remote sensing

MDPI

Article

TanDEM-X Forest Mapping Using Convolutional Neural Networks

Antonio Mazza [1,*], Francescopaolo Sica [2], Paola Rizzoli [2] and Giuseppe Scarpa [1]

[1] Department of Electrical Engineering and Information Technology (DIETI), University Federico II, 80125 Naples, Italy; giscarpa@unina.it

[2] Microwaves and Radar Institute at the German Aerospace Center (DLR), 82234 Wessling, Germany; francescopaolo.sica@dlr.de (F.S.); paola.rizzoli@dlr.de (P.R.)

* Correspondence: antonio.mazza@unina.it; Tel.: +39-081-768-3768

Received: 7 November 2019; Accepted: 9 December 2019; Published: 12 December 2019

Abstract: In this work, we face the problem of forest mapping from TanDEM-X data by means of Convolutional Neural Networks (CNNs). Our 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, we focus 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, our experiments consistently demonstrate the great potential of deep learning in data fusion for information extraction in the context of synthetic aperture radar signal processing and specifically for the task of forest mapping from TanDEM-X images. We compare three state-of-the-art CNN architectures, such as ResNet, DenseNet, and U-Net, obtaining a large performance gain over the baseline approach for all of them, with the U-Net solution being the most effective one.

Keywords: synthetic aperture radar; digital elevation model; image segmemtation; forest classification; target detection; convolutional neural network; data fusion

1. Introduction

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 in 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 Hansen et al. [1], 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 [2] or the use of hyperspectral images [3]. 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 Dobson et al. [4], SAR backscatter data from the ALOS PALSAR mission have been fruitfully applied to global forest mapping in Shimada et al. [5]. 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 [6,7].

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 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 [8]. 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 [9]. 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 References [10,11], where the authors presented a framework for the development of a global TanDEM-X forest/non-forest map [12] as described more in details in Section 2.1.

Deep learning approaches and, in particular, Convolutional Neural Networks (CNNs) 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 [13]. Thanks to the CNNs capability to learn very complex nonlinear relationships from huge labeled datasets with the help of commercial GPUs, unprecedented results have been obtained for many typical tasks such as super-resolution [14,15], segmentation [16], denoising [17], object detection [18,19], classification [20–22], and many others.

Recently, deep learning has started to significantly impact remote sensing applications as well, as testified by the recent survey of Zhu et al. [23]. Established techniques in remote sensing concern, e.g., pansharpening [24,25], vehicle detection [26] with optical images, crop classification [27,28], anomaly detection with hyperspectral data [29], despeckling [30,31], classification [32], or target recognition using SAR data [33,34]. More recently, a multi-temporal SAR-optical fusion method for vegetation index reconstruction was proposed in Scarpa et al. [35].

Motivated by the abovementioned works in References [10,11], in this paper, we now explore 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, our 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 [36], DenseNet [37], and U-Net [38]. For each architecture, we tested different input combinations, 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. Some preliminary results can be found in Mazza and Sica [39].

The paper is organized as follows. Section 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, while the used datasets and experimental results are presented and discussed in Section 4. Finally, the conclusions are drawn in Section 5.

2. Background Concepts

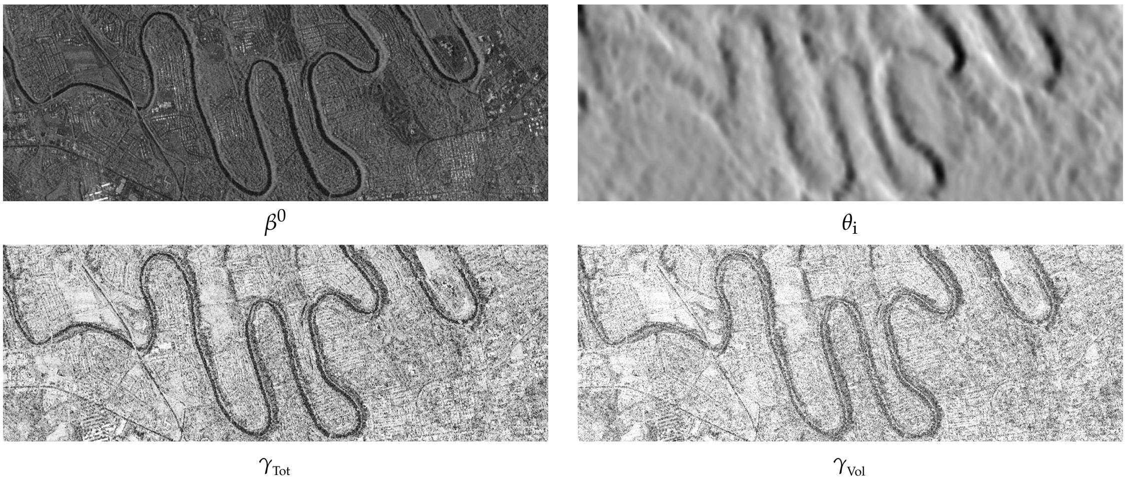
This section provides background concepts that will introduce the reader to the context of the work from both applicative and methodological points of view. In particular, Section 2.1 deals with the definition of some SAR features which can be associated to TanDEM-X data that have been demonstrated to be very effective for forests mapping [11,40]. Then, Section 2.2 recalls general concepts about CNNs, contextualized to the cases of classification and segmentation.

2.1. Baseline Algorithms for Forest Mapping Using TanDEM-X

The work presented in this paper was 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 [10]. 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 [12].

