A Dual-Branch Deep Learning Architecture for Multisensor and Multitemporal Remote Sensing Semantic Segmentation

Luca Bergamasco, *Member, IEEE*, Francesca Bovolo, *Senior Member, IEEE*, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Multisensor data analysis allows exploiting heterogeneous data regularly acquired by the many available remote sensing (RS) systems. Machine- and deep-learning methods use the information of heterogeneous sources to improve the results obtained by using single-source data. However, the state-of-the-art methods analyze either the multiscale information of multisensor multiresolution images or the time component of image time series. We propose a supervised deep-learning classification method that jointly performs a multiscale and multitemporal analysis of RS multitemporal images acquired by different sensors. The proposed method processes very-high-resolution (VHR) images using a residual network with a wide receptive field that handles geometrical details and multitemporal high-resolution (HR) image using a 3-D convolutional neural network that analyzes both the spatial and temporal information. The multiscale and multitemporal features are processed together in a decoder to retrieve a land-cover map. We tested the proposed method on two multisensor and multitemporal datasets. One is composed of VHR orthophotos and Sentinel-2 multitemporal images for pasture classification, and another is composed of VHR orthophotos and Sentinel-1 multitemporal images. Results proved the effectiveness of the proposed classification method.**

***Index Terms*—Deep learning (DL) classification, multiresolution, multisensor data, multitemporal images, remote sensing (RS), very-high-resolution (VHR) images.**

I. Introduction

THE analysis of remote sensing (RS) images with very high spatial resolution (VHR) allows improving the land-use classification [1] and urban monitoring [2], [3], [4] performance. However, VHR images provide many geometrical details that are often poorly handled by standard RS methods. Many methodologies model the high spatial resolution by using parcel-based strategies that analyze portions of VHR images and apply an object-based classification [5], [6]. These methods extract regions from the images using segmentation techniques, but they may penalize the classification performance due to segmentation errors.

The exploitation of deep learning (DL) methods allows accurate classification of VHR RS images. DL models, such as convolutional neural networks (CNNs), automatically learn features during the training phase that analyze the spatial information of the input data. The hierarchical structure of these models allows analyzing the input image in a multiresolution way and increases the receptive field of the model. The receptive field determines the DL model analysis area [7], [8]. The larger it is, the larger the area of the image that the model analyzes. The increase of the receptive fields and the automatic learning of spatial features allow more effective extraction of the geometrical information in VHR images than other non-DL methods. Thus, the use of DL-based methods leads to an increase in classification and land-cover mapping performance [9], [10].

Sensors providing VHR images usually acquire a limited number of spectral bands and have a low acquisition frequency due to the relatively small ground-projected field of view (GFOV). Recent VHR sensors reduce the revisit time by acquiring images with multiple acquisition angles. However, when tasks have to deal with image time series analysis, small variability of the image viewing angle within the time series is usually required to limit undesired radiometric differences that do not correspond to changes on the ground [11]. On the contrary, sensors with a lower spatial resolution acquire more spectral bands by keeping similar viewing angles throughout time, have a higher acquisition frequency, and are more easily accessible than VHR images. The latter aspect depends on data-providing policies and not technical reasons. Hence, it is easier to obtain long image time series acquired by sensors with either medium or high spatial resolution rather than VHR images.

Multiresolution DL models process input images with different spatial and spectral resolutions to improve the scene analysis. These methods merge the high spatial resolution of VHR images with the rich spectral information of images with relatively high spectral resolution [12] to increase the classification accuracy. Other approaches extract and analyze multiscale features through the use of pretrained CNN [6], or atrous convolutional layers with various dilation rates [13]. The latter exploits a strategy similar to [14] to extract multiscale features using atrous convolutional layers with heterogeneous receptive fields that are concatenated to perform a multiscale

Manuscript received 9 November 2022; revised 13 January 2023; accepted 4 February 2023. Date of publication 9 February 2023; date of current version 27 February 2023. *(Corresponding author: Francesca Bovolo.)*

Luca Bergamasco and Francesca Bovolo are with the Center for Digital Society, Fondazione Bruno Kessler, 38123 Trento, Italy (e-mail: lbergamasco@fbk.eu; bovolo@fbk.eu).

Lorenzo Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38050 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it).

Digital Object Identifier 10.1109/JSTARS.2023.3243396

analysis. However, the use of single-date RS images does not provide temporal information that helps the discrimination of time-dependent classes. Thus, processing of the RS image time series is required to account for the time component. Many methods exploit random forest (RF) [15], [16] or support vector machine (SVM) [17] to model the temporal information through an image time series. In [18], the authors exploit a time-weighted dynamic time warping (DTW) to select the images within two image time series that better characterize specific classes, and therefore, perform a land-cover and land-use classification. However, these methods are pixel-based and do not consider spatial information. Hence, their performance drops when classes show strong spatial-context correlations, such as in urban areas. These methods exploit handcrafted features to analyze the time component of an image time series. This can lead to an information loss in the time domain. In the last years, recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures were used to model the time information of image time series [19], [20], [21], and learn automatic temporal features during the training. In [22], the authors exploit a deep RNN to process and extract features from an image time series that are used to retrieve land-cover maps. An alternative way to process the temporal information consists in applying CNNs to the temporal domain [23], [24]. 1-D-CNNs can be also used to analyze temporal profiles extracted for each pixel [24]. However, these methods learn and extract temporal features but analyze the spatial information using only handcrafted spatial features.

Classification methods usually process data acquired by a single sensor. However, some applications require information acquired by multiple sensors with heterogeneous characteristics. Thus, methodologies for analyzing multisensor images with different spatial, spectral, and temporal resolutions should be considered. Some methods merge the temporal information of image time series acquired by heterogeneous sensors by extracting temporal features using RNN-based models and selecting the most informative ones through an attention mechanism [25], [26]. Spatiotemporal features are extracted and combined to segment image time series with homogeneous spatial resolutions using an attention mechanism to learn the most informative features to discriminate the time-based classes [27]. These methods do not properly model the spatial information of the images so they achieve suboptimal results. Some methods combine the spatial context features extracted from CNNs analyzing high-resolution (HR) images with the temporal one processing image time series with RNNs [28] or separately process the temporal and spatial information using 3-D-CNNs and 2-D-CNNs, respectively [29]. In [30], the authors process the spatial information with the temporal one through the use of CNNs to extract spatial features from multisensor data and convolutional RNN (ConvRNN) to process both spatial and temporal information. These methods achieve good results but handle homogeneous spatial resolution only. However, the discrimination of some classes requires a multiresolution analysis of heterogeneous sensors to generalize and improve the classification and perform a multiscale analysis [31]. To this purpose, some methods exploit a deep neural network with two branches [32], [33]: one analyzes the spatial information of VHR images, while the other the spectral information of multispectral data. However, these methods do not consider temporal information. In [34], the authors perform segmentation of multisensor image time series representing crop fields using models based on U-Net [35] exploiting a convolutional LSTM (ConvLSTM) or a 3-D-CNN to jointly analyze the spatial and temporal information. The authors in [36] extend the latter by exploiting bidirectional ConvLSTM to analyze the temporal information in both directions.

Summarizing state-of-the-art (SoA) methods analyze multisensor data single-date multiresolution RS image or multitemporal data with similar spatial resolution. Therefore, in this article, we propose a DL method for supervised classification that analyzes multisensor data with heterogeneous properties in spatial, spectral, and temporal resolutions. The proposed method allows analyzing single-date and multitemporal images acquired by different sensors with heterogeneous spatial, spectral, and temporal resolutions using a DL model having two input branches. The first one is a deep residual network (ResNet) [37] having a wide receptive field that properly models VHR geometrical details, while the other processes the spatial and temporal information of multitemporal images acquired by a sensor with a lower spatial resolution than the previous input data using a 3-D-CNN. The spatial and temporal output features of the two branches are merged and processed by a ResNet-based decoder to obtain the land-cover map. In this way, the classification considers both the multiresolution and multitemporal information and retrieves accurate results.

This article has the following outline. Section II describes the methodology. Section III presents the experimental settings and the results. Finally, Section IV concludes this article.

II. Multiresolution and Multitemporal Classification

The proposed method aims to obtain a classification map $\overline{Y}$ by analyzing a VHR image $X_{\text{VHR}} \in \mathbb{R}^{w_{\text{VHR}} \times h_{\text{VHR}} \times b_{\text{VHR}}}$ acquired at time t_{VHR} and an image time series $X_{\text{TS}} = \{X_1, \ldots, X_i, \ldots, X_I\}$ composed by I images $X_i \in \mathbb{R}^{w_{\text{TS}} \times h_{\text{TS}} \times b_{\text{TS}}}$ $(i = 1, \ldots, I)$ acquired between t_1 and t_I by a sensor with different spatial, spectral, and temporal resolutions with respect to X_{VHR}. X_{VHR} and X_{TS} represent the same geographical area. Assuming that no semantic changes occurs in the analyzed scene (e.g., new buildings and missing objects), X_{VHR} can be acquired at any time with respect to the interval $[t_1, t_I]$. We assume the availability of labeled training samples Y that provide class information about the geographical area represented in X_{VHR} and X_{TS}. A labeled training dataset X is retrieved by sampling N patches corresponding of X_{VHR} and X_{TS} according to Y.

