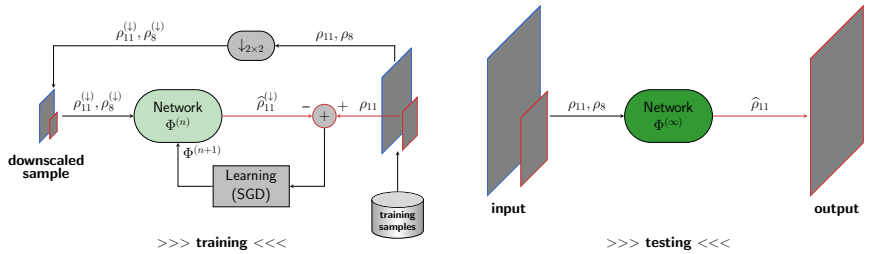


Figure 4.2: Top-level training (left) and testing (right) workflows for model M2.

Wald's protocol, that consists in using as inputs properly downsampled PAN-MS pairs and taking as corresponding output the original MS. This same approach has been adapted to the problem at hand in this work, and it is graphically represented at the top level in Fig. 4.2 (left). The resolution 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. 4.2). The training phase is carried out offline once for all and takes a few hours using GPU cards, while the test can be done in real-time. Moreover, it is preferred to use the L1-norm in place of the L2-norm as it has proven [34] to be more effective in the error backpropagation. Specifically, the loss is computed by averaging over a suitable set (mini batch) of training examples at each updating step of the SGD process:

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

4.1.3 Experimental Results

In order to build a sufficiently general dataset for training three different scenes have been chosen: Guinea, Tunisia, and Italy (Venice). Once left apart some 450×450 clips for testing, 17×17 patches for training were uniformly sampled from all scenes in the remaining segments. Overall, 19000 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).

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 [136]	0.9873	5.949	0.5828	0.0160	0.1762
MTF-GLP-HPM [155]	0.9823	7.245	0.4509	0.0207	0.1370
HPF [135]	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**

Table 4.2: Accuracy of $\widehat{\rho}_{11}$ (Q-index, ERGAS, HCC) and water maps (CER, L-CER).

To assess the performance of the proposed method three full-reference numerical figures commonly used for pansharpening have been evaluated:

- Q-index, an image quality indicator introduced in [152];
- ERGAS, proposed in [153], which reduces to the root mean square error in case of single band;
- HCC, the correlation coefficient between the high-pass components of reference and its estimate [154].

As these indicators require reference, likewise for the training data, reduced resolution test data are produced through Wald's protocol.

The average numerical results obtained for the three scenes of interest are gathered in Tab. 4.2 (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 stress that much the gain provided by M1 as it does HCC which deals with high frequency components that are much affected by the super-resolution and are mostly localized on boundaries. Moving to the methods that make use of the additional band ρ_8 as guide, the proposed M2 compares favourably against state-of-the-art pansharpening methods adapted to the Sentinel-2/SWIR problem as suggested in [36]. 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}, provide an additional gain.

The proposed models are also tested from the application point of view by detecting water basins through the computation of the MNDWI index defined

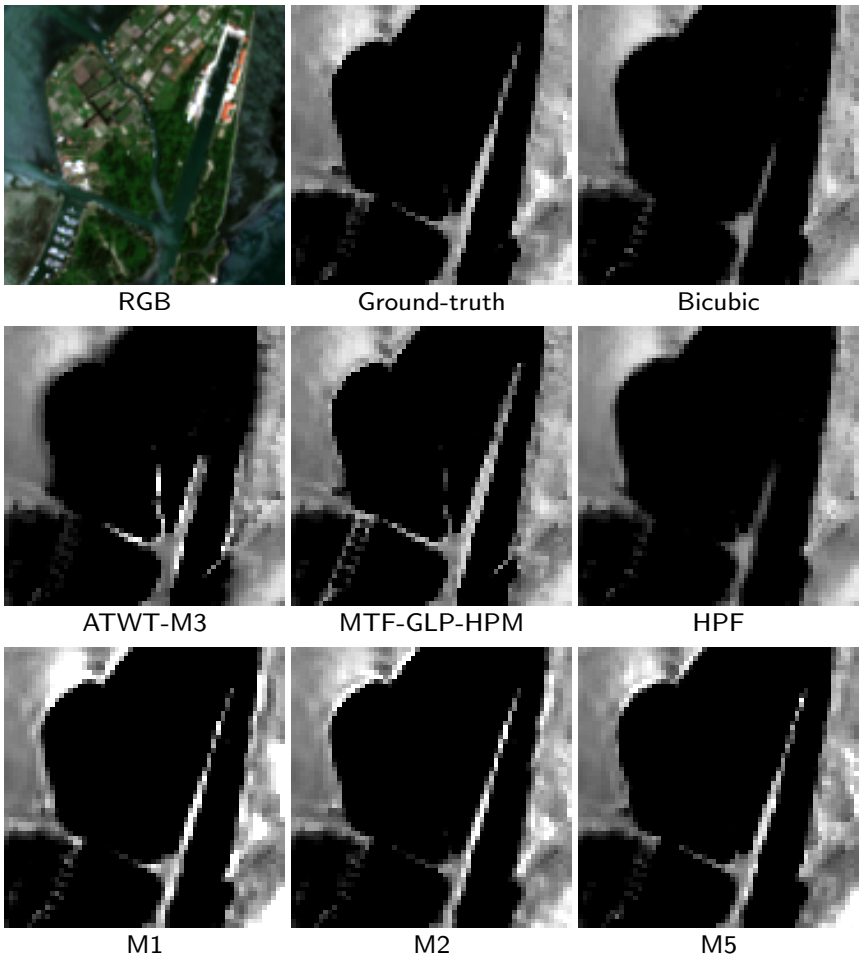


Figure 4.3: MNDWI estimations over a sample detail (from Venice image). In order to have a reference ground-truth Wald's protocol is applied (downgraded resolution).

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 20m or 10m, respectively. Once water (W) is detected by suitably thresholding the MNDWI[2] ($W = I > \alpha$), the classification error rate on the whole image (CER) and locally to boundaries[3] (L-CER) is computed and reported on the right-hand side of Tab. 4.2. These figures provide a further confirm of the superiority of the proposed method.

To conclude this section some sample results of estimated MNDWI are shown (Fig. 4.3) as additional mean to judge the overall behaviour of the proposed approach, leaving the reader free to complete his/her own analysis of the results with any other observation it may concern.

4.2 Super resolution on Sentinel-2 bands

Motivated by the good results obtained in the previous Section for one 20 m band, in the following is described a CNN-based method to provide a fast, upscalable method for the single-sensor fusion of Sentinel-2 (S2) data, whose aim is to provide a 10 m super-resolution of all of the original 20 m bands. Aiming to obtain 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.

In Lanaras et al. [119] a relatively large Sentinel-2 dataset to get good generalization properties is collected. On the other hand, complexity is also an issue that end users care about. In the following is described a solution that employs 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 described 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. The described method for Fast Upscaling of SEntinel-2 (FUSE) images is an evolution of the *proof-of-concept* work presented in [55]. In particular, the major improvements with respect to the method in [55] reside in the following changes:

[2] Hereinafter it has been fixed $\alpha = 0$.

[3] Boundaries are detected using morphological gradient.

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

b. The definition 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 Section is organized as follows. In Section 4.2.1, follows a description of the datasets and proposed method. Evaluation metrics, comparative solutions and experimental results are then gathered in Section 4.2.5. Insights about the performance of the proposed solution and related future perspectives are given in Section 4.2.9. Finally, Section 4.2.10 provides concluding remarks.

4.2.1 Materials and Methods

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;

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

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

4.2.2 Datasets and Labels Generation

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

For the sake of clarity, the main characteristics of the 13 spectral bands of Sentinel-2 are gathered in Table 4.3, while symbols and notations that are used in the following are grouped in Table 4.4.

Except for some cases where unsupervised learning strategies can be applied, a sufficiently large dataset containing input–output examples is usually

Table 4.3: 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

Table 4.4: Notations and symbols.

Symbol	**Meaning**
$\mathbf{x}$	Stack of six S2 spectral bands (B5, B6, B7, Bba, 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 $\mathbf{x}$.
$\mathbf{r}$	Full-resolution *reference* (also referred to as *ground truth* or *label*), usually unavailable.
$x, \hat{x}, r$	generic band of $\mathbf{x}, \hat{\mathbf{x}}, \mathbf{r}$, 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 $\mathbf{x}$, $\mathbf{z}$, $\mathbf{r}$,..., respectively. Whenever unambiguous subscript $_{\downarrow}$ will be dropped.

necessary to to train a deep learning model. This is also the case for super-resolution or pansharpening. In this case, since the goal is to fuse 10 m bands ($\mathbf{z}$) with 20 m ($\mathbf{x}$) to enhance the resolution of $\mathbf{x}$ by a factor of 2 (resolution ratio), which means that the examples of the kind $((\mathbf{x}, \mathbf{z}); \mathbf{r})$ are needed, being $\mathbf{r}$ 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 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., [54, 18, 151, 115, 149, 119, 156]) by means of 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 which 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

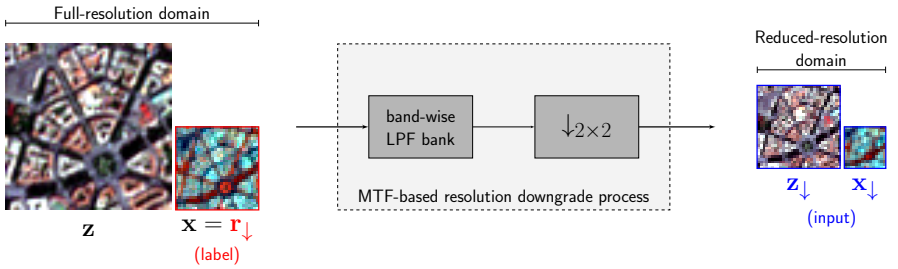


Figure 4.4: Generation of a training sample $((\mathbf{x}_{\downarrow}, \mathbf{z}_{\downarrow}); \mathbf{r}_{\downarrow})$ using Wald's protocol. All images are shown in false-color RGB using subsets of bands for ease of presentation. Each band is low-lass filtered with a different cut-off frequency according with the sensor MTF characteristics.

to a *reduced*-resolution domain. In this specific 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 several bands before subsampling is an open question. Lanaras et al. [119] 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. [18] asserted that the resolution scaling should be done accounting for the sensor Modulation Transfer Function (MTF), in order to generalize properly 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 [140]. 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 the following the approach proposed in [18] making use of sensor MTF is adopted. The process for the generation of a training sample is summarized in Figure 4.4. Each band undergoes a different low-pass filtering, prior to being downsampled, whose cut-off frequency is related to the sensor MTF characteristics. Additional details can be found in [157].

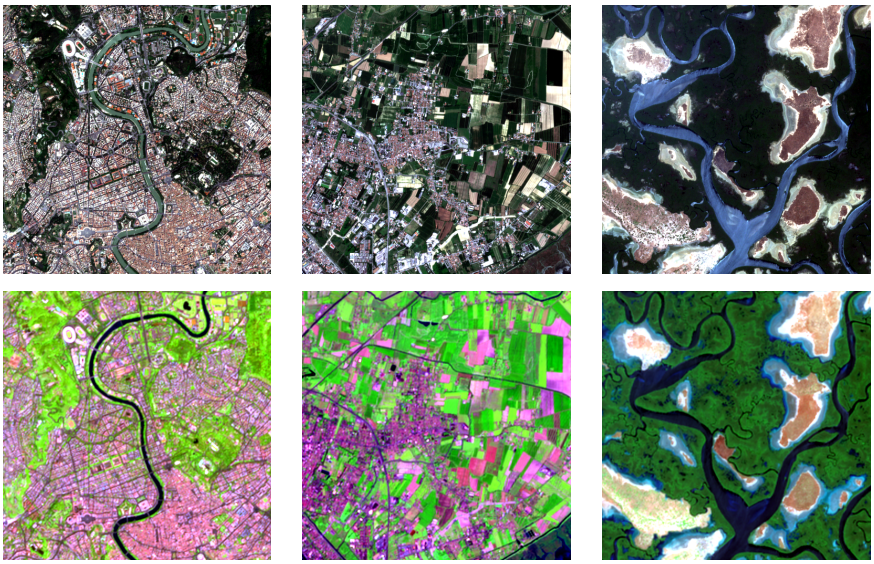
Another rather critical issue is the training dataset selection as it impacts

the capability of the trained models to generalize well on unseen data. In the computer vision domain, a huge effort has been devoted to the collection of very large datasets in order 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, within the remote sensing domain, there are no examples of datasets which are as large as ImageNet or Kitty. This is due to several obstacles, among which the cost of the data and the related labeling which requires domain experts, as well as 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 labeling 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 [18, 148, 156]. The generalization limits of a pansharpening model trained on too few data have been stressed in [115], 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. [119], aiming for a roughly even distribution on the globe and for variety in terms of climate zone, land-cover and biome type. In the following, instead, is described 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 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 be aligned with the work of Lanaras et al. [119], Sentinel-2 data without atmospheric correction (L1C product) are used for the experiments.