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 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: 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 [10,11] and are easy to compute. Baseline solutions considered in the present work, in fact, use volume correlation coefficient and local incidence angle. For these reasons, we decided to include them as additional input layers. In the following, a suitable definition of them with a related description of the meaning is provided.


β^0

θ_i

γ_Tot

γ_Vol

Figure 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.

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

$$\gamma_\mathrm{Tot} = \frac{|E[xy^*]|}{\sqrt{E[|x|^2]E[|y|^2]}}, \tag{1}$$

where $E[\cdot]$ is the statistical expectation, $*$ is the complex conjugate operator, and $|\cdot|$ is the absolute value. x and y represent the master and slave images, respectively. γ_Tot varies between 0 and 1 m 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 Krieger et al. [8], γ_Tot can be factorized as

$$\gamma_\mathrm{Tot} = \gamma_\mathrm{SNR}\gamma_\mathrm{amb}\gamma_\mathrm{quant}\gamma_\mathrm{az}\gamma_\mathrm{rg}\gamma_\mathrm{Vol}\gamma_\mathrm{Temp}, \tag{2}$$

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 differences (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

$$\gamma_{\text{Vol}} = \frac{\gamma_{\text{Tot}}}{\gamma_{\text{SNR}}\gamma_{\text{amb}}\gamma_{\text{quant}}\gamma_{\text{az}}\gamma_{\text{rg}}\gamma_{\text{Temp}}}. \tag{3}$$

In our specific case, all factors at the denominator but γ_{Temp} have been estimated as described in Martone et al. [10]. γ_{Temp} is equal to one, since we are considering 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 Martone et al. [10] 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

$$h_{\text{amb}} = \frac{\lambda r \sin\theta_{\text{i}}}{B_{\perp}}, \tag{4}$$

where λ is the wavelength, r is the slant range distance, and $B_{\perp}$ is the baseline perpendicular to the line of sight. As it has been demonstrated in Martone et al. [40], 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 Martone et al. [10] 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 dataset 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 Reference [10] proposed in Reference [11], aimed to preserve both global classification accuracy and local precision thanks to the introduction of nonlocal filtering. This latter work represents the starting point of our work and will be therefore referred to as the baseline. The same work [11] 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 makes 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+.

2.2. 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), most of which are convolutional ones. Neuronal weight 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. These features allow for 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 feature 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 and hence summarizing the image content with abstract features. Roughly speaking, pooling helps to move from "where" to "what". On the other hand, nonlinear point-wise activation layers, usually coupled with convolutional or FC layers, allow the overall network to mimic very complex nonlinear functions, therefore expanding the network capacity. According to this paradigm, many state-of-the-art CNN models for classification, e.g., AlexNet [13], VGG [41], and GoogLeNet [42], extract hierarchically related feature layers of decreasing scale, usually shown as a pyramidal stack of which the 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

$$\hat{z}_i = g(\mathbf{s})_i = \frac{e^{s_i}}{\sum_{j=1}^{K} e^{s_j}} \qquad \text{for} \quad i = 1, \ldots, K, \tag{5}$$

with $\mathbf{s} = (s_1, \ldots, s_K)$ as the unnormalized score vector entering the softmax layer and $\hat{z}_i$ as the ith 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 Long et al. [16] by converting the FC stages of classification nets, such as in References [13,41,42], 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 Ronneberger et al. [38], Lin et al. [43].

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, the loss must be differentiable and, more in general, must have a shape that speeds up the gradient descent optimization process. In our 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 [44], as the gradient of their combination has good convergence properties. In the binary case, the cross-entropy loss for the ith input–output training example $\mathbf{t} = (\mathbf{x}, \mathbf{z})$, generalized to the case of pixel-wise classification, is defined as

$$L_{\mathbf{t}}^{\text{bce}} = -\frac{1}{N} \sum_{n=1}^{N} \left[z_n \, \log(\hat{z}_n) + (1 - z_n) \, \log(1 - \hat{z}_n) \right], \tag{6}$$

with $\mathbf{x} \in \mathbb{R}^N$ as the N-pixel input image, $\mathbf{z} = \{z_n\} \in \{0,1\}^N$ as the corresponding binary ground-truth map, and $\hat{\mathbf{z}} = \{\hat{z}_n\} = f_\Phi(\mathbf{x}) \in [0,1]^N$ as 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:

$$L^{\text{bce}} = \mathrm{E}_{\mathbf{t} \sim \text{train}} \left[L_{\mathbf{t}}^{\text{bce}} \right]. \tag{7}$$

The cross-entropy loss works pretty well for classification tasks where the predictors are 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 Equation (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 [45], which is defined as

$$L_{\mathbf{t}}^{\mathrm{J}} = 1 - \frac{\sum_{n=1}^{N} z_n \hat{z}_n}{\sum_{n=1}^{N} \left[(z_n + \hat{z}_n) - z_n \hat{z}_n \right]}, \tag{8}$$

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

3. Proposed Models

In light of the great success of deep learning to solve vision problems, we propose and compare three different CNN models to extract forest maps from TanDEM-X data and/or related products. In particular, the proposed models refer to three different network topologies commonly referred to as residual network (ResNet) [36], dense network (DenseNet) [37], and U-shaped network (U-Net) [38]. 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, we consider several input stacking options 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 noticing that, although CNNs are able to learn features end-to-end, the injection of suitably defined handcrafted features can be beneficial for the network performance (see Masi et al. [24]) as eventually confirmed also by our experiments.

For all models, we have fixed the same loss to minimize through the training process, which is a combination of the cross-entropy (Equation (6)) and the Jaccard (Equation (8)) losses:

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

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.