VHR images provide many geometrical details but limited spectral and temporal information. Thus, the modeling of time-dependant classes is challenging. On the contrary, HR image time series have a better tradeoff in terms of spectral and temporal resolutions allowing the modeling of time-dependant classes, but lower spatial resolution than VHR images. Hence, they have fewer geometrical details, and the modeling of classes characterized by small-size elements is less accurate. The proposed architecture combines the spatial context information of VHR

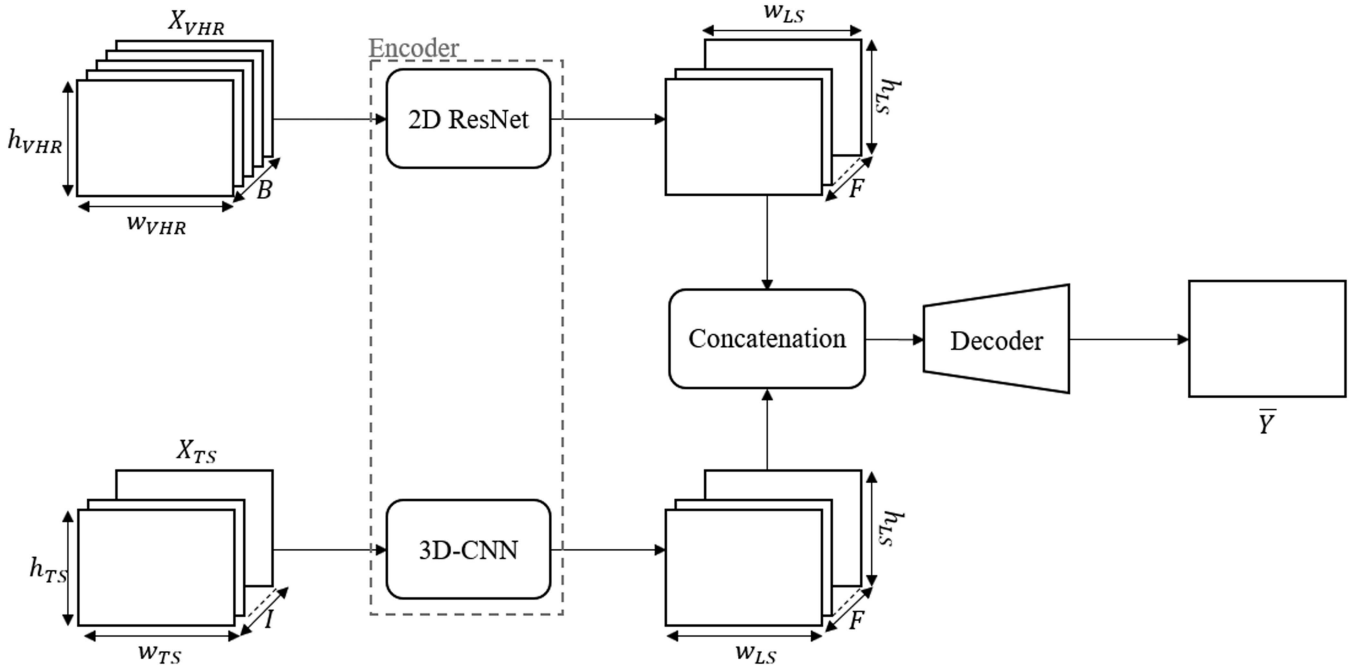


Fig. 1. Block scheme of the proposed DL model using 3-D convolutional layers.

images with the spectral and temporal information provided by the HR image time series to accurately model classes varying over time and characterized by small-size elements. It relies on a deep-learning (DL) model based on ResNet [37] that allows dealing with inputs having different spatial resolutions and different characteristics due to the heterogeneous acquisition sensors. The DL model analyzes them in two branches (see Fig. 1). We define the two branches to obtain the same number of feature maps with the same spatial size. Thus, the output features of the two branches can be easily concatenated and have a balanced contribution to the final classification. The decoder processes the concatenated features to obtain the classification map.

A. VHR Image Analysis With a the Residual Neural Network

The first branch analyzes X_{VHR}, which has the highest spatial resolution and many geometrical details to be processed. We analyze the spatial information of X_{VHR} using an encoder $f(X_{\text{VHR}})$ composed of multiple residual blocks. As in [37], each residual block exploits the bottleneck design to reduce the training time and increase the depth of the model by controlling the number of model parameters. Each block is composed of three 2-D convolutional layers with a kernel size of 1×1, $k \times k$, and 1×1, where k is the kernel size of the middle convolutional layer in the residual block (see Fig. 2). The first and third layers reduce and increase the dimensions, whereas the second one analyzes the spatial information. The depth of the encoder depends on the spatial resolution of X_{VHR}. The higher the X_{VHR} spatial resolution, the deeper the encoder. When the spatial resolution of the input image is very high, the sensor captures many geometrical details, and neighboring pixels show a strong spatial correlation [38], [39]. To effectively process the

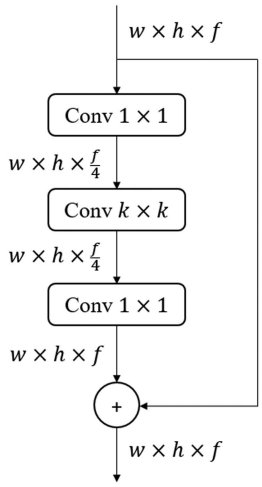


Fig. 2. Block scheme of the residual block.

spatial context information of the image and produce informative feature maps, the model requires a large receptive field. Thus, we use multiple stride convolutional layers to compress the spatial information of X_{VHR}. The output of the encoder is defined by

$$Z_{\text{VHR}} = f(X_{\text{VHR}}) \tag{1}$$

where $Z_{\text{VHR}} \in \mathbb{R}^{w_{\text{LS}} \times h_{\text{LS}} \times F}$ provides F-dimensional latent space (LS) feature maps with a dimension of $w_{\text{LS}} \times h_{\text{LS}}$.

B. Image Time-Series Analysis With 3-D Convolutional Layers

The processing of X_{TS} with size $w_{\text{TS}} \times h_{\text{TS}} \times I \times b_{\text{TS}}$, where I is the number of images in the time series and b_{TS} is the spectral band number, is challenging since it provides both

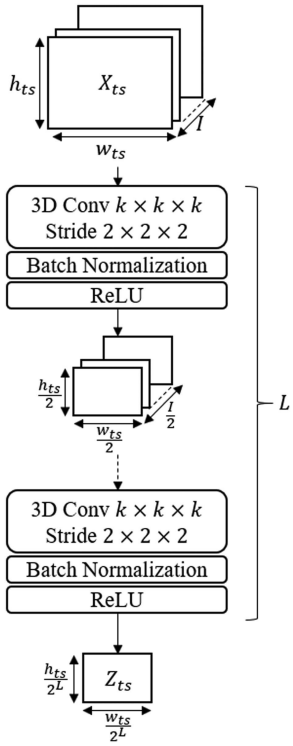


Fig. 3. Block scheme of the 3-D-CNN.

spatial and temporal information. 2-D convolutional layers effectively analyze the spatial information of each image in the time series but do not model the temporal relationship between them. 2-D convolutional layers sum the contribution of each image in the time series, losing the temporal information. To process both the spatial and temporal information of X_{TS}, we use 3-D convolutional layers [40] with 3-D kernels. Thus, we process X_{TS} with a 3-D-CNN) with L layers (see Fig. 3). Each 3-D convolutional layer compresses the spatial and temporal information to increase the receptive field of the model in both the dimensions and obtain informative features. Since the spatial resolution of X_{TS} is lower than the X_{VHR} one, we need a smaller receptive field in the spatial dimensions. On the other side, the more images are in the time series, more layers are needed to model the temporal relationship. Therefore, the number of 3-D convolutional layers is a tradeoff that depends on the spatial resolution and the number of images in the time series. We use stride 3-D convolutional layers that compress the spatial and temporal information and enlarge the receptive field in all dimensions. Let us assume, for simplicity, that the spatial size of the image time series is $w_{\mathrm{TS}} = h_{\mathrm{TS}}$, the LS feature map size is $w_{\mathrm{LS}} = h_{\mathrm{LS}}$, the padding operation is applied to avoid dimension reduction due to the convolution operation, and all the division terms of the following equations are divisible. The spatial dimension of the 3-D-CNN output (i.e., w_{LS}) is defined as a function of the relationship between the spatial size of the input image time series and the spatial stride of the 3-D convolutional layers

$$w_{\mathrm{LS}} = \frac{w_{\mathrm{TS}}}{(s_w)^L} \tag{2}$$

where s_w is the stride value of the spatial dimensions in the 3-D convolutional layer. We can retrieve the number of layers L that are needed to achieve the spatial dimension w_{LS} by exploiting (2).

$$(s_w)^L = \frac{w_{\mathrm{TS}}}{w_{\mathrm{LS}}}$$

$$L = \log_{s_w}\left(\frac{w_{\mathrm{TS}}}{w_{\mathrm{LS}}}\right). \tag{3}$$

Since in the proposed method the data time dimension is reduced to 1 to obtain feature maps providing both spatial and temporal information, we can use an equation similar to (2) to retrieve the number of layers L that are needed to fully process the temporal information

$$\frac{I}{(s_t)^L} = 1$$

$$L = \log_{s_t}(I) \tag{4}$$

where s_t is the stride value of the temporal dimension in the 3-D-convolutional layers. Thus, the total number of 3-D convolutional layers for the joint analysis of spatial and temporal information is given by

$$L = \max\left(\log_{s_w}\left(\frac{w_{\mathrm{TS}}}{w_{\mathrm{LS}}}\right), \log_{s_t}(I)\right). \tag{5}$$

The output of the encoder is defined by

$$Z_{\mathrm{TS}} = g(X_{\mathrm{TS}}) \tag{6}$$

where $g(.)$ is the function of the 3-D-CNN that provides F feature maps of size $w_{\mathrm{LS}} \times h_{\mathrm{LS}}$. The spatial resolution of $Z_{\mathrm{TS}} \in \mathbb{R}^{w_{\mathrm{LS}} \times h_{\mathrm{LS}} \times F}$ has to be harmonized with the one of Z_{VHR} to merge the processed data of the two sensors.

C. Merge of the Data in the Two Branches and Classification

We design the two branches of the DL model to obtain from each of them F-dimensional LS feature maps (i.e., Z_{VHR} and Z_{TS}) with the same spatial dimensions $w_{\mathrm{LS}} \times h_{\mathrm{LS}}$ that are merged to process the information derived from the two sensors together. In this way, the output of the two branches provides a balanced contribution to the final classification, and the information of one sensor does not dominate the other. We concatenate Z_{VHR} and Z_{TS} to process them through a decoder and obtain the land-cover map. After the concatenation, we obtain LS feature maps $Z = \mathrm{concat}([Z_{\mathrm{VHR}}, Z_{\mathrm{TS}}])$ with dimensions $w_{\mathrm{LS}} \times w_{\mathrm{LS}} \times 2F$.

We process $Z \in \mathbb{R}^{w_{\mathrm{LS}} \times w_{\mathrm{LS}} \times 2F}$ using residual blocks composed of atrous convolutional layers to further enlarge the receptive field of the model without compressing the spatial information of Z [41]. We then process Z through a decoder composed of residual blocks using deconvolutional layers to decompress the spatial information of the features and increase their spatial dimensions to achieve $w_{\mathrm{VHR}} \times h_{\mathrm{VHR}}$. A 2-D convolutional layer with C kernels, where C is equal to the number of classes, of 1×1 ends the decoder and retrieves the land-cover map $\overline{Y} \in \mathbb{R}^{w_{\mathrm{VHR}} \times h_{\mathrm{VHR}} \times C}$ using a softmax activation function.