For training and validation, three scenes have been chosen (see Figure 4.5), corresponding to different environmental contexts: Venice, Rome, and Geba River.

In particular, the three scenes have been randomly cropped in 18,996



Rome Venice Geba River

Figure 4.5: Examples of images used for training. (Top row) RGB-composite images using 10 m bands B4(R), B3(G) and B2(B), subset of $\mathbf{z}$; and (Bottom row) corresponding 20 m RGB subset of $\mathbf{x}$, using B5(R), B8a(G) and B11(B).

square tiles of size 33×33 (at 20 m resolution) to be used for training (15,198) and validation (3898). Besides, four more scenes have been selected for the purpose of testing, namely Athens, Tokyo, Addis Abeba, and Sydney, which present different characteristics, hence allowing for a more robust validation of the described model. From such sites, three 512×512 crops at 10 m resolution are singled out, for a total of twelve test samples.

4.2.3 Proposed Method

The proposed solution takes inspiration from two state-of-the-art CNN models for pansharpening, namely PanNet [148] and the target-adaptive version [115] of PNN [18], both conceived for very high resolution sensors such as Ikonos or WorldView-2/3. Both methods rely on a residual learning scheme, while main differences concern loss function, input preprocessing, and overall network backbone shape and size.

Figure 4.6 shows the top-level flowchart of the proposed method. Since

Sentinel-2 images are used, differently from [148] and [115], the input is composed by 10 bands, six lower-resolution ones ($\mathbf{x}$), to be super-resolved, plus four higher-resolution bands ($\mathbf{z}$). For each band x to be super-resolved, a single (relatively small) network is trained as represented at the output in Figure 4.6.

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 make use of a lighter solution with computational advantages. The second reason is related to the experimental observation that training separately the six networks allows 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 own dataset. Turning back to the workflow, observe that both input subsets, $\mathbf{x}$ and $\mathbf{z}$, are high-pass filtered (HPF) as also done by PanNet. This operation relies on the intuition that the missing details that the network is asked to recover lie in the high frequency range of the input image. Next, the HPF component $\mathbf{x}^{\mathrm{hp}}$ is upsampled ($R \times R$) using a standard bicubic interpolation, yielding $\widetilde{\mathbf{x}}^{\mathrm{hp}}$, in order to match the size of $\mathbf{z}^{\mathrm{hp}}$ with which to be concatenated prior to feed 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 combination, obtained through a skip connection that retrieves the low-resolution content of $\hat{x}$ directly from the input, is known as residual learning strategy [106], and has soon became a standard option for deep learning based super-resolution and pansharpening [148, 115, 119], as it is proven to speed-up the learning process.

The CNN architecture is more similar to the pansharpening models [18, 115] than to PanNet [148], making use of just four convolutional layers, whereas PanNet uses ten layers, each singling out 32 features (except for the output layer). Moreover, a batch normalization layer operating on the input stack precedes the convolutional ones. This has proven to make the learning process robust with respect to the statistical fluctuations of the training dataset [116]. In Table 4.5, the network hyper-parameters of the convolutional layers are summarized.

4.2.4 Training

Once the training dataset and model are fixed, a suitable loss function to be minimized needs to be defined in order for the learning process to take place.

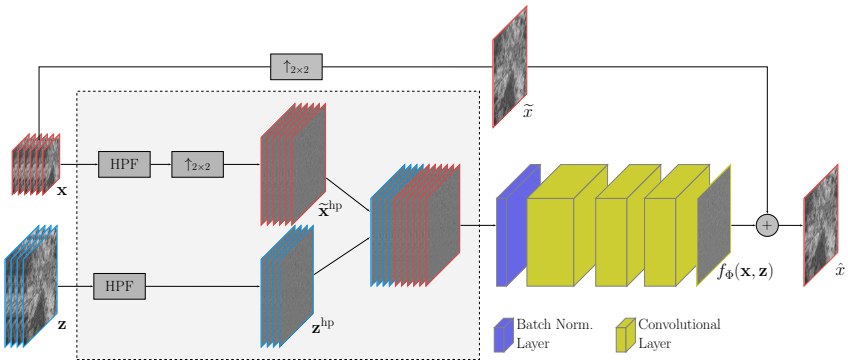


Figure 4.6: Top-level workflow for the super-resolution of any 20 m band of Sentinel-2. The dashed box gathers the shared processing which is the same for all predictors.

Table 4.5: 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

L_2 or L_1 norms are typical choices [54, 18, 115, 119, 55] due to their simplicity and robustness, with the latter being probably more effective to speed-up the training, as observed in [115, 119]. 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 [158], which is an indirect error measurement performed in a suitable feature space generated with a dedicated CNN. In [148], structural consistency is enforced by working directly on detail (HPF) bands. In the proposed solution, in addition to the use HPF components, a combined loss that explicitly accounts for spectral and structural consistency is defined. In particular, inspired by the variational approach [159], it has been used the following loss function:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{Spec}} + \lambda_2 \mathcal{L}_{\text{Struct}} + \lambda_3 \mathcal{L}_{\text{Reg}} \tag{4.1}$$

where three terms, corresponding to fidelity, or spectral consistency ($\mathcal{L}_{\text{Spec}}$), structural consistency ($\mathcal{L}_{\text{Struct}}$) and regularity ($\mathcal{L}_{\text{Reg}}$), 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 [115, 119], the fidelity term is based on the L_1 norm, that is

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

where the expectation $\mathrm{E}\{\cdot\}$ 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, for the sake of simplicity the subscript $\downarrow$ is omitted.

The structural consistency term is given by

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

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 [159]. It has been shown that the gradient distribution for real-world images is better fit with a heavy-tailed distribution such as a hyper-Laplacian ($p(x) \propto e^{-k|x|^p},\ 0 < p < 1$). Accordingly, here a L_p-norm with $p = 1/2$ is used, which could be more effective [159]. 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 regularization term that penalizes the total variation of $\hat{x}$ helps to avoid unstable behaviors:

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

Eventually, the network parameters were (pre-)trained by means of the Adaptive Moment Estimation (ADAM) optimization algorithm [76] applied to the above-defined overall loss (Equation (4.1)). In particular, ADAM has been set with 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 [160]. 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.

4.2.5 Experimental Results

In the following, after a brief recall of the accuracy evaluation metrics (Section 4.2.6) and of the comparative methods (Section 4.2.7), follows a discussion on numerical and visual results (Section 4.2.8).

4.2.6 Accuracy Metrics

The quality assessment of pansharpening algorithms can be carried out in two frameworks, with or without ground-truth. Since normally the ground-truth is unavailable, the former context refers to the application of Wald's protocol [157], which is the same process used for the generation of training samples, as described in Section 4.2.2. 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.

In particular, in the following, reference-based metrics are used:

- Universal Image Quality Index (Q-Index) takes into account three different components: correlation coefficient, mean luminance distance and contrasts [152].
- *Erreur Relative Globale Adimensionnelle de Synthése* (ERGAS) measures the overall radiometric distortion between two images [153].
- Spectral Angle Mapper (SAM) measures the spectral divergence between images by averaging the pixel-wise angle between spectral signatures [161].
- High-pass Correlation Coefficient (HCC) is the correlation coefficient between the high-pass filtered components of two compared images [162].

On the other hand, as no-reference metrics, we use the following [140, 17]:

- 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.

4.2.7 Compared Methods

The FUSE method described above, is compared with several state-of-the-art solutions. On the one side are classical approaches for pansharpening, properly generalized to the case of Sentinel-2, such as the following:

- Generalized Intensity Hue Saturation (GIHS) method [134]
- Brovey transform-based method [163]
- Indusion [164]

- Partial Replacement Adaptive Component Substitution (PRACS) [165].
- A Troús Wavelet Transform-based method (ATWT-M3) [136]
- The High-Pass Filtering (HPF) approach [135]
- Generalized Laplacian Pyramid with High Pass Modulation injection (MTF-GLP-HPM) [166]
- Gram-Schmidt algorithm with Generalized Laplacian Pyramid decomposition (GS2-GLP) [166]

Detailed information about these approaches can be found in the survey work of Vivone et al. [140].

Besides, FUSE is also compared with the following deep learning approaches native for Sentinel-2 images, including two ablations of the proposal:

- A previous CNN-based method (M5) proposed in [149] and described in Section 4.1, extended (training from scratch) to all six 20 m bands.
- The CNN model (DSen2) proposed in [119], which is much deeper than FUSE and has been trained on a very large dataset.
- An enhancement of M5 where High-Pass filtering on the input and other minor changes have been introduced (HP-M5) [55], 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.

4.2.8 Numerical and Visual Results

To assess the performance of the proposed method, twelve 512×512 images (at 10 m resolution) are extracted 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 Table 4.6. 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, it can be seen that the proposed

Table 4.6: 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
Lanaras	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**

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, in this case, three well-established indicators that are usually employed in the pan-sharpening context are considered: 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 Table 4.6, the proposed method, again, outperforms the competitors. A slightly better spectral preservation is given by HP-M5, M5 and ATWT-M3.

Giving a look at some sample results starting from the full-resolution context. Figures 4.7 and 4.8 show some of the 512 $\times$ 512 images used for test, associated with urban and extra-urban contexts, respectively. For the sake of visualization, an RGB false-color subsets of $\mathbf{z}$ and $\mathbf{x}$ is used. In particular, three out of four bands of $\mathbf{z}$ (B2, B3 and B4) are employed, and three out of

six bands of $\mathbf{x}$ (B5, B8a and B11—see Table 4.3). The input components $\mathbf{z}$ and $\widetilde{\mathbf{x}}$ are shown on the left and middle columns, while the super-resolution $\hat{\mathbf{x}}$ obtained with the proposed method is shown on the right.

Although at a first glance these results look pretty nice, a different observation scale would help to gain insight the behavior of the compared solutions. Therefore, Figure 4.9, shows some zoomed details with the corresponding super-resolutions using different selected methods. In particular, for the sake of simplicity, the visual inspection is restricted to the most representative DL and not DL approaches according to both reference-based and no-reference indicators reported in Table 4.6. A careful inspection reveals that some model-based approaches provide higher detail enhancement compared to DL methods. However, it remains difficult to appreciate the spectral preservation capability of the different methods due to the lack of objective references.

Actual errors can be visualized in the reduced-resolution domain instead. Figure 4.10, shows in particular 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 left-most 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 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 pretty well in comparison with model based approaches as the error map is nearly constant gray, whereas for PRACS and ATWT-M3 visible piece-wise color shifts are introduced. This observation does not contrast with the good values of SAM obtained by PRACS, since this indicator accounts for the relative color/band proportions but not for their absolute intensity (some "colorful" error maps in Figure 4.10 are partially due to the band-wise histogram stretching used for the sake of visualization). Overall, by looking at both 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.