3.1. TDX-Res

A well-known bottleneck in deep learning is the computational time for training. The ResNet approach [36] is a notable solution recently proposed to mitigate this problem, which has proven also to be effective to limit overfitting. The idea is to concatenate *residual* blocks as done for example in Figure 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

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

where Φ is 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 [46]. More in general, in He et al. [36], it has been shown that 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 [41], GoogLeNet [42], and BN-InceptionNet [47].

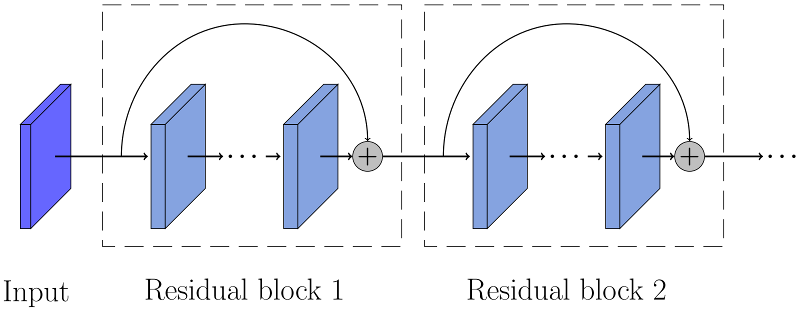


Figure 2. ResNet module.

By following this rationale, we designed our 7-layer residual network, hereinafter referred to as TDX-Res (TanDEM-X ResNet), of which the hyperparameters are gathered in Table 1. All but the last layer are coupled with a Rectified Linear Unit (ReLU) activation [13] 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.

Table 1. TDX-Res hyperparameters. Shape: #kernels × #input channels × kernel support.

Layer	Kernel Shape
Batch Normalization	-
Conv + ReLU	$64 \times b \times (3 \times 3)$
Conv + ReLU	$64 \times 64 \times (3 \times 3)$
Conv + ReLU	$64 \times 64 \times (3 \times 3)$
Conv + ReLU	$64 \times 64 \times (3 \times 3)$
Conv + ReLU	$64 \times 64 \times (3 \times 3)$
Conv + ReLU	$64 \times 64 \times (3 \times 3)$
Conv + Sigmoid	$1 \times 64 \times (1 \times 1)$

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 [37], 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 Figure 3. The ℓth 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. Since k is the number of feature maps produced by each layer, it is easy to verify that the lth layer has $b + k(\ell - 1)$ input feature maps, where b is the number of channels in the input layer. For this reason, the hyperparameter k is referred to as the *growth rate* of the network.

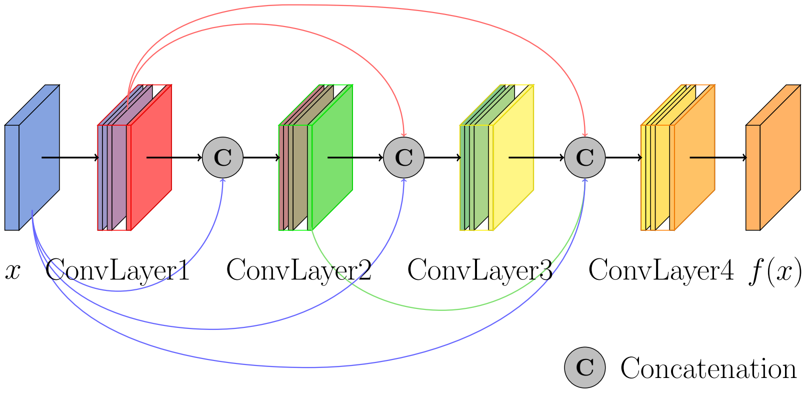


Figure 3. Example of the DenseNet model.

Our 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 Table 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 [47].

Table 2. TDX-Dense hyperparameters. Shape: #kernels × #input channels × kernel support.

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

3.3. TDX-U

The third model for segmentation belongs to the U-Net family originally proposed for medical images [38]. The original idea is to use an encoder–decoder paradigm in order to inheritate well-established CNN classification models pretrained 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, semantic segmentation (see Figure 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 pretrained net, fine-tuned if needed, the decoder is typically trained from scratch. In our case, due to the very specific characteristics of the input images, we decided to train from scratch the whole proposed network on the given dataset, avoiding any transfer learning. Figure 4 refers to our specific implementation, referred to as TDX-U 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 Table 3.

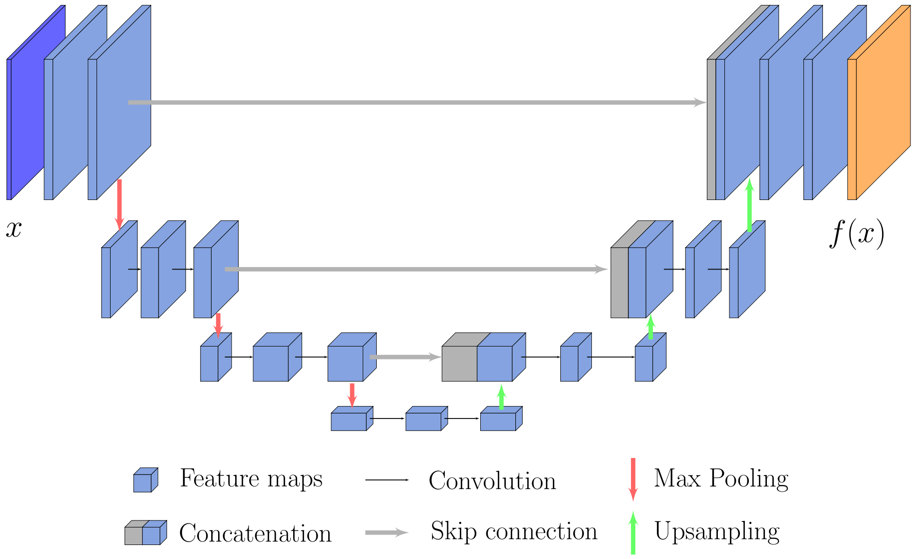


Figure 4. Proposed U-Net structure for forest segmentation from TanDEM-X data.