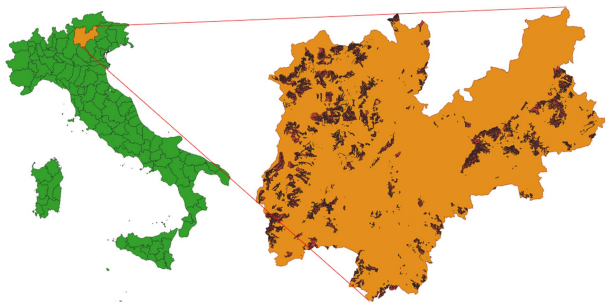


Fig. 4. Study areas (in red) of the two datasets in the Trentino region, Italy.

The output of the DL model is defined by

$$\begin{aligned}\overline{Y} &= d(Z) = d(\text{concat}([Z_{\text{VHR}}, Z_{\text{TS}}])) \\ &= d(\text{concat}([f(X_{\text{VHR}}), g(X_{\text{TS}}])))\end{aligned} \tag{7}$$

where $d(.)$ defines the decoder function.

We use a cross-entropy loss function to train the whole DL model as it proved its effectiveness in many SoA classification methods [42], [43], [44], [45]. We add to the loss function a weight contribution to regularize the model weights during the training and improve the classification performance. The final loss function is given by

$$\text{CE} = -\frac{1}{N}\sum_{n}^{N} Y_n \log(\overline{Y_n}) + \lambda \|W\|_2^2 \tag{8}$$

where λ is a constant that controls the weight regularization and $\|W\|_2^2$ represents the L2 normalization of the model weights W.

III. Experimental Settings and Results

In this section, we introduce the datasets used to test the proposed method, present the experimental setup, and discuss the results.

A. Description of Dataset

To test the proposed method, we used two datasets composed of images acquired over the Trentino region in Italy. Both datasets exploit a VHR orthophoto acquired by aircraft. The first dataset combined the orthophoto with an image time series acquired by Sentinel-1, whereas the second dataset uses multitemporal Sentinel-2 images together with the orthophoto.

1) Dataset Combining VHR Orthophotos and Sentinel-1 Image Time Series: We tested the proposed method by analyzing the temporal information using a synthetic-aperture-radar image time series. For this dataset, we combined one VHR orthophoto with an image time series acquired by Sentinel-1 [see Fig. 5(c)]. The airborne orthophoto was acquired in September 2017 and was organized in 470 tiles. They have a spatial resolution of $20\,\text{cm/pixel}$ that was down-scaled to $1\,\text{m/pixel}$ to make the computation lighter. These tiles were acquired in the red, green, blue, and near-infrared spectral bands [see Fig. 5(a)]. We exploited a DTM to infer the slope degrees using [46]. The Sentinel-1 image time series is VV polarized and has a spatial resolution of $10\,\text{m/pixel}$. It was acquired from January to December 2017 over the Trentino region in Italy. We aimed to discriminate the following six classes.

1) Fruit trees: Areas with various fruit trees types (e.g., apple trees, vineyards, etc.).
2) Artificial areas: Areas with artificial objects, such as houses and farmhouses.
3) Pastures.
4) Forest.
5) Water bodies.
6) Impervious areas: Either rocky or steep areas.

TABLE I
Proportion Between the Classes of the Dataset Using Sentinel-1 Images

Class	Percentage (%)
Fruit trees	0.0008
Artificial areas	0.43
Pastures	56.43
Forest	29.03
Water bodies	0.32
Impervious area	13.79

The class distribution is unbalanced (see Table I), which may lead to some classification errors due to the training bias. We sampled the tiles and the image time series according to the reference map and obtained a dataset composed of 271 155 patches with a spatial size of 120×120 for the VHR orthophoto and 12×12 for the image time series. The dataset was split into a training set with 219 791 patches, a validation set with 24 352 patches, and a test set with 27 012 patches.

2) Dataset Combining VHR Orthophotos and Multitemporal Sentinel-2 Images: This dataset was created for a project aiming to classify the quality of Trentino pasture areas (see Fig. 4). It included five classes: three classes are about the quality of the pasture, and two are about the presence of other kinds of vegetation. The five classes can be divided into the pasture ones and no pasture ones. The pasture ones composed by the following .

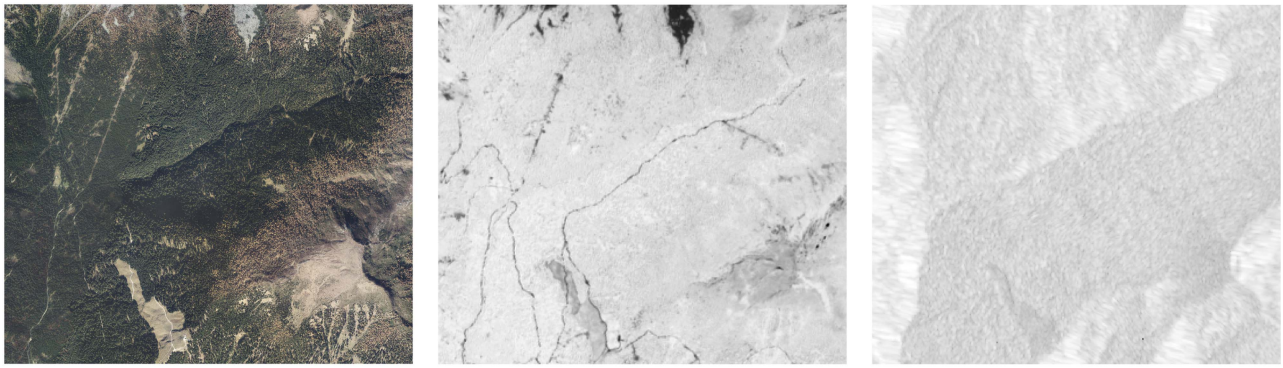
1) Level 0: 100% of grass.
2) Level 20: 80% of grass; shrub and rock presence until the 20%.
3) Level 50: 50% of grass; shrub and rock presence until the 50%.

And the no pasture ones include the following.

1) Forest.
2) Impervious areas: Lands that cannot be used for pasture because of the steep slope or the strong shrub and rock presence.

Note that the five classes are similar and all related to the grass or the presence of vegetation, so their discrimination is challenging both in the spectral and temporal domains. As we can observe in Table II, the proportion between the five classes is slightly unbalanced. This can lead to a bias training that may cause classification errors.

The dataset was composed of a VHR airborne orthophoto with the same spectral bands and acquisition characteristics of the previous dataset and multitemporal HR satellite images acquired by Sentinel-2. The VHR orthophoto provides the



(a) (b) (c)

Fig. 5. Three example areas of the test dataset showing (a) RGB airborne images, (b) NDVI obtained by the one of the Sentinel-2 image in the time series, (c) one of the Sentinel-1 image of the time series.

TABLE II
PROPORTION BETWEEN THE CLASSES OF THE DATASET USING SENTINEL-2 IMAGES

Class	Percentage (%)
Level 0	12.6
Level 20	14.79
Level 50	28.12
Forest	35.52
Impervious area	8.97

spatial context information but not the spectral and temporal one to model the pasture phenological behavior over time. On the contrary, Sentinel-2 images provide lower spatial context information (they have a spatial resolution of $10\,\mathrm{m/pixel}$) to discriminate the mentioned classes accurately since they are highly fragmented. Instead, they bring the temporal information since the time series includes images acquired from June 2018 to September 2018, which is the most significant period for the phenological characterization of alpine pastures. We exploit ten spectral bands of the Sentinel-2 images by excluding the one at $60\,\mathrm{m/pixel}$ that provide atmospheric information only. For each acquisition month, we chose the atmospherically corrected Sentinel-2 images with the least cloud-cover percentage. It is worth noting that even if the airborne data and the multitemporal images were acquired almost one year apart, this was not a problem from the application point of view since no relevant changes were expected in pastures during this timeframe. We randomly sampled only the areas where the classes had no cloud coverage on any of the four dates of the Sentinel-2 images. We sampled patches with spatial dimensions of 120×120 for the VHR orthophotos and 12×12 for the Sentinel-2 multitemporal images. These patch sizes guaranteed that, in each sample, both kinds of data represented the same geographical area. We obtained a dataset composed of 94 048 patches split into training with 76 245 patches, validation with 8 403 patches, and test with 9 400 patches.

B. Design of Experiments

We set up the first branch $f(X_{\mathrm{VHR}})$ of the deep learning (DL) model a ResNet [37] with ten residual blocks having a central convolutional layer with $k = 3$ and as last layer a 2-D convolutional layer with a kernel size of 3×3. To have the same number of feature maps of $f(X_{\mathrm{VHR}})$, $g(X_{\mathrm{TS}})$ was composed by two 3-D convolutional layers with kernel sizes $3 \times 3 \times 3$ and $3 \times 3 \times 2$ and strides $1 \times 1 \times 2$ and $2 \times 2 \times 2$. The decoder was composed of seven residual blocks with $k = 3$ preceded by an upsampling layer that decompresses the spatial information. The last 2-D convolutional layer had a 1×1 kernel size and a filter number equal to the number of classes (see Table III). It exploited a softmax activation function to retrieve the land-cover map. Every convolutional layer (2-D and 3-D) was followed by a rectified linear unit activation function and batch normalization.

We trained the proposed DL model in a supervised way for $E = 300$ using the Adam optimizer [47] with a learning rate equal to 10^{-4}. Due to hardware constraints, we used a batch size equal to 80. We set the weight decay $\lambda = 5 \cdot 10^{-5}$, which is a value used in the SoA to improve the generalization capability of the model [48]. We tested the effectiveness of the proposed method using the two datasets in three experiments.

1) *Experiment 1*—We compared the proposed method with SoA methods. Both the SoA and proposed method were trained using the hyperparameters mentioned previously.
2) *Experiment 2*—We observed the contribution provided by the temporal information for land-cover mapping by comparing the results of the proposed DL model with the ones of a model analyzing only a VHR image. This model was composed of the same layers of $f(.)$ and $d(.)$, but it did not include $g(.)$ to process the temporal information and the concatenation step. Thus, we compared the outcome of the ResNet processing only the VHR images with the proposed method processing both VHR images and multitemporal data.
3) *Experiment 3*—We observed the contribution of the spatial information provided by the VHR orthophotos by comparing the proposed-method results with the outcomes of a model composed only of $g(.)$ and $d(.)$. Hence, this model processed only the temporal information of multitemporal images.

For *Experiments 2* and *3*, we used the same hyperparameters of *Experiment 1*.