4.2.9 Discussion

To assess the impact of the proposed changes with respect to the baseline HP-M5, an additional convolutional layer and a composite loss that adds a regularization term and a structural term to the basic spectral loss (L1-norm), an ablation study is also carried out. In particular, a the three-layer scaled version of FUSE is considered, as well as the four-layer version trained without

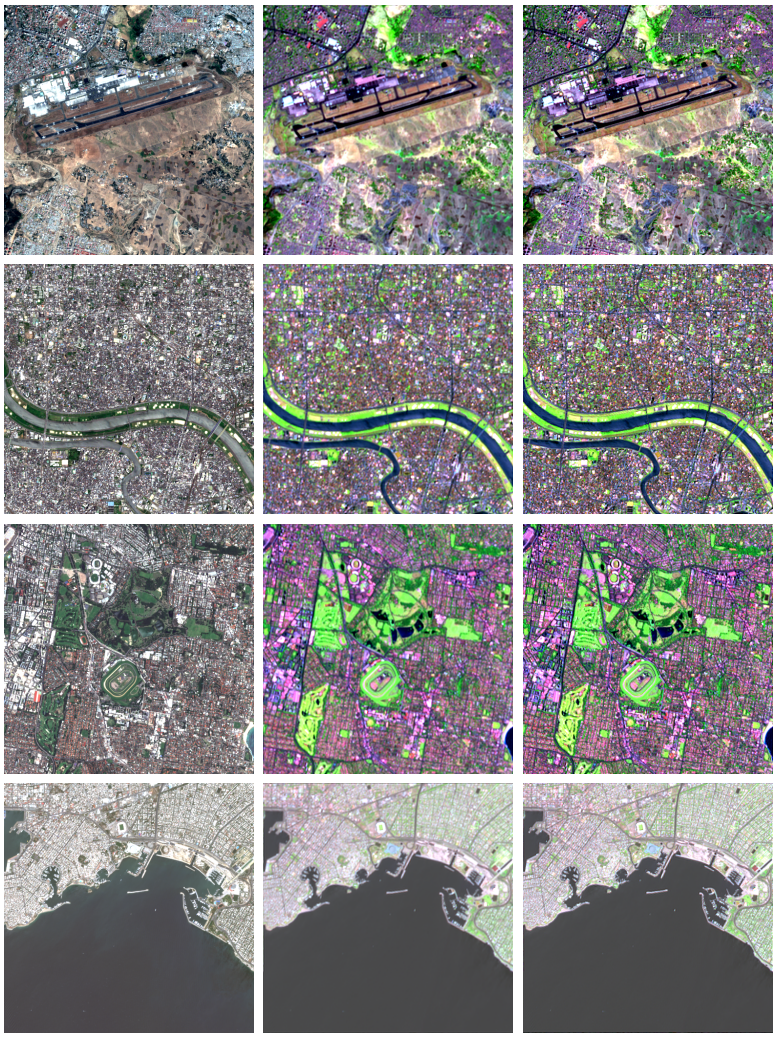


Figure 4.7: Super-resolution of the test images—Urban zones. From top to bottom: Adis Abeba, Tokyo, Sydney, and Athens. From left to right: High-resolution 10 m input component $\mathbf{z}$, low-resolution 20 m component $\widetilde{\mathbf{x}}$ to be super-resolved, and super-resolution $\hat{\mathbf{x}}$ using the FUSE algorithm.

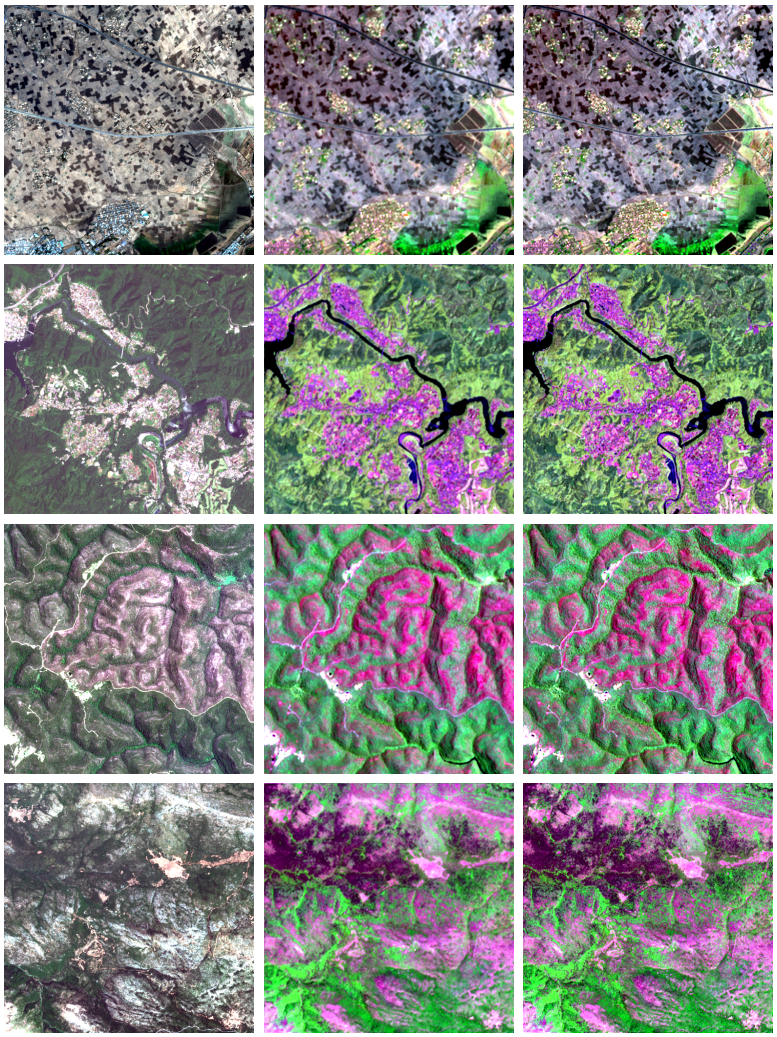


Figure 4.8: Super-resolution of the test images—Extra-urban zones. From top to bottom: Adis Abeba, Tokyo, Sydney, and Athens. From left to right: High-resolution 10 m input component $\mathbf{z}$, low-resolution 20 m component $\widetilde{\mathbf{x}}$ to be super-resolved, and super-resolution $\hat{\mathbf{x}}$ using the FUSE algorithm.

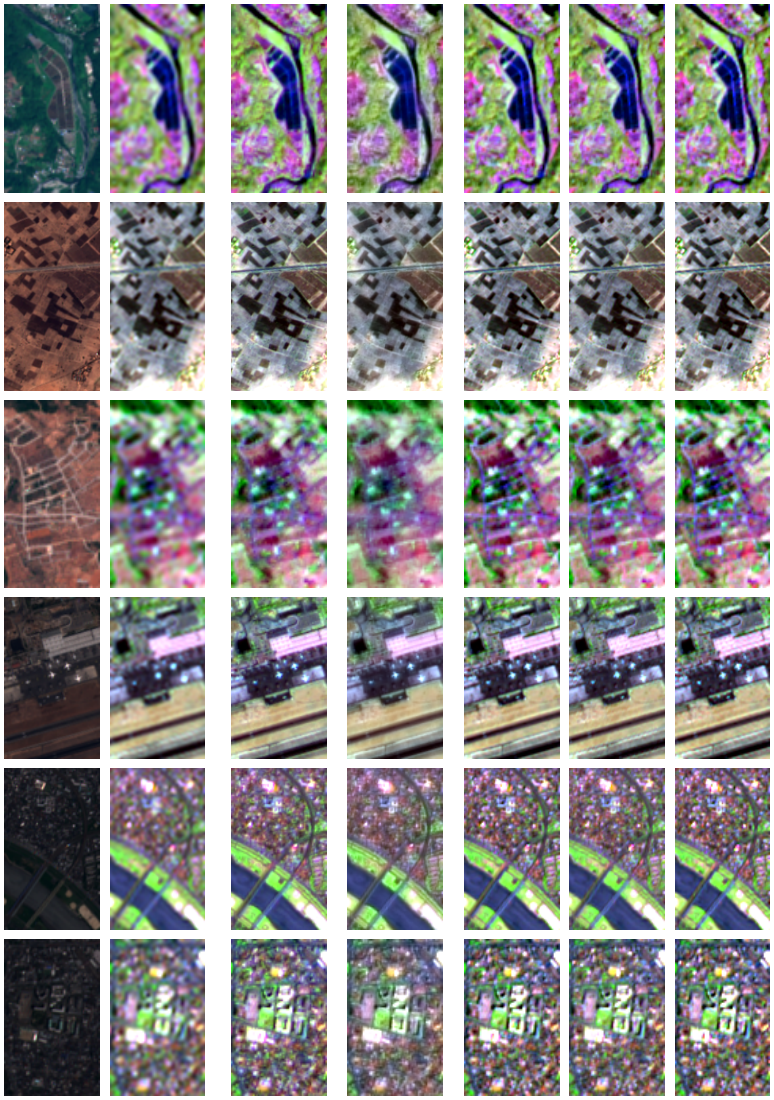


Figure 4.9: Full-resolution results for selected details. For each detail (row) from left to right are shown the two input components to be fused, followed by the corresponding fusions obtained by compared methods.

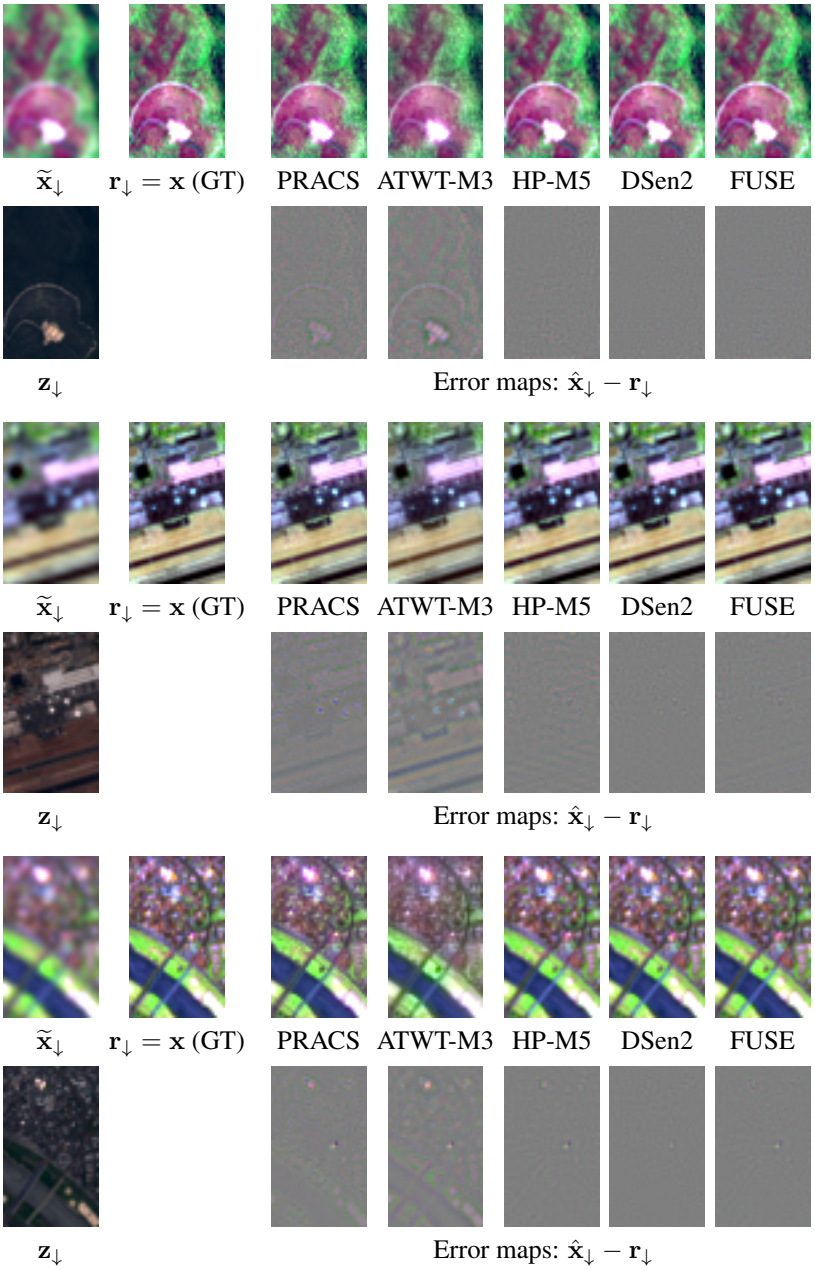


Figure 4.10: Reduced-resolution samples. Bottom images (Columns 3–7) show the difference with the ground-truth (GT).

Table 4.7: 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	8.7 GB	9.2 GB	17.4 GB	out of memory
	3.4 s	4.3 s	7.4 s	9.8 s	-
FUSE	391 MB	499 MB	707 MB	1.1 GB	1.9 GB
	6×0.45 s	6 × 0.47	6 × 0.50 s	6 × 0.55 s	6 × 0.60 s

regularization and structural loss terms. These two solutions are also reported in Table 4.6. 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, it seems that the use of the composite loss has a relatively better impact compared to the network depth increase. This is particular evident looking at the SAM indicator.

The experimental evaluation presented above confirms the great potential of the DL approach in the context of the data fusion problem at hand, as already seen for pansharpening [18] and single-image super-resolution of natural images [54] a few years ago. The numerical gap between DL methods and the others is consistent and confirmed by visual inspection. In particular, it can be observed that the use of the additional structural loss term, the most relevant change with respect to the previous models M5 and HP-M5, allowed 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, the proposed framework is focused on the use of a relatively small CNN model that involves about 28K parameters in contrast to DSen2 which has 2M parameters. In Table 4.7, are gathered a few numbers obtained experimentally on a single GPU Quadro P6000 with 24 GB of memory. For both the proposed and DSen2, are shown the GPU memory load and the computational time for the inference with respect to the image size.