Table 3. TDX-U hyperparameters. Shape: #kernels × #input channels × kernel support.

Layer	Kernel Shape
Batch Normalization	-
Conv + ReLU	64 × b × (3 × 3)
Conv + ReLU	64 × 64 × (3 × 3)
2 × 2 Max Pooling	-
Conv + ReLU	128 × 64 × (3 × 3)
Conv + ReLU	128 × 128 × (3 × 3)
2 × 2 Max Pooling	-
Conv + ReLU	256 × 128 × (3 × 3)
Conv + ReLU	256 × 256 × (3 × 3)
2 × 2 Max Pooling	-
Conv + ReLU	512 × 256 × (3 × 3)
Conv + ReLU	512 × 512 × (3 × 3)
2 × 2 Upsampling	-
Conv + ReLU	256 × 512 × (3 × 3)
Conv + ReLU	256 × 256 × (3 × 3)
2 × 2 Upsampling	-
Conv + ReLU	128 × 256 × (3 × 3)
Conv + ReLU	128 × 128 × (3 × 3)
2 × 2 Upsampling	-
Conv + ReLU	64 × 128 × (3 × 3)
Conv + ReLU	64 × 64 × (3 × 3)
Conv + Sigmoid	1 × 64 × (1 × 1)

4. Experimental Results

In the following, we will discuss the experimental results obtained with the proposed CNN models in comparison with some reference solutions. First, we provide details about the dataset and training in Section 4.1, and then, we give a summary of involved methods and accuracy measures in Section 4.2. Finally, we provide the numerical accuracy assessment of the compared methods in Section 4.3 and a subjective comparison through the visual inspection of some sample results in Section 4.4.

4.1. The Pennsylvania Dataset and Training Details

The dataset of bistatic TanDEM-X images used for the current work 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 [48] 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, etc.) 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 [49]. This dataset 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.

4.2. Methods and Metrics

The proposed solutions described in detail in Section 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, we have tried different input configurations for each model category by selecting up to 4 input channels selected among SAR amplitude β^0, incidence angle θ_i, interferometric 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 [11] briefly described in Section 2.1, namely baseline and baseline+, we have also implemented random forest classifiers with different input configurations as well as for the proposed methods.

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, we will provide precision, recall, F_1-score, and accuracy, which are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

Precision and recall are usually shown together as they represent the trade-off between the need of catching the target class whenever it occurs (FN = 0, Recall = 1) and that of reducing false alarms (FP = 0, 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).

4.3. Numerical Assessment

For ease of presentation, the numerical results are grouped into two tables. The former (Table 4) gathers a meaningful set of proposed CNN models and is useful to understand several design choices. The latter (Table 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 Figure 5.

Table 4. Forest detection accuracy assessment on the test dataset for different proposed Convolutional Neural Network (CNN) models. Input bands are marked in columns 2–5. Bold numbers correspond to the best for each input configuration, while blue correspond to the best configuration.

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 [39]	×	×	×		**85.04%**	81.38%	83.17%	84.97%
TDX-Dense [39]	×	×	×		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%**

Let us start analyzing our design solutions with the help of Table 4. The models are grouped depending on the input layer setting, which is specified in columns 2–5. For the sake of brevity, a more extensive evaluation with respect to the input configuration is presented for the TDX-Res case only, without loss of generality. A few models correspond to the solutions discussed in Mazza and Sica [39].

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, 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 (Equation (3)). In our 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 Figure 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 Figure 5, with TDX-U located on the upper-triangular image section contrarily to TDX-Res/Dense that lies on the lower-triangular part.

Table 5. Numerical comparison with reference methods.Bold numbers correspond to the best for each input configuration, while blue correspond to the best configuration.

Method	β^0	θ_i	γ_{Tot}	γ_{Vol}	Recall	Prec.	F_1-Score	Acc.
Baseline [11]		×		×	76.17%	60.34%	67.34%	67.72%
Baseline+ [11]		×		×	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%**

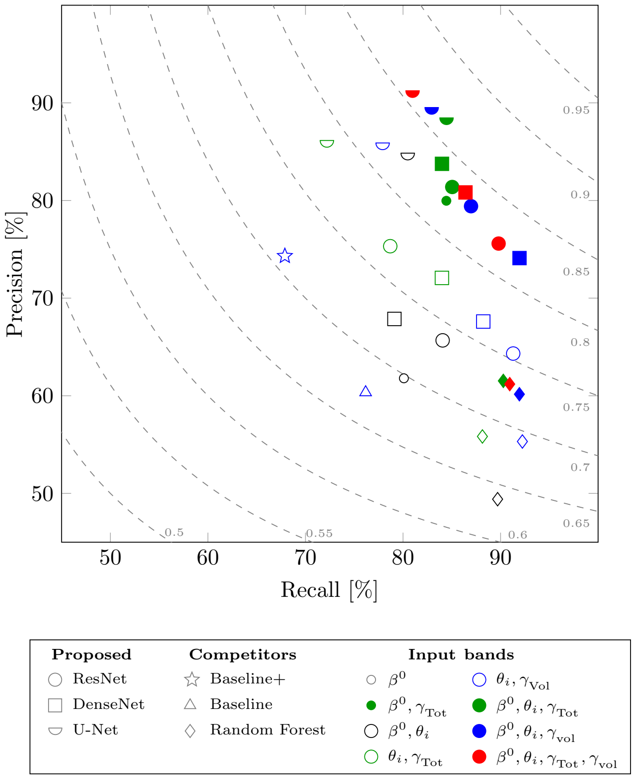


Figure 5. Precision–recall comparison: Dashed lines show F_1-score level curves.