TABLE III
PROPOSED DL MODEL STRUCTURE

	Layer/ Block(# filters)	Kernel size	Stride	Atrous rate	Output size
$f(X_{\text{VHR}})$	Input(X_{VHR})	-	-	-	$120 \times 120 \times 5$
	Conv 2-D	7×7	1×1	1×1	$120 \times 120 \times 32$
	Batch Norm.	-	-	-	$120 \times 120 \times 32$
	Max Pooling	2×2	2×2	1×1	$60 \times 60 \times 32$
	3xRes. blocks (32,32,128)	3×3	1×1	1×1	$60 \times 60 \times 128$
	4xRes. blocks (64,64,256)	3×3	2×2	1×1	$30 \times 30 \times 256$
	3xRes. blocks (128,128,512)	3×3	2×2	1×1	$15 \times 15 \times 512$
	Conv 2-D	3×3	2×2	1×1	$8 \times 8 \times 64$
	Batch Norm.	-	-	-	$8 \times 8 \times 64$
	Bilinear interp.	-	-	-	$6 \times 6 \times 64$
$g(X_{\text{TS}})$	Input(X_{TS})	-	-	-	$12 \times 12 \times I \times 10$
	Conv 3-D	$3 \times 3 \times 3$	$1 \times 1 \times 2$	1×1	$12 \times 12 \times 2 \times 32$
	Batch Norm.	-	-	-	$12 \times 12 \times 2 \times 32$
	Conv 3-D	$3 \times 3 \times 2$	$2 \times 2 \times 2$	1×1	$6 \times 6 \times 1 \times 64$
	Batch Norm.	-	-	-	$6 \times 6 \times 1 \times 64$
	Reshape	-	-	-	$6 \times 6 \times 64$
	Concat.	-	-	-	$6 \times 6 \times 128$
$d(Z)$	Atrous Res.B. (128,128,512)	3×3	1×1	1×1	$6 \times 6 \times 512$
	Atrous Res.B. (128,128,512)	3×3	1×1	2×2	$6 \times 6 \times 512$
	Upsampling	-	2×2	-	$12 \times 12 \times 512$
	Conv 2-D	3×3	1×1	1×1	$12 \times 12 \times 64$
	Batch Norm.	-	-	-	$12 \times 12 \times 64$
	Upsampling	-	2×2	-	$24 \times 24 \times 64$
	4xRes. blocks (64,64,256)	3×3	1×1	1×1	$24 \times 24 \times 256$
	Upsampling	-	2×2	-	$24 \times 24 \times 64$
	3xRes. blocks (32,32,128)	3×3	1×1	1×1	$48 \times 48 \times 128$
	Upsampling	-	2×2	-	$96 \times 96 \times 128$
	Conv 2-D	5×5	1×1	1×1	$96 \times 96 \times 32$
	Batch Norm.	-	-	-	$96 \times 96 \times 32$
	Bilinear interp.	-	-	-	$120 \times 120 \times 32$
	Conv 2-D	1×1	1×1	1×1	$120 \times 120 \times 5$

We evaluated the performance of the land-cover mapping methods by considering the average overall accuracy (OA) [(true positives+true negatives)/number of labeled pixels], F1-score, and kappa coefficient (κ). In *Experiments 2* and *3*, we provided the confusion matrix and the F1 score per class.

C. Experiment 1: SoA Comparison

We compared the proposed method results with the SoA using the datasets defined in Section III-A. The comparison with SoA techniques was performed against the following:

1) a DL model analyzing both spatial and temporal information using convolutional and recurrent layers (TWINNS) [30], and two models analyzing multiresolution RS images;
2) one using a CNN and stacked autoencoder (DMIL) [32];
3) the other one using two 2-D-CNNs (MrFusion) [33].

Since 2) analyzes only images with a homogeneous spatial resolution, we upsampled the Sentinel-2 images to have the spatial dimensions equal to the orthophotos.

1) SoA Comparison Using the Dataset VHR Orthophotos and Sentinel-1 Image Time Series: Using the dataset composed of VHR orthophotos and Sentinel-1 image time series, the proposed method achieved comparable results with the MrFusion approach and improved the classification performance with respect to the other SoA methods (see Table IV) (e.g., improvement of 5.77% with respect to TWINNS). The quantitative results proved that the joint analysis of the spatial context and temporal information performed by 3-D convolutional layers is more efficient in modeling the information provided by image time series and improved the classification performance with respect to SoA methods splitting the spatial context and temporal information analysis, such as DMIL (i.e., improvement of 2.7% with 10.7 M of parameters less). Although MrFusion obtained comparable results with respect to the proposed method, it was more computationally demanding since it required almost six times more training parameters (i.e., 18.4 M instead of 3.2 M). In most cases, the proposed method obtained better classification results than the SoA methods with fewer training parameters (e.g., 3.2 M). Thus, it was more effective than the SoA methods in the analysis of multimodal multiresolution image time series.

TABLE IV
OAS, F1-SCORE, κ COEFFICIENT, AND NUMBER OF PARAMETERS OF THE PROPOSED METHOD AND SOA METHODS USING A DATASET COMPOSED OF VHR ORTHOPHOTOS AND SENTINEL-1 IMAGE TIME SERIES

Method	OA	F1-score	κ	N. par.
TWINNS [30]	$84.67 \pm 6.59\%$	$72.31 \pm 16.14\%$	0.69 ± 0.14	13.0M
DMIL [32]	$87.74 \pm 4.1\%$	$74.72 \pm 16.35\%$	0.74 ± 0.12	13.9M
MrFusion [33]	$\mathbf{91.01 \pm 3.33\%}$	$\mathbf{77.23 \pm 16.49\%}$	$\mathbf{0.81 \pm 0.11}$	18.4M
Proposed	$90.44 \pm 3.72\%$	$76.86 \pm 16.53\%$	0.8 ± 0.11	**3.2M**

TABLE V
OAS, F1-SCORE, κ COEFFICIENT, AND NUMBER OF PARAMETERS OF THE PROPOSED METHOD AND SOA METHODS USING A DATASET COMPOSED OF VHR ORTHOPHOTOS AND MULTITEMPORAL NDVI

Method	OA	F1-score	κ	N. par.
TWINNS [30]	$80.69 \pm 6.56\%$	$80.89 \pm 6.6\%$	0.7 ± 0.1	13.2M
DMIL [32]	$77.31 \pm 7.48\%$	$77.53 \pm 7.42\%$	0.65 ± 0.1	17.9M
MrFusion [33]	$80.92 \pm 6.57\%$	$81.16 \pm 6.48\%$	0.71 ± 0.099	18.4M
Proposed	$\mathbf{81.25 \pm 6.7\%}$	$\mathbf{81.35 \pm 6.66\%}$	$\mathbf{0.71 \pm 0.097}$	**3.0M**

The qualitative results confirmed the quantitative ones. The discrimination between the pasture and forest areas was the most common classification error [e.g., Fig. 6(b), the bottom part of Fig. 7(d), the center of Fig. 8(b)]. However, the proposed method alleviated this problem by jointly analyzing the spatial context and temporal information [e.g., Fig. 6(e), the bottom part of Fig. 7(e), the center of Fig. 8(e)], although it overestimated water bodies.

2) SoA Comparison Using the Dataset VHR Orthophotos and Multitemporal Sentinel-2 Images: Using the dataset composed of VHR orthophotos and Sentinel-2 image time series, the proposed method achieved better classification performance than the SoA ones (see Table V). MrFusion [33] and TWINNS [30] were the SoA methods with the best F1 score (i.e., 80.92% and 80.69%, respectively). TWINNS analyzed the spatial and temporal information using a convolutional RNN with an attention mechanism, whereas MrFusion used a 2-D-CNN. The proposed method slightly improved the F1 score with respect to TWINNS and MrFusion. This proved how the joint analysis of spatial and temporal information of 3-D convolutional layers used in the proposed method improved the classification performance with

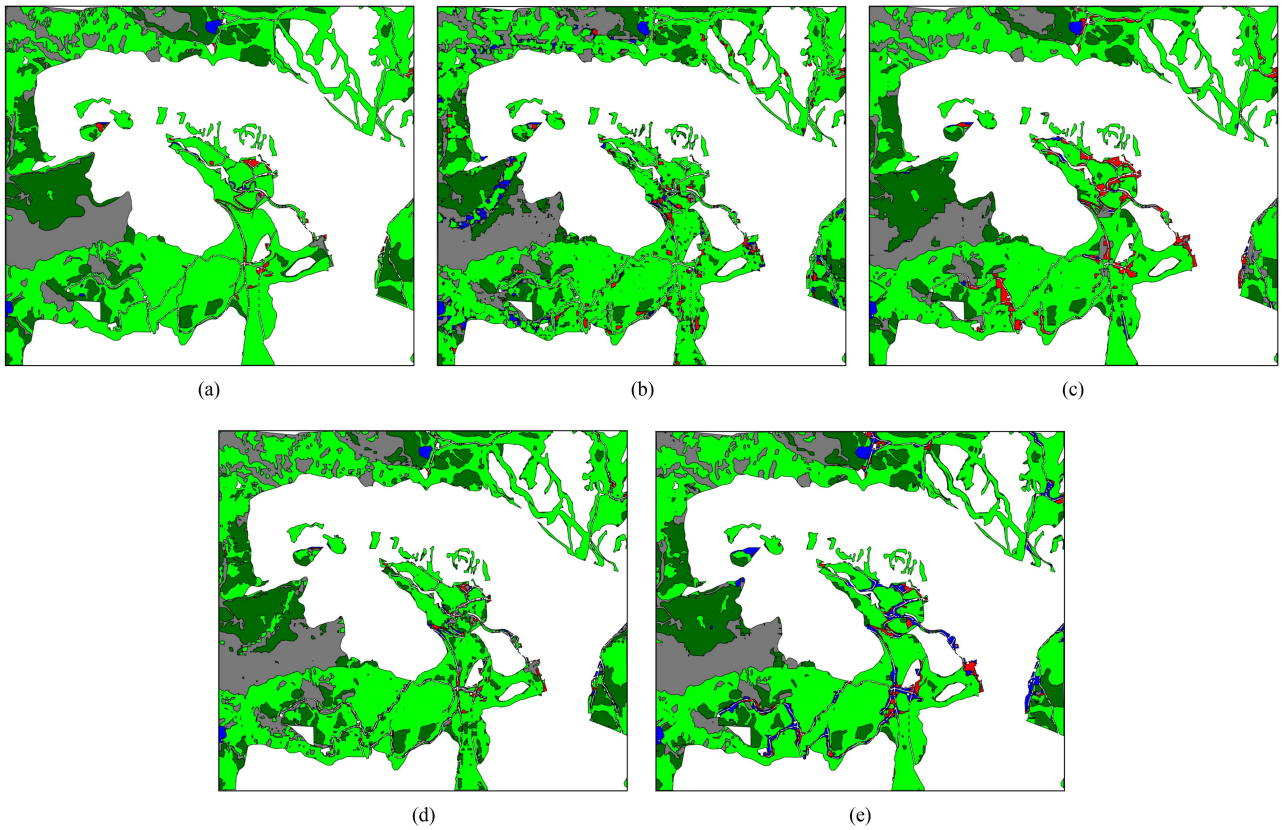


Fig. 6. Comparison between (a) reference maps of the first Sentinel-1 dataset area and the classification maps retrieved using (b) TWINNS model [30], (c) DMIL model [32], (d) MrFusion model [33], and (e) proposed method using Sentinel-1 image time series (yellow is Fruit trees, red is Artificial areas, green is Pastures, dark green is Forest, blue is Water bodies, gray is Impervious areas, and white is masked pixels).

respect to 2-D-CNN used in MrFusion, which processed only the spatial information. The joint analysis of the spatial and temporal resolution performed by the proposed method using the 3-D-CNN improved the F1-score of 3.82% with respect to DMIL [32] that processed only the temporal information using an SAE. The proposed method was less computationally demanding (i.e., 3.0 M parameters) than the SoA ones (i.e., 17.9 M for DMIL, 18.4 M for MrFusion, and 13.2 M for TWINNS) and it achieved the best classification performance with the lowest number of parameters (see Table V).