As the proposed model is replicated, with different parameters, for each of the six bands to be super-resolved, are assumed either a sequential GPU usage (as done in the table) or a parallel implementation, therefore with 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 $\times$ 512 image, FUSE can super-resolve a 16$\times$ larger image (2048 $\times$ 2048) in the same time slot. In addition, 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, which would further highlight the advantage of using a smaller network. To have a rough idea of this, according to [119], DSen2 was trained in about three days on a NVIDIA Titan Xp 12 GB GPU, whereas the training of this model took about 3 h using a Titan X 12 GB.

4.2.10 Conclusions

In this Section has been proposed and validated experimentally a CNN-based super-resolution method for the 20 m bands of Sentinel-2 images, which blends high-resolution spatial information from the 10 m bands of the same sensor. The proposed network is relatively small compared to other state-of-the-art CNN-based models, such as DSen2, achieving comparable accuracy levels in both numerical and subjective visual terms. Overall, it is worth noticing that DL methods overcome model-based approaches especially in terms of spectral distortion (see Figure 4.10), which is rather interesting considering that the two band sets to be fused are only partially overlapped/correlated, as can be seen in Table 4.3. In light of this, it will be interesting to explore in future work s the extension to 60 m bands of the proposed approach.

Chapter 5

Preliminary results

In the first part of this chapter are presented some preliminary results for the task of cloud detection. In particular, leveraging on the capability of CNN to accurately approximate complex relationships between raw data and higher-level products, a U-Net-like solution is proposed conceived for Sentinel-2 images. In order to face the scarcity of training data, a proper domain adaptation strategy has been pursued, which resorts to a labeled Landsat-8 dataset. Preliminary results show a consistent improvement over standard tools.

Moreover, in the last part of the chapter follows a study on the impact of the training set design for the despeckling task using CNNs. For the despeckling task, the dataset choice is a critical task since there is no 'clean' reference to use as example. Indeed, the speckle is an inherent an unavoidable property of coherent imaging systems. For these reason a comparison between a reference obtained performing a temporal multilooking, using different acquisitions of SAR images, with a simulated-speckle approach.

5.1 Cloud detection

Cloud detection is of critical importance for many monitoring applications based on passive imaging. Examples are vegetation/forest monitoring [167, 168] canopy chlorophyll mapping [169], coastal monitoring [170], land-cover classification [171] and so forth. Whenever clouds do not represent the actual objective for the given application, they become an issue, playing as occluding objects. In the latter case, cloud detection helps to define to what extent and where an image has to be taken into consideration to infer the estimation of any physical parameter of interest. Moreover, when dealing with multitem-

poral series, the knowledge of the spatial distribution of the clouds within an image allows one to selectively interpolate the temporal sequence for filling its spatio-temporal gaps due to clouds [84, 85].

Clouds can cause more or less severe visibility issues as they can be opaque, semi-transparent, or nearly transparent [172] depending on the spectral bands. Such a variability makes the detection problem hard to be solved since true positives can be very similar to true negatives. Needless to say, the naive solution which resorts to the manual annotation by domain experts can be very accurate but is also a very expensive solution which is unsuited in many practical situations. Therefore many algorithms have been proposed in the past years [173, 174, 175, 176, 177]. In [173] a multitemporal approach has been proposed. An early machine learning solution based on decision trees is given in [174] instead. By suitably combining the spectral bands, on the basis of physical considerations, an adaptation to the Sentinel-2 case of the Fmask algorithm for Landsat images is proposed in [175]. Recently, with the advent of deep learning, a paradigm shift from model-based to data-driven approaches has been observed and specifically for cloud detection it is worth mentioning [176], which deals with Landsat-8 images, and [177] working on both Landsat-8 and Sentinel-2 datasets.

The goal of the presented framework is to face the cloud detection problem on Sentinel-2 images, benefitting also of available labeled datasets collected by the Landsat-8 mission. Indeed, the scarcity of training data is one of the main issues that prevent the use of data-driven methods in the remote sensing domain. The free distribution of Sentinel-2 data partly addresses such as issue as a reliable labeling (cloud masks in this case) is not available. However, coarse cloud masks can be retrieved through ESA's utility Sen2Cor,[1] a model-based algorithm performing a pixel-wise spectral-based cloud detection. Although these masks provide a low accuracy level, thanks to a suitable domain adaptation strategy, they can be used in combination with labeled Landast-8 datasets as it will be shown.

The description of the framework is organized as follows. Section 5.1.1 describes the proposed solution. Section 5.1.2 gathers and discusses the related experimental results. Finally, concluding remarks are given in Section 5.2.

[1] https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithm

5.1.1 Proposed method

In this section, firstly is presented the general architecture of the convolutional neural network (CNN) employed for detection. Then, in the last part are described different training configurations.

Network architecture

The modeling approach follows the U-Net framework [108] for two main reasons: it has proved to be very effective on diverse segmentation tasks and, second, it fits with the input characteristics. In fact, Sentinel-2 images comprise 13 spectral bands which can be split in three sets acquired at three distinct spatial resolutions, 10, 20 and 60 m, respectively. On the other hand an U-Net model is designed to work on different resolution levels as well. Therefore, is proposed a U-Net model based on three resolution levels whose top-level flowchart is depicted in Figure 5.1. On the left-hand side (encoding or contracting path) convolutional stages are interleaved by downscaling operations where the flowing features are concatenated with related input components. On the right-hand side (decoding or expansive path), symmetrically disposed stages allow to progressively integrate more abstract features (from upscaling) with localized ones (from skip connections). Details on stages and scaling are summarized in Tab. 5.1. Therefore, the four 10-m Sentinel-2 bands feed Stage 1, whose output (that also skips toward Stage 5) is passed to the 20-m resolution level using maxpooling. Here, six 20-m Sentinel-2 bands concatenate with the downscaled features prior to feed Stage 2. The same scheme is then repeated at the coarsest level, where the three remaining 60-m Sentinel-2 bands join the feature flow. Finally, the expansive path allows to restore the feature resolution to 10 meters, simultaneously converting them in a single cloud probability map in output. The hard cloud mask is therefore provided by thresholding the soft map at 0.5.

In this study different working modalities have been hypothesized to assess the impact of the different bands of Sentinel-2 on cloud detection. In particular, there will be also considered the cases where the 20-m and/or 60-m band sets are excluded in input. In these cases the architecture shown in Fig. 5.1 changes accordingly, by skipping the interested concatenation nodes on the contracting section.

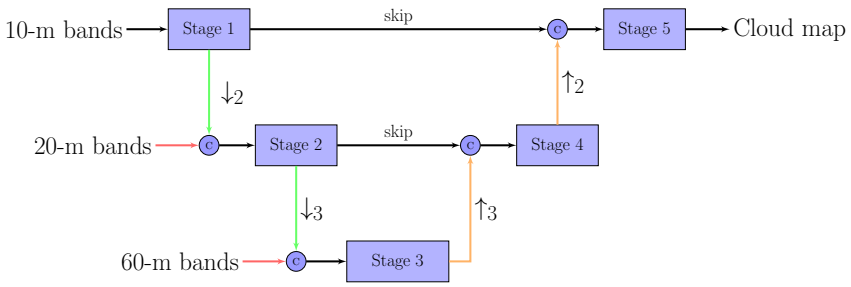


Figure 5.1: Network architecture. C-nodes stand for concatenation. Down and up arrows perform downscaling and upscaling, respectively.

Datasets and traning

The above presented network is trained following two approaches. The first one makes exclusive use of Sentinel-2 images, the second one combines Sentinel-2 and Landsat-8 datasets. In the first case, Level-1C Sentinel-2 products are considered as source for input samples, while complementary reference cloud maps have been retrieved from the corresponding Level-2A products. These maps are actually outcomes of ESA's Sen2Cor tool, and are not very accurate. Now, a reasonable question arises which is why to learn a coarse predictor such as Sen2Cor if it is of low quality and anyhow already available? Surprisingly, as it will be shown in the next section, is possible to overcome the same Sen2Cor that provided us the reference maps for training. Before continuing, in the following some information on the involved dataset are given.

For training purposes three different scenes that are representative of different environmental contexts have been chosen:

- Munich (Germany). Includes agricultural fields, plains and lakes. Elevation: $\sim$600 m.
- Gobabeb (Namibia). Desert site including bright spots of sand and mountains. Elevation: $\sim$370 m.
- Pretoria (South Africa). Grassland, savannah, and dry woodland. Elevation: $\sim$1500 m.

From each scene 3721 not overlapping 180×180 patches are randomly selected whose 80% is organized in mini-batches with 32 examples each to be used in training. The rest is used for validation. For the testing phase we have

Stage	Layers	kernel	features	activation
1	2×(BN+Conv)	3×3	8	ReLU
2	2×(BN+Conv)	3×3	16	ReLU
3	4×(BN+Conv)	3×3	32	ReLU
4	2×(BN+Conv)	3×3	16	ReLU
5	2×(BN+Conv)	3×3	8	ReLU
	+(BN+Conv)	1×1	1	Softmax

(a)

Scaling	Layer	kernel
$\downarrow_R$	MaxPool	$R\times R$
$\uparrow_R$	Deconv	$R\times R$

(b)

Table 5.1: Details on stages (a) and scaling (b).

manually created very accurate ground-truths for five selected 540×540 scenes not used for training/validation.[2]

As loss function to minimize a combination of the binary cross-entropy L_{bce} and of the Jaccard loss L_{Jacc} is employed, *i.e.*,

$$L = L_{\text{bce}} + L_{\text{Jacc}}, \tag{5.1}$$

with

$$L_{\text{bce}} = -\sum_i y_i \log(p_i) + (1-y_i)\log(1-p_i) \tag{5.2}$$

and

$$L_{\text{Jacc}} = \sum_i 1 - \frac{y \cap p}{y \cup p}, \tag{5.3}$$

where y_i and p_i are the i-th pixel values of the binary reference y and of the predicted probability map p, respectively. On one side, by minimizing the binary cross-entropy favors global accuracy whereas, on the other side, reducing Jaccard's loss allows to achieve good local precision [178]. The optimization is then carried out by running Adam algorithm [76] for 200 epochs with an initial learning rate of 10^{-4}.

[2] All image sizes are specified at full scale. For Sentinel-2 it means 10 m.

In order to asses the influence of the different spectral bands on the detection result, four input configurations have been envisaged and a corresponding network has been trained with the above described procedure:

- U-Cloud (UC): all thirteen Sentinel-2 bands.
- UC-10: only four 10-m bands.
- UC-10-20: four 10-m and six 20-m bands.
- UC-10-60: four 10-m and three 60-m bands.

In addition to these models, given the unsatisfactory performance observed (see next section), which is manly due to the low-quality of the Sen2Cor references, is employed a domain adaptation strategy to allow the use of a more accurate training dataset which is publicly available [176] but coming from a different sensor, *i.e.*, Landsat-8. Such a dataset is composed of 95 images and which 18 scenes with associated, manually extracted, cloud masks have been used for the training. About 7000 192×192 patches were extracted from these images, and grouped in 32-minibatches for the purpose of training. A data augmentation (flipping, rotation, zoom) was also applied for regularization. This dataset actually does not provide all Landsat-8 bands but just four channels that roughly correspond to the 10-m bands of Sentinel-1, with the difference that here the resolution is 30 m. Therefore, a training from scratch on the Landsat-8 dataset has been performed only for model UC-10 introduced above.

Once converged, this trained model is used to move to the Sentinel-2 domain (limited to 10-m bands) running a few additional training epochs to let the model getting adapted to the actual target domain.