In Table 5, we compare our best architecture, TDX-U, with the reference methods, differentiating the analysis with respect to the input configuration. Since the baseline methods apply to the pair $(\theta_i, \gamma_{\text{Vol}})$ of TanDEM-X by-products, we have also built a CNN model for this case for a fair comparison. 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, we compare with random forest classifiers registering large gains in this case, as well.

4.4. Visual Comparison

Besides numerical evaluation, it is worth to analyse some sample results by visual inspection. Let us first compare our approach with the baseline methods, therefore assuming the same input configuration which is the pair $(\theta_i, \gamma_{\text{Vol}})$. Three samples representative of different contexts are shown in Figure 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 our best model using the same input is in the last column. Observe preliminarly that baseline predicts as forest the water (middle sample) and built-up areas (top and bottom samples). As already underlined above, γ_{Vol} does not allow to discriminate forest from these two classes, and for this reason, in Martone et al. [11], the masked version of baseline, that is baseline+ here, of which the predictions are shown in the third column, was also proposed. 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 γ_{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; therefore, we skip further discussions about it for the sake of brevity.

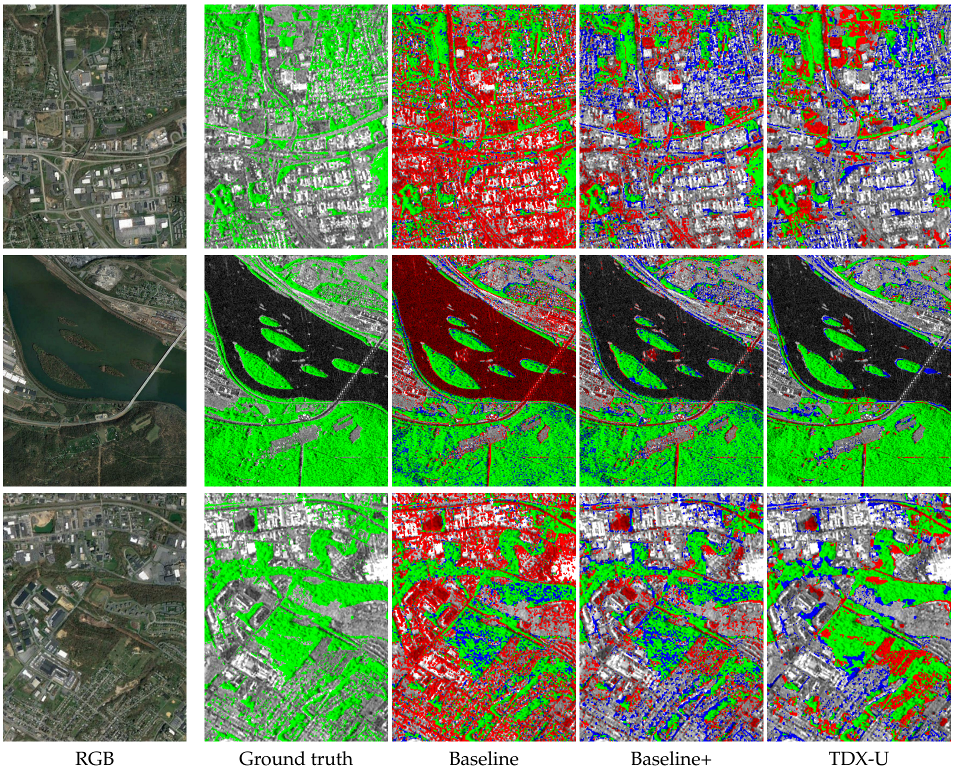


Figure 6. Forest mapping comparison among baseline, baseline+ and TDX-U, using $(\theta_i, \gamma_{\text{Vol}})$ as input. From the left to the right: optical reference, ground truth, baseline, baseline+ and TDX-U methods. Correctly classified forest pixels (TP) are shown in green; non-forest pixels erroneously classified as forest (FP) are in red; and the blue indicates missed forest pixels (FN).

Let us now concentrate on deep learning models and see what happens with other input configurations, focusing for simplicity on TDX-U without loss of generality. In Figure 7, we compare three input settings for the proposed model on the previous running samples. All three configurations include the incidence angle channel θ_i, which is concatenated with γ_{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 we replace γ_{Vol} with γ_{Tot} or simply add the latter to the input. In general, the use of β^0 seems to improve the accuracy on fine details, likely because of the coarser resolution of the other input features (see Figure 1).