The qualitative results confirmed the quantitative ones. The proposed method better discriminated the difference between pasture and nonpasture classes than the SoA ones (e.g., the right side of the classification maps in Fig. 9). The inaccurate classification between Level 50, Forest, and Impervious areas was the most common error [e.g., right part of Fig. 9(c) and (d)] that was alleviated by the proposed method [see Fig. 9(e)]. We can observe how the analysis of both spatial and temporal information performed by 3-D convolutional layers performed by the proposed method achieved better results [e.g., the upper part of Fig. 10(e) and the bottom left part of Fig. 11(e)] in the three pasture classes discrimination than the 2-D convolutional layers that learned only spatial features and not temporal ones [see Figs. 10(d) and 11(d)].

D. Experiment 2: Effectiveness of the Temporal Information

We compared the results obtained by processing a single-date VHR image and using the proposed method to jointly process a VHR orthophoto and multitemporal data (i.e., Sentinel-1 image time series or multitemporal Sentinel-2 image) from the two datasets to observe if the inclusion of the temporal component in the spatial analysis improves the classification accuracy with respect to the use of single-date VHR image. We proved that the temporal information analysis performed by the proposed method improved the model classification performance with respect to the method processing only single-date VHR images in both datasets (see Table VI). The Sentinel-1 temporal information increased the OA of 4.86%, whereas the addition of multitemporal Sentinel-2 images improved the pasture classification OA of 1.81%. The temporal information analysis of the Sentinel-2 images allowed to better model the phenological behavior of the classes over time. The inclusion of the Sentinel-1 temporal information in the classification analysis improved the method capability in the minority class discrimination, such as

TABLE VI
OAs, F1-Score, and κ of a DL Model Analyzing Only Single-Date Images and the Proposed Method With the Two Datasets

# DS	Method	OA	F1-score	κ
1	VHR images	$85.58 \pm 6.88\%$	$73.36 \pm 17.89\%$	0.71 ± 0.13
	VHR images + Sentinel-1 time series	$90.44 \pm 3.72\%$	$76.86 \pm 16.53\%$	0.8 ± 0.11
2	VHR images	$79.44 \pm 7.21\%$	$79.88 \pm 6.92\%$	0.69 ± 0.11
	VHR images + multi-temp. Sentinel-2	$\mathbf{81.25 \pm 6.7\%}$	$\mathbf{81.35 \pm 6.66\%}$	$\mathbf{0.71 \pm 0.097}$

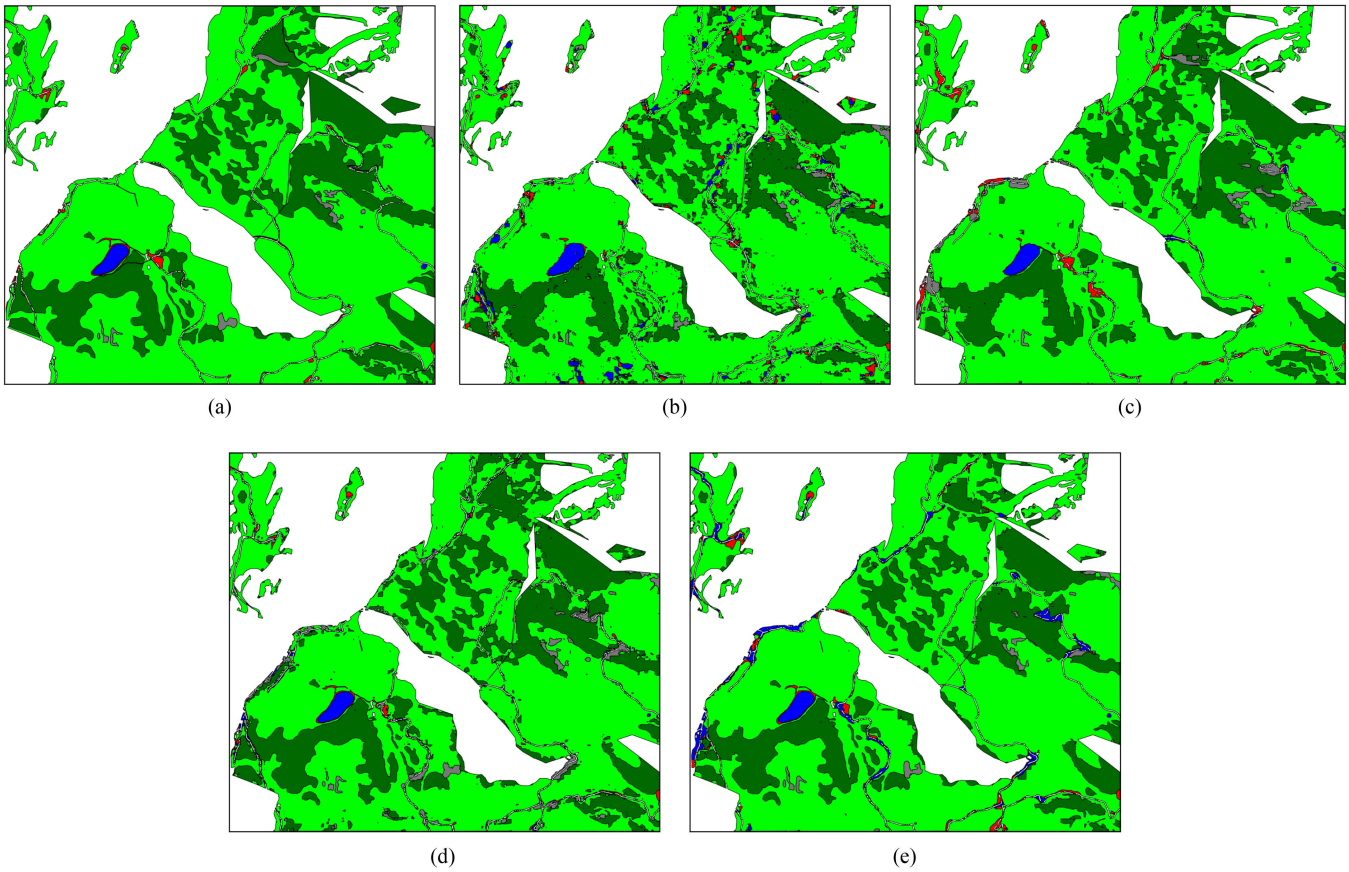


Fig. 7. Comparison between (a) reference maps of the second Sentinel-1 dataset area and the classification maps retrieved using (b) the TWINNS model [30], (c) DMIL model [32], (d) MrFusion model [33], and (e) proposed method using Sentinel-1 image time series (Yellow is Fruit trees, red is Artificial areas, green is Pastures, dark green is Forest, blue is Water bodies, gray is Impervious areas, and white is masked pixels).

fruit trees, and further increased the classification performance with the majority classes [see Fig. 12(a) and (c)]. The proposed method tended to misclassify minority classes (i.e., artificial areas and water bodies) [see Fig. 12(c)] since few labeled samples belonging to these classes were in the training set leading to overfitting them. The F1 score of each class increased by adding the Sentinel-2 temporal information. We can observe that the proposed method classified better pasture and no pasture class than the one using only single-date VHR images [see Fig. 13(a)]. However, the proposed method underestimated the class Level 20 and overestimated Levels 0 and 50 [see Fig. 13(c)] due to their similarity.

The qualitative results (see Figs. 14 and 15) confirmed the quantitative ones and showed the classification performance improvement by adding temporal to the spatial context analysis. In the first dataset, the Sentinel-1 temporal information analysis allowed us to characterize better the impervious areas and water bodies [e.g., left side of Fig. 15(b) and (c)]. In the second dataset, we can observe that the Sentinel-2 temporal information allowed us to discriminate better similar classes, such as Pasture Level 50 and Forest [e.g., the right side of Fig. 15(b) and (c)]. It also improved the classification of similar classes, such as Level 0, 20, and 50 [e.g., the bottom side of Fig. 15(b) and (c)].