5.1.2 Experimental results

Five 540×540 images have been manually segmented to serve for testing purposes. In order to objectively assess the detection and segmentation accuracy commonly employed numerical figures such as F_1-score, intersection over union (IoU), precision and recall are used. The averaged results obtained on the test images are summarized in Tab. 5.2. Starting with a focus on the models that do not undergo domain adaptation, at first glance, it seems that some ablated configurations, UC-10 and UC-10-60, provide higher scores compared to the full-input model UC. Indeed, it has to be recalled that the training was carried out using Sen2Cor masks as reference rather than manually drawn ground-truths as those used for testing. In this perspective, if Sen2Cor masks are used

	IoU	Precision	Recall	F_1
Sen2Cor	0.1982	0.9996	0.1982	0.3308
UC	0.2408	0.8307	0.2532	0.3881
UC-10-20	0.1031	0.8978	0.1043	0.1869
UC-10-60	0.2928	0.8347	0.3109	0.4531
UC-10	0.2672	0.9893	0.2679	0.4217
UC-10-DA	0.5797	0.9494	0.5982	0.7339

Table 5.2: Numerical results.

as reference for tests, then UC would be the best one. This can be further appreciated by inspecting Fig. 5.2, where some sample results are shown. As it can be seen, UC results are the closest to Sen2Cor, while the partial variants either over estimate (UC-10, UC-10-60) or under estimate (UC-10-20) Sen2Cor maps. In any case, it is worth to observe that all but UC-10-20 provide maps which are closer than Sen2Cor to the actual ground-truth, as confirmed by the numerical scores of Tab. 5.2 (Sen2Cor gets the highest precision score due to its too conservative setting). This can be partly interpreted considering that Sen2Cor maps are built using a pixel-wise processing, whereas our convolutional networks, being capable to also describe spatial interactions, have eventually learned shape priors for clouds, boosting further the segmentation accuracy.

In particular, UC-10 provides the best segmentation map, detecting also rather small cloud spots that are neglected by others. This can be partly explained by observing that UC-10 makes exclusive use of the highest resolution Sentinel-2 bands, whereas its competitors make also use of lower resolution bands, some of which, directly related to the cloud phenomena, may dominate the network decision with a consequent cost in resolution. In fact, 60-m bands are sensible to aerosols, water-vapour and cirrus, while 20-m bands provide responses for snow, ice and clouds.

Moving to the domain adapted solution UC-10-DA a considerable jump on IoU, recall and F_1-score can be read. Such a superiority can be also appreciated by visually inspecting the predictions of Fig. 5.2. In particular, the first test image (top row) where none of the other solutions is capable to "see" any cloud, likely because of their small size, whereas UC-10-DA provides a consistent map.

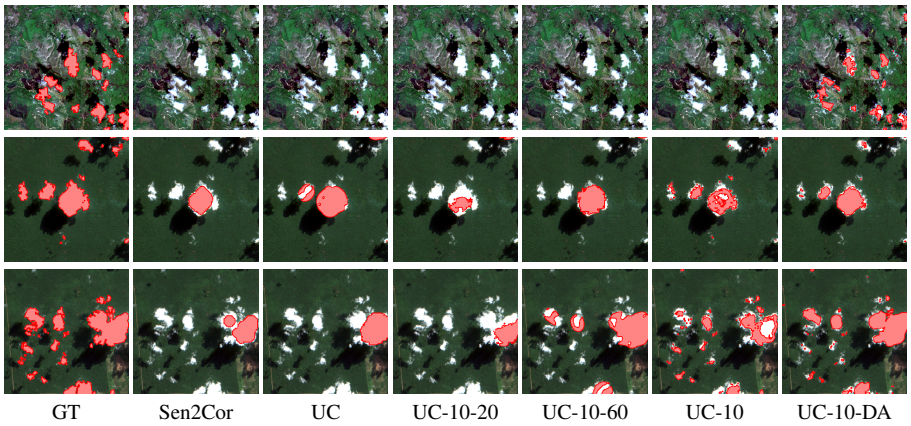


Figure 5.2: Cloud detection results. From left to right: the ground-truth (GT), and the predictions by Sen2Cor, U-Cloud (UC), UC-10-20, UC-10-60, UC-10, and the UC-10 trained with the proposed domain adaptation strategy, namely UC-10-DA.

5.2 Conclusions

A framework for cloud detection on Sentinel-2 images based on an U-Net architecture has been presented. The experimental analysis allows to draw three main concluding remarks. First, despite their different spatial resolution, all Sentinel-2 spectral bands count for cloud segmentation accuracy. Nevertheless, second remark, proper domain adaptation strategies leveraging on rich and accurate labeled datasets from other sensors can considerably boost the segmentation accuracy. Finally, third remark, the availability of a large and accuratately labeled Sentinel-2 dataset remains necessary to fully exploit the information conveyed by all available spectral channels. In addition, it is useful recall that the Copernicus program provides also SAR images through Sentinel-1 which could complement the optical acquisitions by means of a CNN-based data-fusion approach. All above remarked considerations address future research on this topic.

5.3 SAR Despeckling

Synthetic aperture radar (SAR) images are precious sources of information for the most diverse Earth Observation applications. Despite their many desir-

able features, their usability is severely limited by the intense speckle noise affecting them, intrinsic of coherent imaging mechanisms. Speckle removal is necessary to reliably perform image processing primitives, from edge detection to segmentation or classification. In fact, SAR despeckling keeps being a very active field of research, with a large number of despeckling methods proposed over the years, from spatially-adaptive filters, to transform-domain methods, variational techniques, and more recently nonlocal filters, arguably the current state of the art [179, 180].

In the last few years, all image processing tasks have been re-examined in the light of deep learning (DL), and SAR despeckling is no exception. A large number of deep learning-based methods have been proposed for this task [181] and some of the most sound and popular are proposed in [182, 183, 184, 185, 186, 187]. However, only limited performance improvements have been achieved thus far. The main reason is certainly the very high noise intensity (here the focus is on the most interesting single-look case) which makes this problem especially challenging. However, another major reason is the absence of reliable reference data to train deep neural networks.

In fact, a primary requirement of learning-based methods is the availability of examples from which to learn. For despeckling, this means having speckle-free images to use as ground truth for the noisy observations. However, speckle is an inherent an unavoidable property of coherent imaging systems, hence there is no such thing as a "clean" SAR image. So, the fundamental issue in deep learning-based SAR despeckling is not architectures but training. How is it possible to teach a deep network to perform a noisy-to-clean mapping in the absence of desired clean examples?

In the literature, two approaches are commonly followed to address this problem:

- approximating as closely as possible a real-world clean SAR image by means of temporal multilooking;
- simulating pairs of real and noisy SAR images based on available data and suitable models.

This Section aims at analyzing these approaches experimentally, pointing out their strengths and weaknesses, and establishing useful guidelines.

In next Section are described in some more depth these two approaches, then are presented and discussed experimental results for both of them in Sections 5.3.2 and 5.3.3, to eventually draw conclusions in Section 5.3.4.

5.3.1 Training set design in the literature

Supervised image restoration relies on pairs of noisy/clean images to be used as examples to train a deep neural network. Now, although clean SAR images do not exist, they could be obtained, in principle, by averaging an infinite number of SAR images, characterized by the same signal but incoherent realizations of speckle. Indeed, since the speckle is due to the presence of a large number of independent scatterers in the same cell, even tiny variations in the acquisition geometry will give rise to the desired independent realizations of speckle. Therefore, if one acquires a large number of images of the same scene at different times and averages them, a process called *temporal multilooking*, a good approximation of the desired reference image is obtained.

Although appealing, this approach presents a number of technical problems. First of all, stacks of multitemporal images are not widely available. In any case, satellites revisit the same area and acquire new images only after several days, therefore, the number of available images is intrinsically limited. Then, the acquired images must be accurately co-registered to a common master, a non-trivial process which may affect the statistical properties of slave images. Finally, a major problem with multitemporal fusion is represented by temporal changes, due both to natural phenomena (think of seasonal changes) and human activity (new buildings, deforestation) leading to unreliable references. These regions can be excluded from the training set, but the more images are used, the less likely it is to find unchanged regions.

Probably because of all the above problems, only a few research groups have adopted the temporal multilooking approach. In [182, 188, 186] a stack of 25 single-look COSMO-SkyMed images is used for training with a leave-one-out strategy: 24 images are multilooked to provide the desired reference for the remaining one. A similar procedure is also used in [189] with a stack of 52 TerraSAR-X images.

The majority of papers on DL-based SAR despeckling follow a simpler approach for dataset creation, based on *simulation*. The idea is to collect clean images, with the same statistics expected of clean SAR images, and corrupt them with simulated noise with the same statistics of real speckle. If the speckle is fully developed it follows a gamma distribution, with parameter L equal to the number of looks. Assuming it also spatially white, generating the desired speckle field becomes straightforward. As for the clean SAR image, it is usually approximated by general-purpose or remote sensing optical images.

This approach, followed for example in [183] and [184], allows one to generate a virtually unlimited number of clean-noisy pairs, with the desired level

of noise, thus ensuring a very accurate training phase and ample design freedom. On the other hand, the simulated images are very far from the real ones, giving rise to a large gap between the training and the testing domains. First of all, the statistics of (clean) SAR and optical images are very different, due to the different nature of the two imaging mechanisms. Just think of corner reflectors and double reflection lines, abundant in SAR images but absent in optical images. In addition, in real-world SAR images, the speckle is not always fully developed but varies spatially, and often is spatially correlated, which may impact significantly on results. All these sources of mismatch shade serious doubts on the actual transferability to SAR images of models trained with the fully simulated approach. A step towards more realistic simulation consists in averaging [185] on more in general filtering [190, 191] a stack of single-look SAR images to obtain the clean reference. This preserves the signal statistics, even though a field of fully developed speckle is still injected on it to simulate the noisy instance.

Finally, it is worth mentioning some recent works on unsupervised DL-based SAR despeckling [187, 192] which avoid the need of ground truths altogether. Though departing from the classical approach, this is a very interesting development which fits especially well the case of SAR despeckling.

5.3.2 Experiments: temporal multilooking

In all the following analyses, is considered a fixed and relatively small SAR despeckling architecture, SAR-CNN [182]. This makes full sense, since the focus here, is on the effects of training, and expect only minor dependencies on the specific architecture.

With temporal multilooking approach reference data is obtained by averaging the largest possible number of co-registered SAR images. So, a number of key questions arise:

- how many temporal instances should be averaged to obtain a satisfactory reference?
- what is the impact of undetected temporal changes?
- what is the impact of imperfect co-registration?
- can multitemporal/spatial filtering improve results?

To address the first point the reference images are generated by averaging an increasing number of temporal instances, keeping fixed all other hyperparameters. However, since only a small number of co-registered real SAR

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

Figure 5.3: Results for a COSMO-SkyMED clip. Left: input noisy image (a) and 24-look multitemporal reference (b). Right: images despeckled by SAR-CNN trained on 24-look SAR (c) single-look SAR (d) UC-Merced (e) UC-Merced equalized (f).

Table 5.3: PSNR as a function on multilooking depth

	∞	32	16	8	4	2	1
with bias	26.16	26.11	26.10	26.09	25.93	25.31	22.82
compens.	26.16	26.12	26.12	26.12	26.05	26.02	25.87

images are available, this analysis (hardly affected by the domain gap problem) is carried out on simulated images in which also the references are obtained through temporal multilooking. Thus, the SAR-CNN network is trained, using images drawn from the UC-Merced dataset [193] for land-use classification. The dataset includes 21 subsets of 100 256×256-pixel images each, corresponding to different semantic classes. From each subset are kept 80 images for training, 10 for validation and 10 for testing, extracting 40×40-pixel patches with stride 10, for a total of about 1 million patches. Spatially white gamma-distributed speckle, with parameter L=1, is generated to simulate the noisy patches. Then, different realizations of the same noisy patch are averaged to obtain the version with L-look speckle, with L going from 1 to 1024 in powers of 2. Eventually the clean patches are used to simulate the $L = \infty$ case. SAR-CNN is then trained anew for each case, always for 50 epochs, using ADAM with initial learning rate set to 0.001 decreased by a factor 10 after the 30th epoch.

Tab.5.3 shows numerical results on the testing images in terms of peak signal-to-noise ratio (PSNR) with respect to the clean reference. Performance (row 1) remains quite stable as L decreases from ∞ to 8, with a sharp impairment when it reduces further. However, visual inspection makes clear that this impairment is only due to a bias on the output image scale. After compensating this bias with respect to the available noisy input, the performance drop for small values of L becomes almost negligible. This result is not really new. In the seminal Noise2Noise paper [194] it was already observed that a CNN can be trained effectively for image restoration using only one or more noisy realizations of the same clean image, provided that the noise has zero mean. This is exactly the case, since SAR-CNN works in the log-domain, where speckle is additive and the non-zero mean can be trivially compensated. Nonetheless, confirmation of this finding for SAR images has quite remarkable consequences. It means that only a pair of co-registered SAR images is necessary to train successfully the network, obtaining about the same results observed with a clean ground truth. Moreover, it largely de-emphasizes the importance of all other issues: there is no need to collect and co-register *many* images; with temporally close instances, changes become rare events; and neither temporal nor spatial filtering can possibly improve the reference. In fact, even with the perfect clean reference, results improve only marginally w.r.t. the $L = 1$ case.