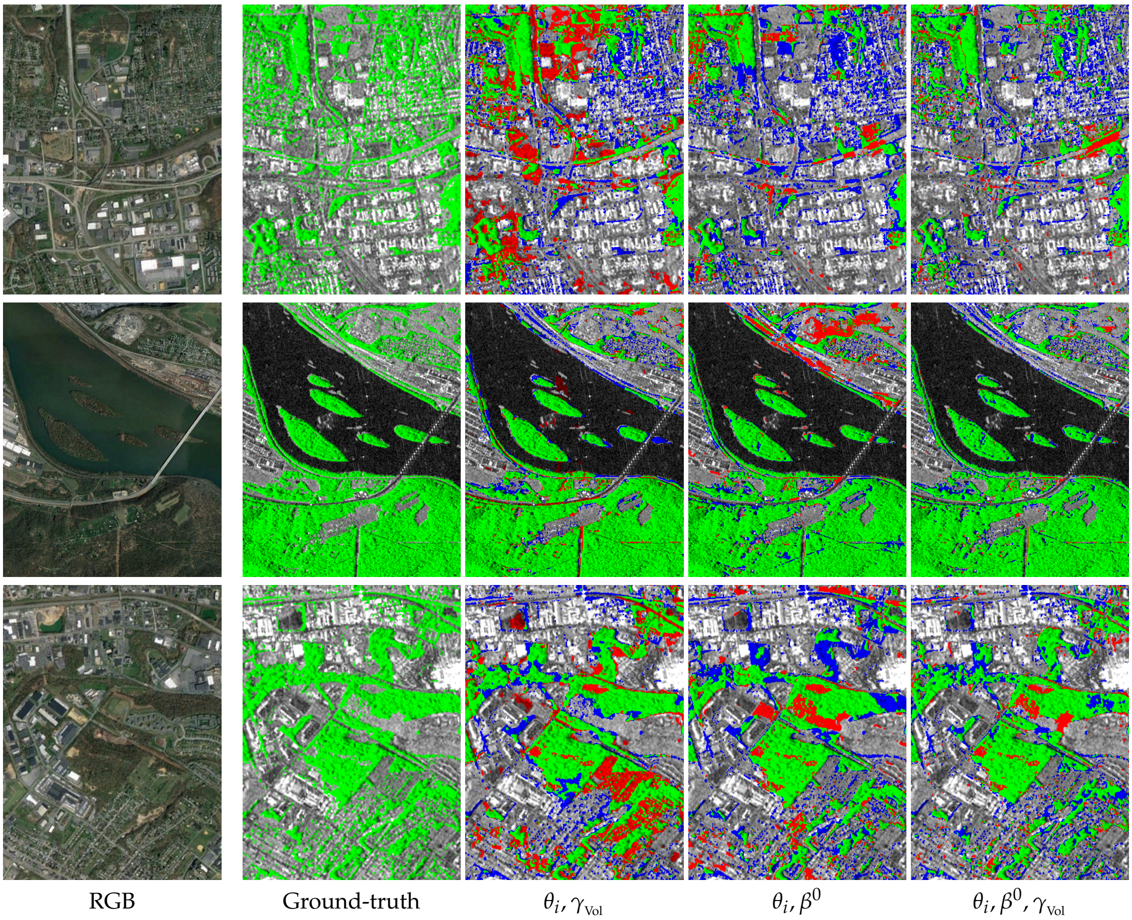

RGB
Ground-truth
$\theta_i, \gamma_{\text{Vol}}$
θ_i, β^0
$\theta_i, \beta^0, \gamma_{\text{Vol}}$

Figure 7. Detection results provided by TDX-U using different input settings. From the left to the right: optical reference, ground truth, TDX-U using $(\theta_i, \gamma_{\text{Vol}})$, (θ_i, β^0) and $(\theta_i, \gamma_{\text{Vol}}, \beta^0)$ as input respectively. Correctly classified forest pixels (TP) are shown in green; non-forest pixels erroneously classified as forest (FP) are in red; and the blue indicates missed forest pixels (FN).

Finally, in Figure 8, we compare the three proposed models in the best input configuration according to the numerical results of Table 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 Table 4.