TABLE VII
OAs, F1-Score, and κ of a DL Model Analyzing Sentinel-1 Image Time Series, and the Proposed Method Using the Two Datasets

# DS	Method	OA	F1-score	κ
1	Sentinel-1 time series	$73.79 \pm 8.47\%$	$64.15 \pm 16.14\%$	0.49 ± 0.16
	VHR images + Sentinel-1 time series	$\mathbf{90.44 \pm 3.72\%}$	$\mathbf{76.86 \pm 16.53\%}$	$\mathbf{0.8 \pm 0.11}$
2	Multi-temp. Sentinel-2	$72.98 \pm 7.88\%$	$73.42 \pm 7.91\%$	0.59 ± 0.12
	VHR images + multi-temp. Sentinel-2	$\mathbf{81.25 \pm 6.7\%}$	$\mathbf{81.35 \pm 6.66\%}$	$\mathbf{0.71 \pm 0.097}$

E. Experiment 3: Effectiveness of Multiscale Information

We compared the results obtained by analyzing only Sentinel-1 image time series or multitemporal Sentinel-2 images with the proposed method ones using the datasets defined in Section III-A to prove the improvement obtained by processing multiresolution images. The use of multiresolution images sharply improved the classification performance with respect to multitemporal data analysis in both datasets (see Table VII). Sentinel-1 image time series and multitemporal Sentinel-2 images did not provide many geometrical details due to their spatial resolution. Thus, the

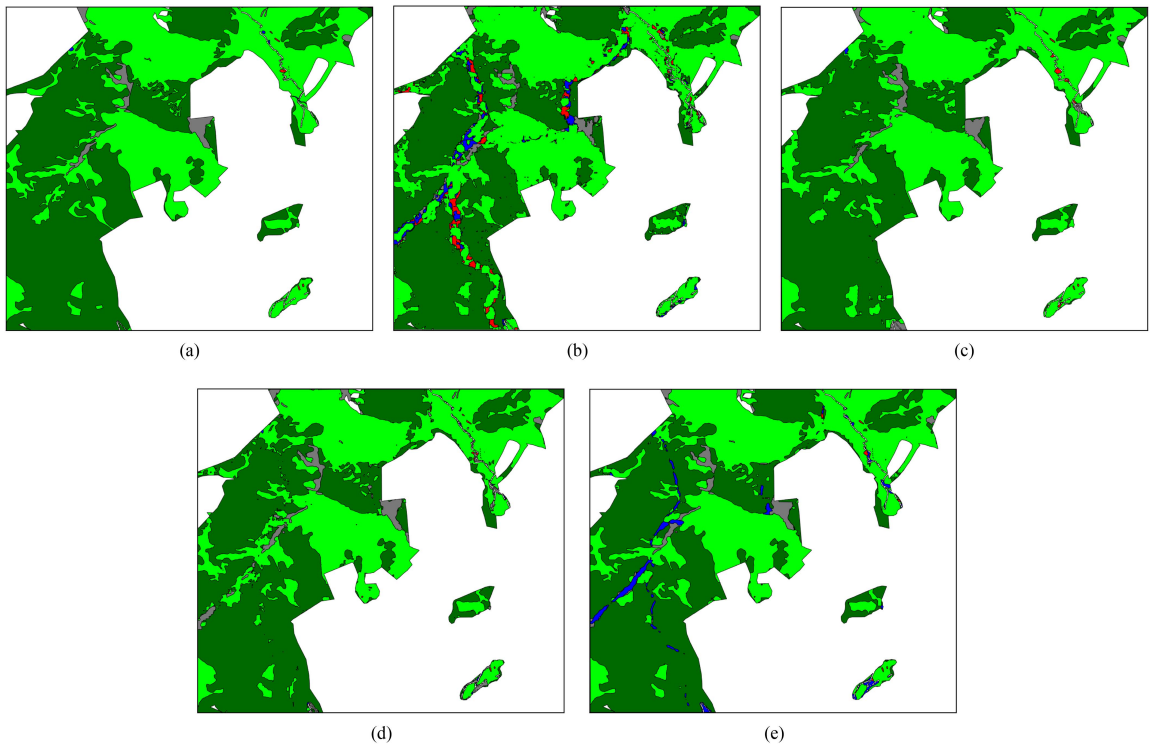


Fig. 8. Comparison between (a) reference maps of the third Sentinel-1 dataset area and the classification maps retrieved using (b) TWINNS model [30], (c) DMIL model [32], (d) MrFusion model [33], and (e) proposed method using Sentinel-1 image time series (yellow is Fruit trees, red is Artificial areas, green is Pastures, dark green is Forest, blue is Water bodies, gray is Impervious areas, and white is masked pixels).

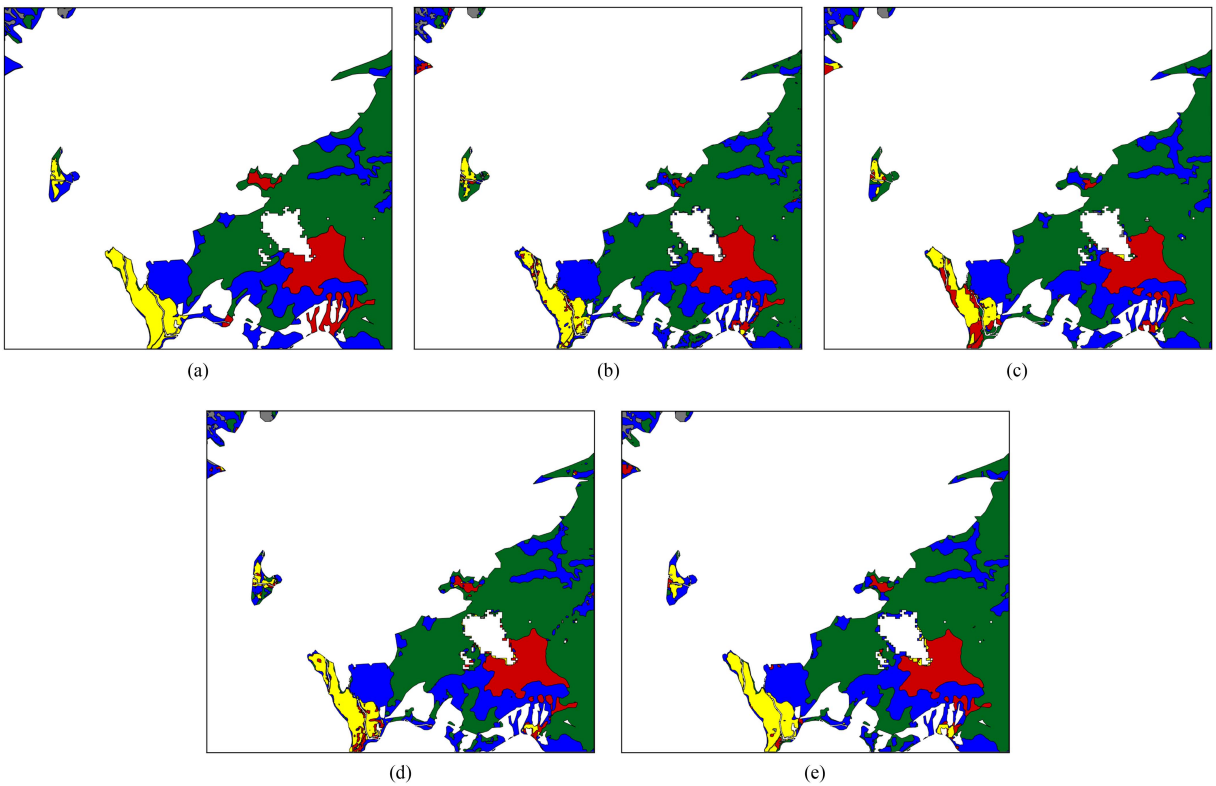


Fig. 9. Comparison between (a) reference maps of the first Sentinel-2 dataset area, and the classification maps retrieved using (b) TWINNS model [30], (c) DMIL model [32], (d) MrFusion model [33], and (e) proposed method (yellow is Level 0, red is Level 20, blue is Level 50, dark green is Forest, gray is Impervious areas, and white is masked pixels).

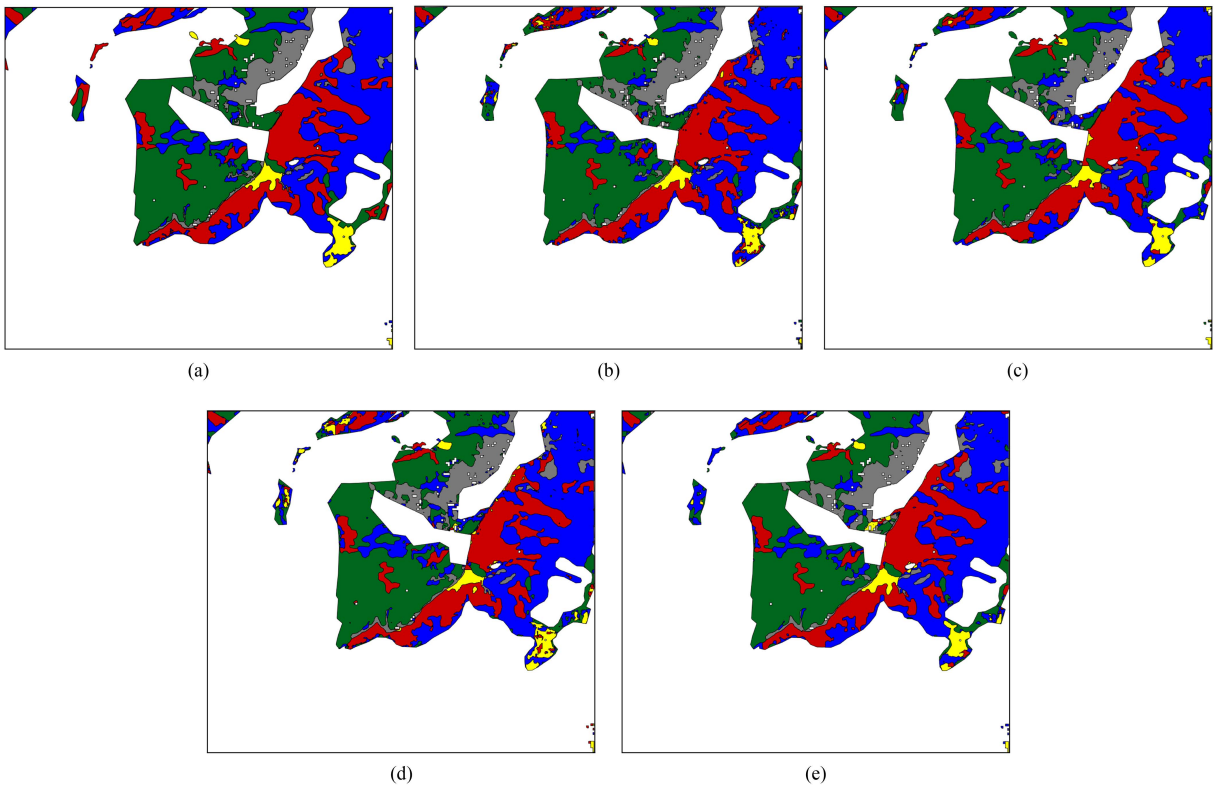


Fig. 10. Comparison between (a) reference maps of the second NDVI dataset area, and the classification maps retrieved using (b) TWINNS model [30], (c) DMIL model [32], (d) MrFusion model [33], and (e) proposed method (yellow is Level 0, red is Level 20, blue is Level 50, dark green is Forest, gray is Impervious areas, and white is masked pixels).