Well aware of the domain-gap, the same experiment are repeated on a stack of multitemporal COSMO-SkyMED SAR images, for $L = 1, 2, 4, 8, 16, 24$. Even in the absence of objective measures, due the lack of a clean reference, the same behavior emerges clearly. Fig.5.3 shows, for a small test clip (a), the output of SAR-CNN trained with $L = 24$ (c) and $L = 1$ (d) (with compensation). The despeckled images are not identical, but have comparable quality. Speckle is rejected quite effectively in both cases, while major structures and regions are faithfully preserved. Of course, the comparison with the 24-look reference (b) shows that fine-grain details are lost but these are probably out of reach of any despeckling method given the overwhelming noise present in the input image, calling for more advanced approaches, involving other sources of information, for significant improvements.

5.3.3 Experiments: simulated

To analyze the approach based on SAR image simulation, is considered again the UC-Merced dataset, as customary in the literature. Training is carried out as described in the previous Section. Then, the trained SAR-CNN network is

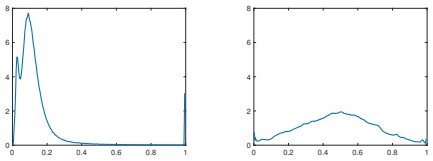


Figure 5.4: Empirical pdfs of clean SAR and optical data.

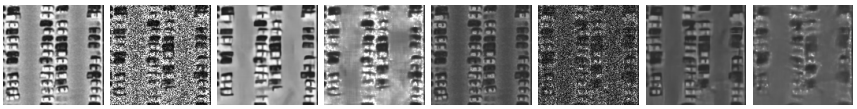


Figure 5.5: Results on simulated images, w/o (left) and with (right) equalization. Reference, noisy, aligned training, SAR training.

used to despeckle the same SAR clip considered before. The output image, shown in Fig.5.3(e), speaks by itself. Speckle is only partially suppressed, noise patterns of the single-look input are often interpreted as image structures, man-made structures are poorly represented, with a clear shift in dynamics (many bright areas are desaturated). Such a poor result is not due to poor training but to a strong domain gap. SAR-CNN works very well on aligned test data (see Fig.5.5), but cannot deal correctly with input images having wildly different statistics. The empirical pdfs of the 24-look SAR data (clipped) and of the clean optical data, Fig.5.4, highlight a strong mismatch in dynamics. To compensate for this mismatch, we repeated the training procedure starting from amplitude-equalized UC-Merced data, such to have the same pdf as SAR data. The result of Fig.5.3(f) shows a clear improvement under this respect, with man-made structures recover much better. However, the other problems remain, as they are related with mismatch in higher order statistics.

Lacking a SAR ground truth, it cannot be possible to measure objectively the impairment due to such a domain gap. On the other hand, it is reasonable to

Table 5.4: PSNR (SSIM) on simulated data

train \ test	optical	optical-eq
SAR	20.20 (0.571)	23.85 (0.638)
optical	**26.23 (0.722)**	28.17 (0.714)
optical-eq	24.19 (0.671)	**28.29 (0.721)**

assume that a similar impairment arises if we exchange the role of source and target domains. Therefore, in next experiment is used the network trained on real SAR data to despeckle the simulated data. Since the ground truth is now available, the full-reference performance measures can be evaluated. Synthetic results are reported in Tab.5.4 in terms of PSNR and SSIM (in parentheses). When plain optical images are used for the simulation, the network trained on the same data (bold) achieves a PSNR of 26.23 dB, 6 dB better than the network trained on SAR data. Part of this loss is certainly due to the dynamics mismatch. However, even when equalized optical data are used for simulation the network trained on SAR data keeps showing a loss of 4.5 dB. Similar results are observed for SSIM. In Fig.5.5, are shown also some visual results to gain a better insight on the problem. Note that, both with the original and equalized ground truth, the aligned network provides a despeckled image of very good quality. On the contrary, the network trained on SAR data does a very poor job in both cases.

5.3.4 Conclusions

In this Section an experimental study on the main approaches used to generate datasets for SAR image despeckling has been described. Temporal multilooked reference work quite well. In addition, even a few co-registered dates allow training the network satisfactorily, much reducing the problems related with data collection and pre-processing and temporal changes. On the contrary, the approach based on simulation is quite risky if the simulated data are not really aligned with the test data. This is certainly the case when optical images with injected white speckle are used: the mismatch in statistics, both of the first and higher orders, eventually leads to poor results. New architectures for SAR despeckling should be tested in the appropriate conditions before drawing conclusions on their effectiveness.

Conclusions

In this thesis, it has been explored the contribution of Deep Learning techniques, and in particular of Convolutional Neural Networks in different data-fusion tasks. It appears clear that Deep Learning can be used effectively for several tasks of data-fusion, showing a relevant impact on performances with respect to State-of-the-Art. In particular, with the use of CNNs, the fusion of the data is learnt automatically during the training process and do not require a human interaction as well as a burdensome preprocessing of the data. This lighten-up the coupling procedure of different data such as optical and SAR images.

Indeed, for the NDVI regression presented in Chapter 2 it appears clear that the CNN allows to benefit of the rich information about vegetation provided by SAR images to estimate a rough NDVI, despite its optical origin. Moreover, the use of CNN helps to fuse multitemporal optical images performing a complex interpolation. Therefore, the missing NDVI is retrieved from a joint cross-sensor/multitemporal fusion, using a single learning machinery that allows to achieve a very good accuracy.

Furthermore, in the forest classification task, by means of TanDEM-X SAR data described in Chapter 3, the CNN-based approach helps to fuse the SAR backscatter with uneven features such as volume correlation coherence and incidence angle that helps to further boost the performances. This is probably due to the ability of CNNs to account for the textural content conveyed by the additional features provided to the net.

In Chapter 4, for the super-resolution problem, the training process and the definition of an appropriate loss function have allowed, with a relatively small network, to obtain a boost of the performance with a smaller spectral distortion of the super-resolved images compared to other proposed solutions. Moreover, an ad hoc training procedure allowed to have a faster training with a relatively small architecture and a reduced size dataset with a negligible impact on quality.

In Section 5.1 a more general fusion process has been described. First of all, a CNN architecture is used to perform a fusion of Sentinel-2 optical bands, in order to obtain a cloud mapping, despite the available reference was not reliable. To overcome this issue, by training the same CNN architecture, it has been proposed a domain adaptation strategy which takes advantage of a Landsat-8 dataset with a reliable reference obtaining very promising results.

Finally, in Section 5.3 an experimental study on the main approaches used to generate SAR image despeckling dataset has been described. The fusion of multitemporal acquisition of coregistered SAR images works quite well, differently from the simulation-based approach that provides poor results when the resulting trained networks are used on real SAR images, because of the training-test data misalignment.

In conclusion, in this thesis work, by exploring diverse remote sensing data fusion problems, the CNN-based approach have shown a great potential, balancing the simplicity of the problem modeling with the capability to learn complex relationships from labeled datasets. These properties have allowed the spreading of the CNN-based solutions over a large number of data-fusion tasks, opening to new perspective and providing superior performances compared to most of the model-based approaches. However, the training process of a Convolutional Neural Network needs important requirements in terms of abundance of training examples and/or computational resources. More precisely, the more complex is the task to be solved, the deeper or the larger should be the network, therefore, the larger must be the training dataset. Fortunately, an increasing number of remote sensing sensors are providing their data for free, hence limiting the data problem to the labeling aspects.

Bibliography

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[2] Giorgio Franceschetti. *Electromagnetics: theory, techniques, and engineering paradigms*. Springer Science & Business Media, 2013.

[3] Giorgio Franceschetti and Riccardo Lanari. *Synthetic aperture radar processing*. CRC press, 2018.

[4] Charles Toth and Grzegorz Jóźków. Remote sensing platforms and sensors: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:22–36, 2016.

[5] F. Paul, S. H Winsvold, Andreas 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*, 8(7):575, 2016.

[6] 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)*, pages 902–905, July 2016.

[7] Bin Wu, Huichun Ye, Wenjiang Huang, Hongye Wang, Peilei Luo, Yu Ren, and Weiping Kong. Monitoring the vertical distribution of maize canopy chlorophyll content based on multi-angular spectral data. *Remote Sensing*, 13(5):987, 2021.

[8] Ming Lu, Bin Chen, Xiaohan Liao, Tianxiang Yue, Huanyin Yue, Shengming Ren, Xiaowen Li, Zhen Nie, and Bing Xu. Forest types classification based on multi-source data fusion. *Remote Sensing*, 9(11):1153, 2017.

[9] Chuanmin Hu, Yingcheng Lu, Shaojie Sun, and Yongxue Liu. Optical remote sensing of oil spills in the ocean: What's really possible? *J. Remote Sens.*, page 9141902, 2021.

[10] N. Ødegaard, A. O. Knapskog, C. Cochin, and J. Louvigne. Classification of ships using real and simulated data in a convolutional neural network. In *2016 IEEE Radar Conference (RadarConf)*, pages 1–6, May 2016.

[11] 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.

[12] G. Chierchia, M. El Gheche, G. Scarpa, and L. Verdoliva. Multitemporal sar image despeckling based on block-matching and collaborative filtering. *IEEE Transactions on Geoscience and Remote Sensing*, 55(10):5467–5480, Oct 2017.

[13] 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 Trans. Geosci. Remote Sens.*, 45(10):3012–3021, Oct 2007.

[14] Luan Pierre Pott, Telmo Jorge Carneiro Amado, Raí Augusto Schwalbert, Geomar Mateus Corassa, and Ignacio Antonio Ciampitti. Satellite-based data fusion crop type classification and mapping in rio grande do sul, brazil. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:196–210, 2021.

[15] 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*, 7(7):2763–2775, July 2014.

[16] C. Pohl and J. L. Van Genderen. Review article multisensor image fusion in remote sensing: Concepts, methods and applications. *International Journal of Remote Sensing*, 19(5):823–854, 1998.

[17] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva. Multispectral and panchromatic data fusion assessment with-

out reference. *Photogramm. Eng. Remote Sens.*, 74(2):193–200, February 2008.

[18] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016.

[19] 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*, 14(5):639–643, May 2017.

[20] 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*, volume 4, pages IV – 81–IV – 84, July 2008.

[21] 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*, 6(5):2159–2173, Oct 2013.

[22] Angela Errico, Cesario Vincenzo Angelino, Luca Cicala, Giuseppe Persechino, Claudia Ferrara, Massimiliano Lega, Andrea Vallario, Claudio Parente, Giuseppe Masi, Raffaele Gaetano, Giuseppe Scarpa, Donato Amitrano, Giuseppe Ruello, Luisa Verdoliva, and Giovanni 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*, 36(13):3345–3367, 2015.

[23] Monidipa Das and Soumya K. Ghosh. Deep-step: A deep learning approach for spatiotemporal prediction of remote sensing data. *IEEE Geosci. Remote Sensing Lett.*, 13(12):1984–1988, 2016.

[24] 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*, 14(3):284–288, March 2017.

[25] 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*, 14(1):3–7, Jan 2017.

[26] Nicola Clerici, Cesar Augusto Valbuena Calderón, and Juan Manuel 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*, 13(2):718–726, 2017.

[27] 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*, pages 711–718, 2017.

[28] 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*, 44(10):2828–2838, Oct 2006.

[29] Cristina Márquez, M. Isabel López, Itziar Ruisánchez, and M. Pilar Callao. Ft-raman and nir spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud. *Talanta*, 161:80 – 86, 2016.

[30] 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*, 46(5):1457–1466, May 2008.

[31] Johannes Reiche, Sytze de Bruin, Dirk Hoekman, Jan Verbesselt, and Martin Herold. A bayesian approach to combine landsat and alos palsar time series for near real-time deforestation detection. *Remote Sensing*, 7(5):4973–4996, 2015.

[32] Peijun Du, Sicong Liu, Junshi Xia, and Yindi Zhao. Information fusion techniques for change detection from multi-temporal remote sensing images. *Information Fusion*, 14(1):19 – 27, 2013.

[33] 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, 6–8 March 2017.

[34] G. Scarpa, S. Vitale, and D. Cozzolino. Target-adaptive CNN-based pansharpening. *ArXiv e-prints*, 2017.

[35] R. Gaetano, G. Masi, G. Poggi, L. Verdoliva, and Scarpa G. Marker controlled watershed based segmentation of multi-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 53(6):1987–3004, June 2015.