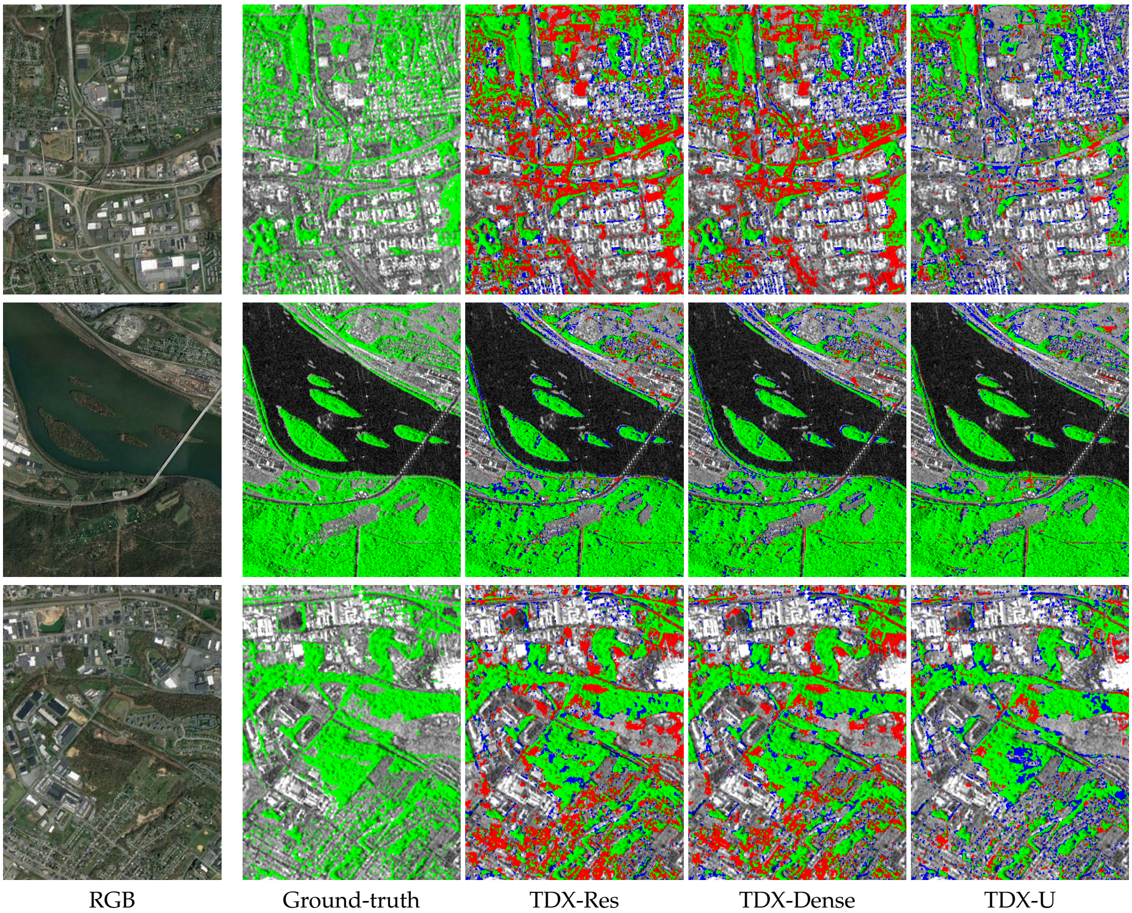


Figure 8. Segmentation results provided by TDX-Res, TDX-Dense, and TDX-U under the best input configuration: $(\beta^0, \theta_i, \gamma_{\text{Tot}})$. From the left to the right: optical reference, ground truth, TDX-Res, TDX-Dense, TDX-U. Correctly classified forest pixels (TP) are shown in green; non-forest pixels erroneously classified as forest (FP) are in red; and the blue indicates missed forest pixels (FN).

5. Conclusions

In this work, we have 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 [10,11], 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. In our opinion, this is due to 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 opens 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. Encouraged by the results shown in this work, our future research will aim to extend the proposed approach to diverse scenarious. In order to do this, a key enhabling 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.

Author Contributions: Conceptualization, A.M., F.S., P.R., and G.S.; methodology, A.M., F.S., P.R., and G.S.; software, A.M.; validation, A.M.; investigation, A.M.; data curation, A.M., F.S., and P.R.; writing—original draft preparation, A.M. and G.S.; writing—review and editing, F.S. and P.R.; supervision, G.S.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hansen, M.C.; Potapov, P.V.; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehamn, S.V.; Goetz, S.J.; Loveland, T.R.; et al. High-resolution global maps of 21st century forest coverage change. *Science* **2013**, *342*, 850–853. [CrossRef] [PubMed]
2. Machala, M.; Zejdová, L. Forest Mapping Through Object-based Image Analysis of Multispectral and LiDAR Aerial Data. *Eur. J. Remote Sens.* **2014**, *47*, 117–131, doi:10.5721/EuJRS20144708. [CrossRef]
3. Hong, D.; Yokoya, N.; Chanussot, J.; Zhu, X.X. An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing. *IEEE Trans. Image Process.* **2019**, *28*, 1923–1938, doi:10.1109/TIP.2018.2878958. [CrossRef] [PubMed]
4. Dobson, M.C.; Ulaby, F.T.; Pierce, L.E. Land-cover classification and estimation of terrain attributes using synthetic aperture radar. *Remote Sens. Environ.* **1995**, *51*, 199–214. [CrossRef]
5. Shimada, M.; Itoh, T.; Motooka, T.; Watanabe, M.; Shiraishi, T.; Thapa, R.; Lucas, R. New global forest/non-forest maps from ALOS PALSAR data (2007–2010). *Remote Sens. Environ.* **2014**, *155*, 13–31. [CrossRef]
6. Engdahl, M.E.; Hyyppa, J.M. Land-cover classification using multitemporal ERS-1/2 InSAR data. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 1620–1628. [CrossRef]
7. Sica, F.; Pulella, A.; Nannini, M.; Pinheiro, M.; Rizzoli, P. Repeat-pass SAR interferometry for land cover classification: A methodology using Sentinel-1 Short-Time-Series. *Remote Sens. Environ.* **2019**, in press. [CrossRef]
8. Krieger, G.; Moreira, A.; Fiedler, H.; Hajnsek, I.; Werner, M.; Younis, M.; Zink, M. TanDEM-X: A Satellite Formation for High-Resolution SAR Interferometry. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3317–3341. [CrossRef]
9. Rizzoli, P.; Martone, M.; Gonzalez, C.; Wecklich, C.; Borla Tridon, D.; Braeutigam, B.; Bachmann, M.; Schulze, D.; Fritz, T.; Huber, M.; et al. Generation and Performance Assessment of the global TanDEM-X digital elevation model. *J. Photogr. Remote Sens.* **2017**, *132*, 119–139. [CrossRef]
10. Martone, M.; Rizzoli, P.; Wecklich, C.; González, C.; Bueso-Bello, J.L.; Valdo, P.; Schulze, D.; Zink, M.; Krieger, G.; Moreira, A. The global forest/non-forest map from TanDEM-X interferometric SAR data. *Remote Sens. Environ.* **2018**, *205*, 352–373, doi:10.1016/j.rse.2017.12.002. [CrossRef]
11. Martone, M.; Sica, F.; González, C.; Bueso-Bello, J.L.; Valdo, P.; Rizzoli, P. High-Resolution Forest Mapping from TanDEM-X Interferometric Data Exploiting Nonlocal Filtering. *Remote Sens.* **2018**, *10*, 1477. [CrossRef]
12. The TanDEM-X Forest/Non-Forest Map. Available online: https://geoservice.dlr.de/web/maps/tdm:forest. (accessed on 10 November 2019).
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1106–1114.
14. Dong, C.; Loy, C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef] [PubMed]
15. Gargiulo, M.; Mazza, A.; Gaetano, R.; Ruello, G.; Scarpa, G. Fast Super-Resolution of 20 m Sentinel-2 Bands Using Convolutional Neural Networks. *Remote Sens.* **2019**, *11*, 2635, doi:10.3390/rs11222635. [CrossRef]
16. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440, doi:10.1109/CVPR.2015.7298965. [CrossRef]

17. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef] [PubMed]
18. Zhang, N.; Donahue, J.; Girshick, R.; Darrell, T. Part-Based R-CNNs for Fine-Grained Category Detection. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
19. Maltezos, E.; Doulamis, N.; Doulamis, A.; Ioannidis, C. Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds. *J. Appl. Remote Sens.* **2017**, *11*, 042620. [CrossRef]
20. Jiao, L.; Liang, M.; Chen, H.; Yang, S.; Liu, H.; Cao, X. Deep Fully Convolutional Network-Based Spatial Distribution Prediction for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5585–5599, doi:10.1109/TGRS.2017.2710079. [CrossRef]
21. Fotiadou, K.; Tsagkatakis, G.; Tsakalides, P. Deep Convolutional Neural Networks for the Classification of Snapshot Mosaic Hyperspectral Imagery. *Electron. Imaging* **2017**, *2017*, 185–190. [CrossRef]
22. Jahan, F.; Awrangjeb, M. Pixel-Based Land Cover Classification by Fusing Hyperspectral and LIDAR Data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, 711–718. [CrossRef]
23. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
24. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by Convolutional Neural Networks. *Remote Sens.* **2016**, *8*, 594, doi:10.3390/rs8070594. [CrossRef]
25. Vitale, S. A CNN-Based Pansharpening Method with Perceptual Loss. In Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2019), Saint Petersburg, Russia, 1–4 July 2019; pp. 3105–3108, doi:10.1109/IGARSS.2019.8900390. [CrossRef]
26. Chen, X.; Xiang, S.; Liu, C.; Pan, C. Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1797–1801, doi:10.1109/LGRS.2014.2309695. [CrossRef]
27. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
28. Li, Y.; Zhang, H.; Shen, Q. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sens.* **2017**, *9*, 67, doi:10.3390/rs9010067. [CrossRef]
29. Li, W.; Wu, G.; Du, Q. Transferred Deep Learning for Anomaly Detection in Hyperspectral Imagery. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 597–601, doi:10.1109/LGRS.2017.2657818. [CrossRef]
30. Chierchia, G.; Cozzolino, D.; Poggi, G.; Verdoliva, L. SAR image despeckling through convolutional neural networks. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 5438–5441, doi:10.1109/IGARSS.2017.8128234. [CrossRef]
31. Vitale, S.; Ferraioli, G.; Pascazio, V. A New Ratio Image Based CNN Algorithm For SAR Despeckling. *arXiv* **2019**, arXiv:1906.04111.
32. Zhang, Z.; Wang, H.; Xu, F.; Jin, Y. Complex-Valued Convolutional Neural Network and Its Application in Polarimetric SAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 7177–7188, doi:10.1109/TGRS.2017.2743222. [CrossRef]
33. Bentes, C.; Frost, A.; Velotto, D.; Tings, B. Ship-Iceberg Discrimination with Convolutional Neural Networks in High Resolution SAR Images. In Proceedings of the EUSAR 2016 11th European Conference on Synthetic Aperture Radar, Hamburg, Germany, 6–9 June 2016; pp. 1–4.
34. Ødegaard, N.; Knapskog, A.O.; Cochin, C.; Louvigne, J. Classification of ships using real and simulated data in a convolutional neural network. In Proceedings of the 2016 IEEE Radar Conference (RadarConf), Philadelphia, PA, USA, 1–6 May 2016. doi:10.1109/RADAR.2016.7485270. [CrossRef]
35. Scarpa, G.; Gargiulo, M.; Mazza, A.; Gaetano, R. A CNN-Based Fusion Method for Feature Extraction from Sentinel Data. *Remote Sens.* **2018**, *10*, 236, doi:10.3390/rs10020236. [CrossRef]
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
37. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

38. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
39. Mazza, A.; Sica, F. Deep Learning Solutions for Tandem-X-Based Forest Classification. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 2631–2634, doi:10.1109/IGARSS.2019.8900441. [CrossRef]
40. Martone, M.; Rizzoli, P.; Krieger, G. Volume Decorrelation Effects in TanDEM-X Data. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1812–1816. [CrossRef]
41. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
42. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arXiv:1409.4842.
43. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
44. Shore, J.; Johnson, R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37. [CrossRef]
45. Yuan, Y.; Chao, M.; Lo, Y.C. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Trans. Med. Imaging* **2017**, *36*, 1876–1886. [CrossRef] [PubMed]
46. Scarpa, G.; Vitale, S.; Cozzolino, D. Target-Adaptive CNN-Based Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5443–5457, doi:10.1109/TGRS.2018.2817393. [CrossRef]
47. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
48. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
49. O'Neil-Dunne, J.; MacFaden, S.; Royar, A.; Reis, M.; Dubayah, R.; Swatantran, A. An object-Based Approach to Satewide Land Cover Mapping. In Proceedings of the ASPRS Annual Conference, Louisville, KY, USA, 23–28 March 2014.