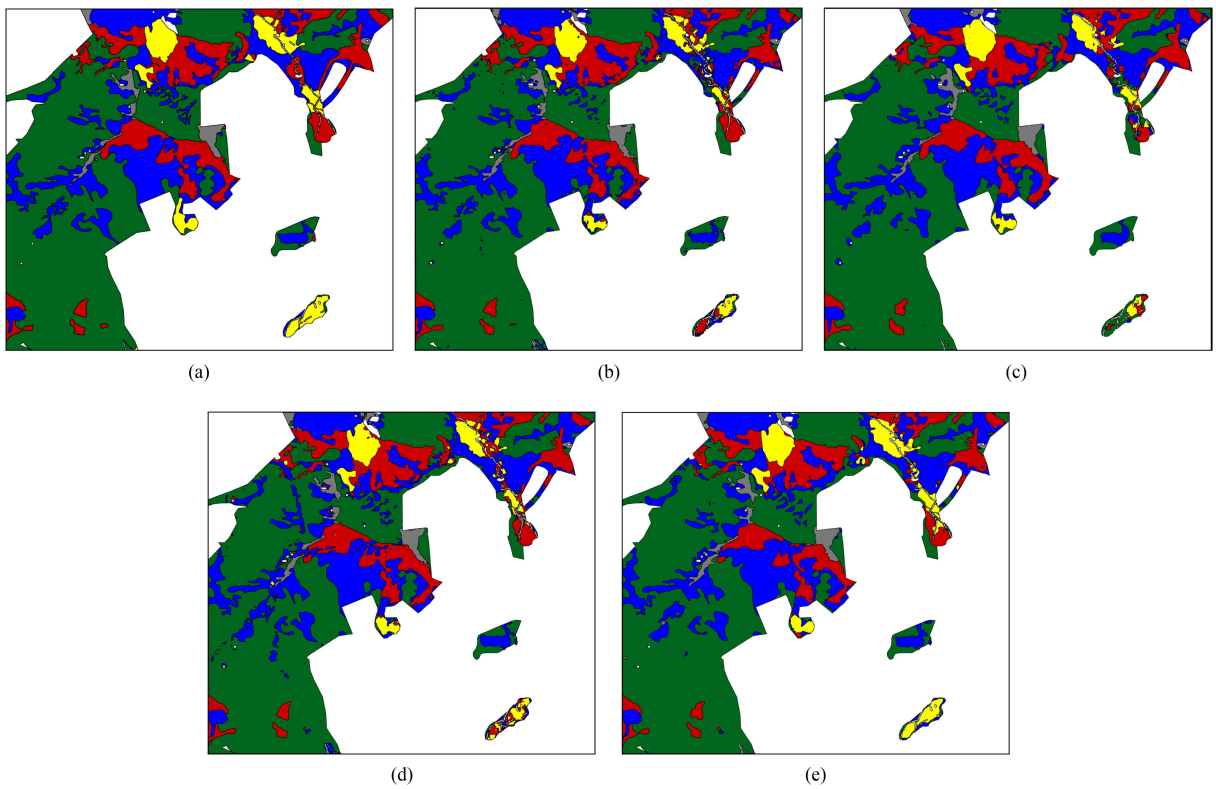


Fig. 11. Comparison between (a) reference maps of the third NDVI dataset area, and the classification maps retrieved using (b) TWINNS model [30], (c) DMIL model [32], (d) MrFusion model [33], and (e) proposed method (yellow is Level 0, red is Level 20, blue is Level 50, dark green is Forest, gray is Impervious areas, and white is masked pixels).

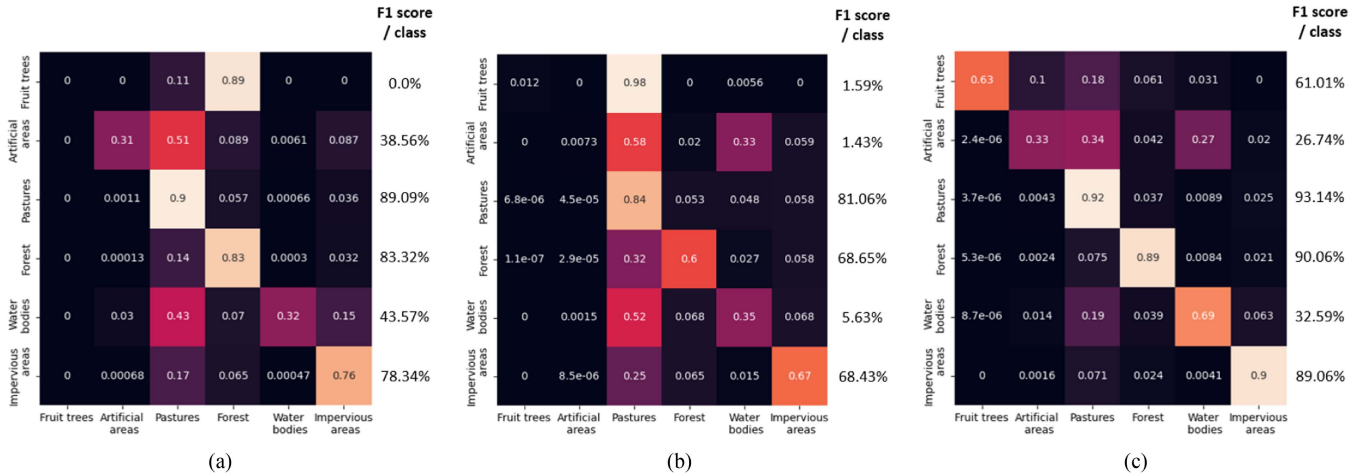


Fig. 12. Confusion matrices and F1 score of the various classes (a) processing only VHR single-date orthophotos, (b) processing only multitemporal Sentinel-1 images, and (c) proposed method processing VHR orthophotos and Sentinel-1 image time series.

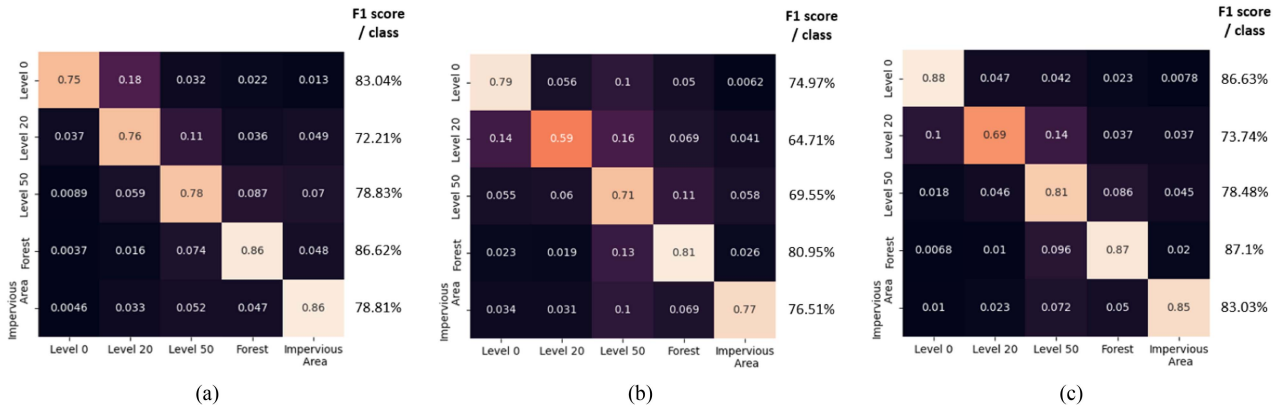


Fig. 13. Confusion matrices and F1 score of the various classes (a) processing only VHR single-date orthophotos, (b) processing only multitemporal Sentinel-2 images, and (c) proposed method processing VHR orthophotos and multitemporal Sentinel-2 images.

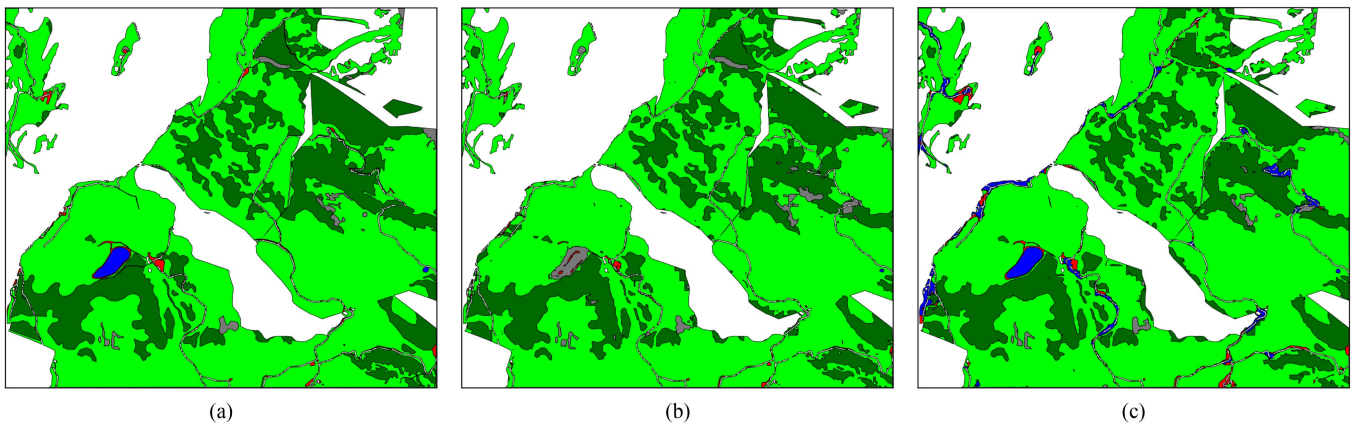
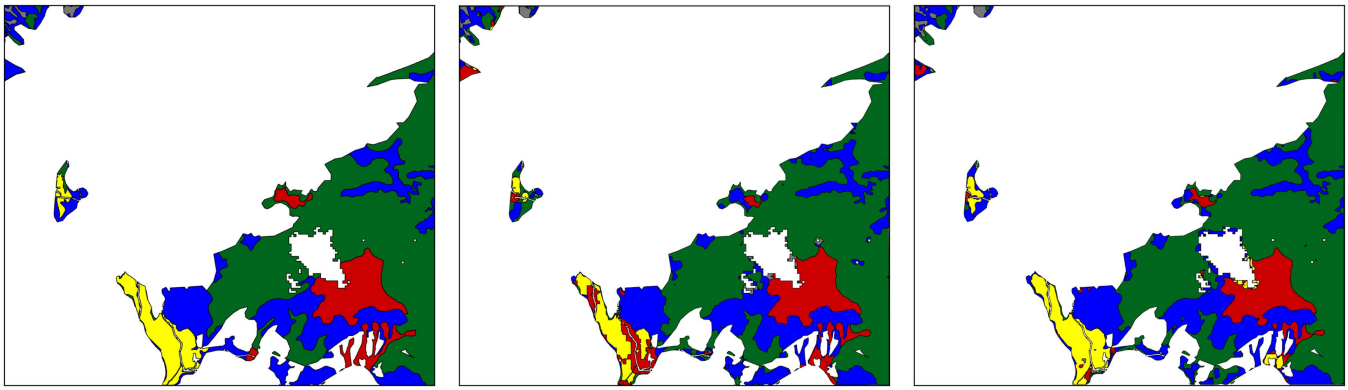
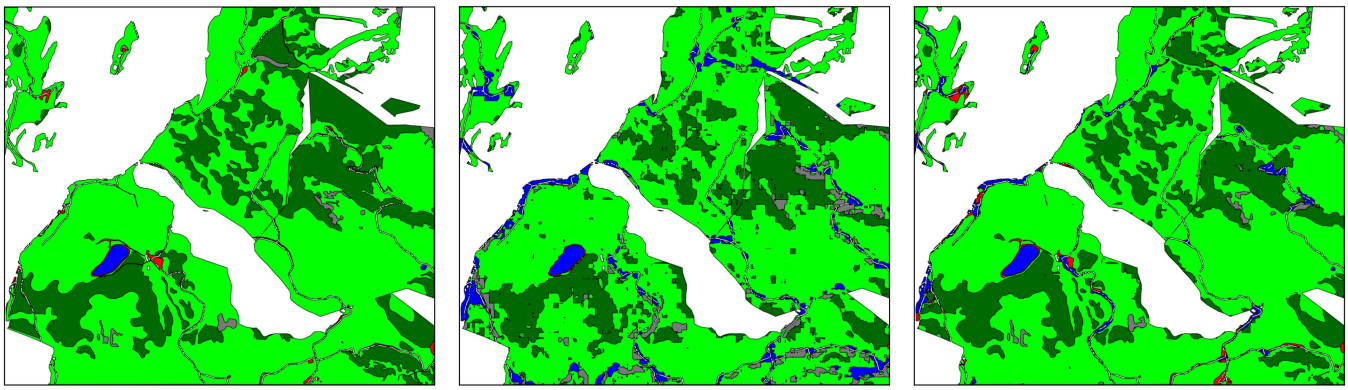