[36] Yun Du, Yihang Zhang, Feng Ling, Qunming Wang, Wenbo Li, and Xiaodong 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*, 8(4):354, 2016.

[37] 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)*, pages 444–448, Nov 2016.

[38] Wensong Liu, Jie Yang, Jinqi Zhao, and Le Yang. A novel method of unsupervised change detection using multi-temporal polsar images. *Remote Sensing*, 9(11):1135, 2017.

[39] Y. Han, F. Bovolo, and L. Bruzzone. Segmentation-based fine registration of very high resolution multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2884–2897, May 2017.

[40] 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*, 42(3):577–587, March 2004.

[41] 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*, 51(7):3899–3909, July 2013.

[42] Elane F Borges, Edson E Sano, and Euzébio 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*, 8(1):083580–083580, 2014.

[43] 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*, 12(5):1061–1065, May 2015.

[44] Qixia Man, Pinliang Dong, and Huadong Guo. Pixel-and feature-level fusion of hyperspectral and lidar data for urban land-use classification. *International Journal of Remote Sensing*, 36(6):1618–1644, 2015.

[45] 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*, 61:281–297, 2007.

[46] 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*, 41(12):2792–2801, Dec 2003.

[47] N. N. Baghdadi, M. El 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*, 9(3):1229–1243, March 2016.

[48] 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)*, pages 5252–5255, July 2016.

[49] Jan Jelének, Veronika Kopačková, Lucie Koucká, and Jan Mišurec. Testing a modified pca-based sharpening approach for image fusion. *Remote Sensing*, 8(10):794, 2016.

[50] Mar Bisquert, Gloria Bordogna, Mirco Boschetti, Pascal Poncelet, and Maguelonne Teisseire. Soft fusion of heterogeneous image time series. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 67–76. Springer, 2014.

[51] M Susan Moran, Daniel C Hymer, Jiaguo Qi, and Edson E Sano. Soil moisture evaluation using multi-temporal synthetic aperture radar (sar) in semiarid rangeland. *Agricultural and Forest Meteorology*, 105(1):69 – 80, 2000.

[52] 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*, 55(7):3885–3899, July 2017.

[53] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.

[54] 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*, 38(2):295–307, Feb 2016.

[55] Massimiliano Gargiulo, Antonio Mazza, Raffaele Gaetano, Giuseppe Ruello, and Giuseppe Scarpa. Fast super-resolution of 20 m sentinel-2 bands using convolutional neural networks. *Remote Sensing*, 11(22), 2019.

[56] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3431–3440, June 2015.

[57] 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*, 26(7):3142–3155, July 2017.

[58] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *Proceedings of European Conference on Computer Vision*, 2014.

[59] 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*, 11(4):042620, 2017.

[60] 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*, 55(10):5585–5599, Oct 2017.

[61] Konstantina Fotiadou, Grigorios Tsagkatakis, and Panagiotis Tsakalides. Deep convolutional neural networks for the classification of snapshot mosaic hyperspectral imagery. *Electronic Imaging*, 2017(17):185–190, 2017.

[62] X. X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, Dec 2017.

[63] Sergio Vitale. A cnn-based pansharpening method with perceptual loss. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3105–3108, July 2019.

[64] X. Chen, S. Xiang, C. Liu, and C. Pan. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 11(10):1797–1801, Oct 2014.

[65] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[66] Ying Li, Haokui Zhang, and Qiang Shen. Spectral-spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing*, 9(1), 2017.

[67] W. Li, G. Wu, and Q. Du. Transferred deep learning for anomaly detection in hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 14(5):597–601, May 2017.

[68] G. Chierchia, D. Cozzolino, G. Poggi, and L. Verdoliva. Sar image despeckling through convolutional neural networks. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5438–5441, July 2017.

[69] Sergio Vitale, Giampaolo Ferraioli, and Vito Pascazio. A new ratio image based CNN algorithm for SAR despeckling. *CoRR*, abs/1906.04111, 2019.

[70] Z. Zhang, H. Wang, F. Xu, and Y. Jin. Complex-valued convolutional neural network and its application in polarimetric sar image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7177–7188, Dec 2017.

[71] C. Bentes, A. Frost, D. Velotto, and B. Tings. Ship-iceberg discrimination with convolutional neural networks in high resolution sar images. In

Proceedings of EUSAR 2016: 11th European Conference on Synthetic Aperture Radar, pages 1–4, June 2016.

[72] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[73] Ilya Sutskever, James Martens, George E Dahl, and Geoffrey E Hinton. On the importance of initialization and momentum in deep learning. *ICML (3)*, 28:1139–1147, 2013.

[74] Dan C. Cireşan, Luca M. Gambardella, Alessandro Giusti, and JÃ¼rgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *In NIPS*, pages 2852–2860, 2012.

[75] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[76] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[77] M. H. Lee, S. B. Lee, Y. D. Eo, S. W. Kim, J.-H. Woo, and S. H. Han. A comparative study on generating simulated Landsat NDVI images using data fusion and regression method – the case of the Korean Peninsula. *Environmental Monitoring and Assessment*, 189(7):333, 2017.

[78] F. Gao, J. Masek, M. Schwaller, and F. Hall. On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance. *IEEE Trans. Geosci. Remote Sens.*, 44(8):2207–2218, Aug 2006.

[79] T. Hilker, M. A. Wulder, N. C. Coops, J. Linke, G. McDermid, J. G. Masek, F. Gao, and J. C. White. A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. *Remote Sensing of Environment*, 113(8):1613 – 1627, 2009.

[80] Jordi Inglada, Marcela Arias, Benjamin Tardy, Olivier Hagolle, Silvia Valero, David Morin, Gérard Dedieu, Guadalupe Sepulcre, Sophie Bontemps, Pierre Defourny, and Benjamin Koetz. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7(9):12356–12379, 2015.

[81] S. t. Wu and S. A. Sader. Multipolarization SAR data for surface feature delineation and forest vegetation characterization. *IEEE Trans. Geosci. Remote Sens.*, GE-25(1):67–76, 1987.

[82] A. R. Huete 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*, 9(15):1–15, 2005.

[83] Jan Haas and Yifang Ban. Sentinel-1a sar and sentinel-2a msi data fusion for urban ecosystem service mapping. *Remote Sensing Applications: Society and Environment*, 8(Supplement C):41 – 53, 2017.

[84] G. Scarpa, M. Gargiulo, A. Mazza, and R. Gaetano. A CNN-Based Fusion Method for Feature Extraction from Sentinel Data. *Remote Sensing*, 10(2), 2018.

[85] Antonio Mazza, Massimiliano Gargiulo, Giuseppe Scarpa, and Raffaele Gaetano. Estimating the ndvi from sar by convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1954–1957. IEEE, 2018.

[86] ESA. ESA Sentinel Application Platform (SNAP) software. `http://step.esa.int/main/toolboxes/snap`, (accessed on 13 December 2017).

[87] THEIA home page. `http://www.theia-land.fr`, (accessed on 13 December 2017).

[88] Olivier Hagolle, Mireille Huc, David Villa Pascual, and Gerard 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*, 7(3):2668–2691, 2015.

[89] Orfeo Toolbox: Temporal gap-filling. `http://tully.ups-tlse.fr/jordi/temporalgapfilling`, (accessed on 13 December 2017).

[90] H. Zhang and B. Huang. Support vector regression-based downscaling for intercalibration of multiresolution satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 51(3):1114–1123, March 2013.

[91] M. S. Moran, D. C. Hymer, J. Qi, and Y. Kerr. Comparison of ERS-2 SAR and Landsat TM imagery for monitoring agricultural crop and soil conditions. *Remote Sensing of Environment*, 79(2):243 – 252, 2002.

[92] E. Santi, et al. The potential of multifrequency SAR images for estimating forest biomass in Mediterranean areas. *Remote Sensing of Environment*, 200(Supplement C):63 – 73, 2017.

[93] Antonio Mazza and Francescopaolo Sica. Deep learning solutions for tandem-x-based forest classification. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 2631–2634, July 2019.

[94] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehamn, S. J. Goetz, T. R. Loveland, and J. R. G. Kommareddy. High-resolution global maps of 21st century forest coverage change. *Science*, 342:850–853, Nov. 2013.

[95] Martin Machala and Lucie ZejdovÃ¡. Forest mapping through object-based image analysis of multispectral and lidar aerial data. *European Journal of Remote Sensing*, 47(1):117–131, 2014.

[96] 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*, 28(4):1923–1938, April 2019.

[97] M. C. Dobson, F. T. Ulaby, and L. E. Pierce. Land-cover classification and estimation of terrain attributes using synthetic aperture radar. *Remote Sensing of Environment*, 51(1):199–214, Jan. 1995.

[98] M. Shimada, T. Itoh, T. Motooka, M. Watanabe, T. Shiraishi, R. Thapa, and R. Lucas. New global forest/non-forest maps from ALOS PALSAR data (2007-2010). *Remote Sensing of Environment*, 155:13–31, 2014.

[99] E. Marcus, Juha M. Engdahl, and Hyyppa. Land-cover classification using multitemporal ers-1/2 insar data. *IEEE Transactions on Geoscience and Remote Sensing*, 41(7):1620–1628, 2003.

[100] F. Sica, A. Pulella, M. Nannini, M. Pinheiro, and P. Rizzoli. Repeat-pass sar interferometry for land cover classification: a methodology using sentinel-1 short-time-series. *Remote Sensing of Environment*, 2019 (In press).

[101] G. Krieger, A. Moreira, H. Fiedler, I. Hajnsek, M. Werner, M Younis, and M. Zink. TanDEM-X: A satellite formation for high-resolution SAR interferometry. *IEEE Trans. Geosci. Remote Sens.*, 45(11):3317–3341, Nov. 2007.

[102] P. Rizzoli, M. Martone, C. Gonzalez, C. Wecklich, D. Borla Tridon, B. Braeutigam, M. Bachmann, D. Schulze, T. Fritz, M. Huber, B. Wessel, G. Krieger, M. Zink, and A. Moreira. Generation and performance assessment of the global tanDEM-X digital elevation model. *ISPRS Journal of Photogr. and Rem. Sens.*, 132:119–139, Aug. 2017.

[103] M. Martone, P. Rizzoli, C. Wecklich, C. González, J.-L. Bueso-Bello, P. Valdo, D. Schulze, M. Zink, G. Krieger, and A. Moreira. The global forest/non-forest map from tandem-x interferometric SAR data. *Remote Sensing of Environment*, 205:352 – 373, 2018.

[104] M. Martone, F. Sica, C. González, J.-L. Bueso-Bello, P. Valdo, and P. Rizzoli. High-resolution forest mapping from tandem-x interferometric data exploiting nonlocal filtering. *Remote Sensing*, 10:1477, 2018.

[105] The TanDEM-X Forest/Non-Forest Map. `https://geoservice.dlr.de/web/maps/tdm:forest`.

[106] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[107] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[108] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[109] M. Martone, P. Rizzoli, and G. Krieger. Volume Decorrelation Effects in TanDEM-X Data. *IEEE Geoscience and Remote Sensing Letters*, 13:1812 – 1816, 2016.

[110] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[111] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[112] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, July 2017.

[113] John Shore and Rodney Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26(1):26–37, 1980.

[114] Yading Yuan, Ming Chao, and Yeh-Chi Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE transactions on medical imaging*, 36(9):1876–1886, 2017.

[115] G. Scarpa, S. Vitale, and D. Cozzolino. Target-adaptive cnn-based pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9):5443–5457, Sep. 2018.

[116] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[117] J. O'Neil-Dunne, S. MacFaden, A. Royar, M. Reis, R. Dubayah, and A. Swatantran. An object-based approach to satewide land cover mapping. In *Proceedings of the ASPRS Annual Conference*, March 2014.

[118] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120(Supplement C):25 – 36, 2012. The Sentinel Missions - New Opportunities for Science.

[119] Charis Lanaras, JosÃ© Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305 – 319, 2018.

[120] Matteo Mura, Francesca Bottalico, Francesca Giannetti, Remo Bertani, Raffaello Giannini, Marco Mancini, Simone Orlandini, Davide Travaglini, and Gherardo 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*, 66:126 – 134, 2018.

[121] Jose Alan A. Castillo, Armando A. Apan, Tek N. Maraseni, and Severino 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*, 134:70 – 85, 2017.

[122] Jan G. P. W. Clevers, Lammert Kooistra, and Marnix M. M. Van den Brande. Using sentinel-2 data for retrieving lai and leaf and canopy chlorophyll content of a potato crop. *Remote Sensing*, 9(5):405, 2017.

[123] Cindy Delloye, Marie Weiss, and Pierre 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*, 216:245 – 261, 2018.

[124] Kaire Toming, Tiit Kutser, Alo Laas, Margot Sepp, Birgot Paavel, and Tiina Nõges. First experiences in mapping lake water quality parameters with sentinel-2 msi imagery. *Remote Sensing*, 8(8):640, 2016.

[125] Markus Immitzer, Francesco Vuolo, and Clement Atzberger. First experience with sentinel-2 data for crop and tree species classifications in central europe. *Remote Sensing*, 8(3):166, 2016.

[126] Martino Pesaresi, Christina Corbane, Andreea Julea, Aneta J. Florczyk, Vasileios Syrris, and Pierre Soille. Assessment of the added-value of sentinel-2 for detecting built-up areas. *Remote Sensing*, 8(4):299, 2016.

[127] Massimiliano Gargiulo, Domenico Antonio Giuseppe Dell'Aglio, Antonio Iodice, Daniele Riccio, and Giuseppe Ruello. A cnn-based super-

resolution technique for active fire detection on sentinel-2 data. *arXiv preprint arXiv:1906.10413*, 2019.

[128] Dimitra Tzelidi, Stavros Stagakis, Zina Mitraka, and Nektarios 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*, 13(1):016514, 2019.

[129] Mingzheng Zhang, Wei Su, Yuting Fu, Dehai Zhu, Jing-Hao Xue, Jianxi Huang, Wei Wang, Jiayu Wu, and Chan Yao. Super-resolution enhancement of sentinel-2 image for retrieving lai and chlorophyll content of summer corn. *European Journal of Agronomy*, 111:125938, 2019.

[130] 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*, 50(2):528–537, Feb 2012.

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

[132] E. Ibarrola-Ulzurrun, L. Drumetz, J. Marcello, C. Gonzalo-Martín, and J. Chanussot. Hyperspectral classification through unmixing abundance maps addressing spectral variability. *IEEE Transactions on Geoscience and Remote Sensing*, 57(7):4775–4788, July 2019.

[133] V. P. Shah, N. H. Younan, and R. L. King. An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets. *IEEE Trans. Geosci. Remote Sens.*, 46(5):1323–1335, May 2008.

[134] Te-Ming Tu, Shun-Chi Su, Hsuen-Chyun Shyu, and Ping S. Huang. A new look at ihs-like image fusion methods. *Information Fusion*, 2(3):177 – 186, 2001.

[135] P.S. Chavez and J.A. Anderson. Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic. *Photogrammetric Engineering and Remote Sensing*, 57(3):295 – 303, 1991.

[136] T. Ranchin and L. Wald. Fusion of high spatial and spectral resolution images: the ARSIS concept and its implementation. *Photogrammetric engineering and remote sensing*, 66(1):49–61, 2000.

[137] D. Fasbender, J. Radoux, and P. Bogaert. Bayesian data fusion for adaptable image pansharpening. *IEEE Trans. Geosci. Remote Sens.*, 46(6):1847–1857, June 2008.

[138] A. Garzelli. Pansharpening of multispectral images based on non-local parameter optimization. *IEEE Trans. Geosci. Remote Sens.*, 53(4):2096–2107, April 2015.

[139] F. Palsson, J.R. Sveinsson, and M.O. Ulfarsson. A new pansharpening algorithm based on total variation. *Geoscience and Remote Sensing Letters, IEEE*, 11(1):318–322, Jan 2014.

[140] 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 Trans. Geosci. Remote Sens.*, 53(5):2565–2586, May 2015.

[141] Qunming Wang, Wenzhong Shi, Zhongbin Li, and Peter M. Atkinson. Fusion of sentinel-2 images. *Remote Sensing of Environment*, 187:241 – 252, 2016.

[142] 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*, XLI-B7:723–730, 2016.

[143] Honglyun Park, Jaewan Choi, Nyunghee Park, and Seokkeun Choi. Sharpening the vnir and swir bands of sentinel-2a imagery through modified selected and synthesized band schemes. *Remote Sensing*, 9(10):1080, 2017.

[144] Mateo Gašparović and Tomislav Jogun. The effect of fusing sentinel-2 bands on land-cover classification. *International Journal of Remote Sensing*, 39(3):822–841, 2018.

[145] N. Brodu. Super-resolving multiresolution images with band-independent geometry of multispectral pixels. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8):4610–4617, Aug 2017.

[146] Charis Lanaras, Jose Bioucas-Dias, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of multispectral multiresolution images from a single sensor. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[147] 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)*, pages 2589–2592, July 2017.

[148] 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.

[149] 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*, pages 4713–4716, July 2018.

[150] H. Xu. Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery. *International journal of remote sensing*, 27(14):3025–3033, 2006.

[151] J. K. Lee 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*, pages 1646–1654, 2016.

[152] Zhou Wang and A.C. Bovik. A universal image quality index. *Signal Processing Letters, IEEE*, 9(3):81–84, March 2002.

[153] Lucien Wald. *Data Fusion. Definitions and Architectures - Fusion of Images of Different Spatial Resolutions*. Presses de l'Ecole, Ecole des Mines de Paris, Paris, France, 2002. ISBN 2-911762-38-X.

[154] Mohammad Fallah Yakhdani and Ali Azizi. *Quality assessment of image fusion techniques for multisensor high resolution satellite images (case study: IRS-P5 and IRS-P6 satellite images)*. na, 2010.

[155] 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*, pages 90–94, May 2003.

[156] 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*, 11:978–989, March 2018.

[157] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolution: Assessing the quality of resulting images. *Photogramm. Eng. Remote Sensing*, pages 691–699, 1997.

[158] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.

[159] Q. Jiang, L. Cao, M. Cheng, C. Wang, and J. Li. Deep neural networks-based vehicle detection in satellite images. In *2015 International Symposium on Bioelectronics and Bioinformatics (ISBB)*, pages 184–187, Oct 2015.

[160] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[161] Chein-I Chang. Spectral information divergence for hyperspectral image analysis. In *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No. 99CH36293)*, volume 1, pages 509–511. IEEE, 1999.

[162] J. Zhou, D. L. Civco, and J. A. Silander. A wavelet transform method to merge landsat tm and spot panchromatic data. *International Journal of Remote Sensing*, 19(4):743–757, 1998.

[163] Alan R Gillespie, Anne B Kahle, and Richard E Walker. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment*, 22(3):343 – 365, 1987.

[164] 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*, 5(1):98–102, Jan 2008.

[165] J. Choi, K. Yu, and Y. Kim. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE Trans. Geosci. Remote Sens.*, 49(1):295–309, Jan 2011.

[166] 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*, 72(5):591–596, 2006.

[167] S. Bhatnagar, L. Gill, S. Regan, O. Naughton, P. Johnston, S. Waldren, and B. Ghosh. Mapping vegetation communities inside wetlands using sentinel-2 imagery in ireland. *Int. J. Appl. Earth Obs.*, 88:102083, 2020.

[168] P. Mondal, S. S. McDermid, and A. Qadir. A reporting framework for sustainable development goal 15: Multi-scale monitoring of forest degradation using modis, landsat and sentinel data. *Remote Sensing of Environment*, 237:111592, 2020.

[169] A. M. Ali, R. Darvishzadeh, A. Skidmore, T. W. Gara, B. O'Connor, C. Roeoesli, M. Heurich, and M. Paganini. Comparing methods for mapping canopy chlorophyll content in a mixed mountain forest using sentinel-2 data. *Int. J. Appl. Earth Obs.*, 87:102037, 2020.

[170] E. W. J. Bergsma and R. Almar. Coastal coverage of esa'sentinel 2 mission. *Advances in Space Research*, 2020.

[171] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage. Sentinel-2 data for land cover/use mapping: A review. *Remote Sensing*, 12(14):2291, 2020.

[172] S. Foga, P. L. Scaramuzza, S. Guo, Z. Zhu, R. D. Dilley Jr, T. Beckmann, G. L. Schmidt, J. L. Dwyer, M. J. Hughes, and B. Laue. Cloud detection algorithm comparison and validation for operational landsat data products. *Remote sensing of environment*, 194:379–390, 2017.

[173] G. Mateo-García, L. Gómez-Chova, J. Amorós-López, J. Muñoz-Marí, and G. Camps-Valls. Multitemporal cloud masking in the google earth engine. *Remote Sensing*, 10(7):1079, 2018.

[174] A. Hollstein, K. Segl, L. Guanter, M. Brell, and M. Enesco. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in sentinel-2 msi images. *Remote Sensing*, 8(8):666, 2016.

[175] Z. Zhu, S. Wang, and C. E. Woodcock. Improvement and expansion of the fmask algorithm: Cloud, cloud shadow, and snow detection for landsats 4–7, 8, and sentinel 2 images. *Remote Sensing of Environment*, 159:269–277, 2015.

[176] S. Mohajerani and P. Saeedi. Cloud-net+: A cloud segmentation cnn for landsat 8 remote sensing imagery optimized with filtered jaccard loss function. *arXiv preprint arXiv:2001.08768*, 2020.

[177] M. Luotamo, S. MetsÃ¤mÃ¤ki, and A. Klami. Multiscale cloud detection in remote sensing images using a dual convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[178] A. Mazza, F. Sica, P. Rizzoli, and G. Scarpa. Tandem-x forest mapping using convolutional neural networks. *Remote Sensing*, 11(24), 2019.

[179] F. Argenti, A. Lapini, T. Bianchi, and L. Alparone. A tutorial on speckle reduction in synthetic aperture radar images. *IEEE Geoscience and Remote Sensing Magazine*, 1(3):6–35, 2013.

[180] C. Deledalle, L. Denis, G. Poggi, F. Tupin, and L. Verdoliva. Exploiting patch similarity for SAR image processing: the nonlocal paradigm. *IEEE Signal Processing Magazine*, 31:69–78, 2014.

[181] G. Fracastoro, E. Magli, G. Poggi, G. Scarpa, D. Valsesia, and L. Verdoliva. Deep learning methods for synthetic aperture radar image despeckling: An overview of trends and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 9, 2021.

[182] G. Chierchia, D. Cozzolino, G. Poggi, and L. Verdoliva. SAR image despeckling through convolutional neural networks. In *IEEE IGARSS*, pages 5438–5441, 2017.

[183] Q. Zhang, Q. Yuan, J. Li, Z. Yang, X. Ma, H. Shen, and L. Zhang. Learning a dilated residual network for SAR image despeckling. *Remote Sensing*, 10:1–18, february 2018.

[184] P. Wang, H. Zhang, and V. M. Patel. SAR image despeckling using a convolutional neural network. *IEEE Signal Processing Letters*, 24(12):1763–1767, Dec 2017.

[185] F. Lattari, B. Leon, F. Asaro, A. Rucci, C. Prati, and M. Matteucci. Deep learning for SAR image despeckling. *Remote Sensing*, 11:1532, june 2019.

[186] D. Cozzolino, L. Verdoliva, G. Scarpa, and G. Poggi. Nonlocal CNN SAR image despeckling. *Remote Sensing*, 12:1006, march 2020.

[187] A. Bordone Molini, D. Valsesia, G. Fracastoro, and E. Magli. Speckle2Void: Deep Self-Supervised SAR Despeckling with Blind-Spot Convolutional Neural Networks. *arXiv:2007.02075*, 2020.

[188] D. Cozzolino, L. Verdoliva, G. Scarpa, and G. Poggi. Nonlocal SAR image despeckling by convolutional neural networks. In *IEEE IGARSS*, pages 5117–5120, 2019.

[189] X. Tang, L. Zhang, and Xiaoli D. SAR image despeckling with a multilayer perceptron neural network. *International Journal of Digital Earth*, 12(3):354–374, 2019.

[190] X. Yang, L. Denis, F. Tupin, and W. Yang. SAR image despeckling using pre-trained convolutional neural network models. In *JURSE*, pages 1–4, 2019.

[191] E. Dalsasso, L. Denis, and F. Tupin. SAR image despeckling by deep neural networks: from a pre-trained model to an end-to-end training strategy. *Remote Sensing*, 12, august 2020.

[192] E. Dalsasso, L. Denis, and F. Tupin. SAR2SAR: a self-supervised despeckling algorithm for SAR images. *arXiv:2006.15037*, 2020.

[193] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *18th SIGSPATIAL Int.l Conference on Advances in Geographic Information Systems*, pages 270–279, 2010.

[194] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila. Noise2noise: Learning image restoration without clean data. In *ICML*, pages 2965–2974, 2018.