Fig. 14. Comparison between (a) reference maps of the first dataset, and the classification maps retrieved by (b) processing only VHR single-date orthophotos, and the proposed method processing VHR orthophotos and (c) Sentinel-1 image time series (yellow is Fruit trees, red is Artificial areas, green is Pastures, dark green is Forest, blue is Water bodies, gray is Impervious areas, and white is masked pixels).



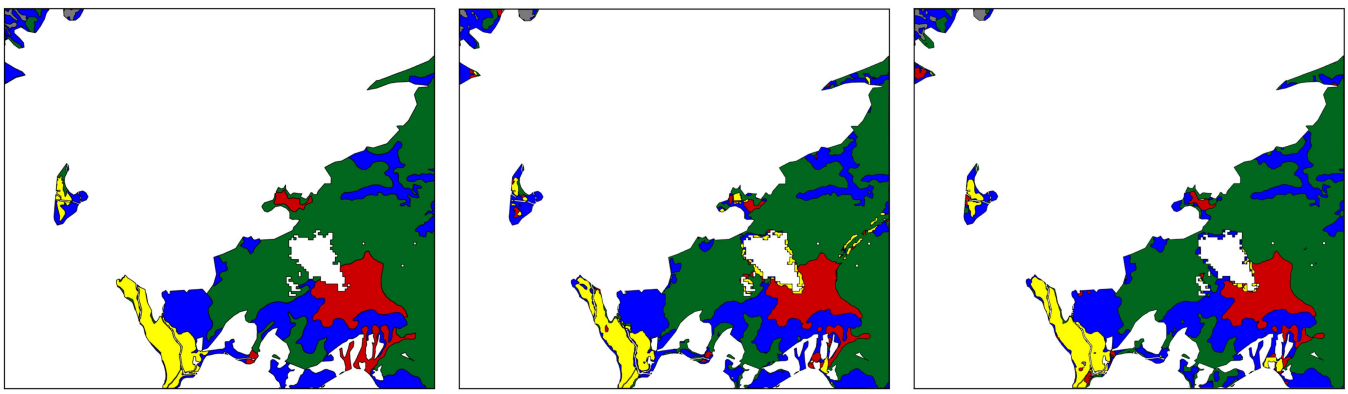
(a) (b) (c)

Fig. 15. Comparison between (a) reference maps of the second dataset, and the classification maps retrieved by (b) processing only VHR single-date orthophotos, and (c) proposed method processing VHR orthophotos and multitemporal Sentinel-2 images (yellow is Level 0, red is Level 20, blue is Level 50, dark green is Forest, gray is Impervious areas, and white is masked pixels).



(a) (b) (c)

Fig. 16. Comparison between (a) reference maps of the first dataset, and the classification maps retrieved by (b) processing only Sentinel-1 image time series, and (c) proposed method processing VHR orthophotos and Sentinel-1 image time series (yellow is Fruit trees, red is Artificial areas, green is Pastures, dark green is Forest, blue is Water bodies, gray is Impervious areas, and white is masked pixels).



(a) (b) (c)

Fig. 17. Comparison between (a) reference maps of the second dataset, and the classification maps retrieved by (b) processing only multi-temporal Sentinel-2 images, and (c) proposed method processing VHR orthophotos and multitemporal Sentinel-2 images (yellow is Level 0, red is Level 20, blue is Level 50, dark green is Forest, gray is Impervious areas, and white is masked pixels).

classification accuracy decreased (i.e., the OA of the first dataset is 73.79%, while the OA of the second one is 72.98%). The introduction of VHR images in the classification analysis added the spatial information unavailable in the multitemporal data and improved the discrimination capability of the classification method of 16.65% and 8.27% for the first and second datasets, respectively. It is possible to observe that F1 score of each class in both datasets improved with the VHR orthophoto inclusion in the classification process (see Figs. 12 and 13). The first dataset classes with fewer training samples and requiring high spatial resolution for the modeling were misclassified [see Fig. 12(b)]. The use of VHR images improved the discrimination of these classes using the same number of training samples. This proved the relevance of the multiresolution analysis.

As we can observe in the qualitative results, the multiscale classification analysis better improved the discrimination between Pastures, Forests, Water Bodies, and Impervious areas, in the first dataset [e.g., Fig. 16(b) and (c)]. These classes are the most fragmented ones and are challenging to discriminate accurately due to the spatial resolution of Sentinel-1 images. In the second dataset, the confusion between Forest and Level 50 was one of the most common classification errors [e.g., the right side of Fig. 17(b)]. This was due to the slight difference in the tree density between Level 50 and Forest, which is difficult to observe in Sentinel-2 images. Class Level 50 presented many trees in some areas and can be easily confused with the Forest class. The poor spatial resolution also caused problems in the discrimination of similar pasture classes (i.e., Level 0, Level 20, and Level 50) given their similarity [e.g., the middle-left part of Fig. 17(b)]. The multiscale analysis performed by the proposed method alleviated these problems [e.g., the middle-left part of Fig. 17(c)].

IV. Conclusion

In this article, we proposed a supervised DL classification method that processes multisensor and multiresolution RS image time series to combine the spatial context information derived by VHR images and the temporal information obtained from the HR image time series and retrieve a land-cover map. We tested our method using two datasets composed of one VHR orthophoto, Sentinel-2, or Sentinel-1 image time series acquired over the Trentino region in Italy. In the tests, we observed the improvement provided by the multiscale analysis in the correct discrimination of complex classes characterized by small objects. We proved that the temporal information analysis improved the modeling of classes with behavior varying over the year that cannot be represented by a single-date image. The proposed method improved the modeling of the temporal information with respect to the SoA ones and achieved more accurate classification performance using a less computationally demanding model. This improvement is due to the fusion of multiscale features extracted by multisensor multiresolution images and temporal features retrieved by analyzing the image time series with 3-D convolutional layers. These layers proved to better model spatial and temporal information than 2-D convolutional and SAE layers exploited in SoA classification methods.

In future activities, we plan to test the proposed method on other kinds of classes using images acquired by different sensors. Moreover, we want to exploit convolutional LSTM layers instead of 3-D convolutional ones to observe if they optimize the modeling of spatial and temporal information.

Acknowledgment

The authors would like to thank Agenzia per le Erogazioni in Agricoltura (AGEA) and Agenzia Provinciale per i Pagamenti in Agricoltura (APPAG) to provide the data and the photointerpreted maps to test the proposed methods.

References

[1] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[2] J. Nichol and C. Lee, "Urban vegetation monitoring in Hong Kong using high resolution multispectral images," *Int. J. Remote Sens.*, vol. 26, no. 5, pp. 903–918, 2005.

[3] M. Kuffer and J. Barrosb, "Urban morphology of unplanned settlements: The use of spatial metrics in VHR remotely sensed images," *Procedia Environ. Sci.*, vol. 7, pp. 152–157, 2011.

[4] X. Zhang, S. Du, Q. Wang, and W. Zhou, "Multiscale geoscene segmentation for extracting urban functional zones from VHR satellite images," *Remote Sens.*, vol. 10, no. 2, 2018, Art. no. 281.

[5] J. A. López et al., "Land cover classification of VHR airborne images for citrus grove identification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 66, no. 1, pp. 115–123, 2011.

[6] Y. Chen, D. Ming, and X. Lv, "Superpixel-based land cover classification of VHR satellite image combining multi-scale CNN and scale parameter estimation," *Earth Sci. Inform.*, vol. 12, no. 3, pp. 341–363, 2019.

[7] A. Coates and A. Ng, "Selecting receptive fields in deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 2528–2536.

[8] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 4905–4913.

[9] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2017.

[10] C. Persello and A. Stein, "Deep fully convolutional networks for the detection of informal settlements in VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2325–2329, Dec. 2017.

[11] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 609–630, Mar. 2013.

[12] J. R. Bergado, C. Persello, and A. Stein, "Recurrent multiresolution convolutional networks for VHR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6361–6374, Nov. 2018.

[13] C. Zhang, S. Wei, S. Ji, and M. Lu, "Detecting large-scale urban land cover changes from very high-resolution remote sensing images using CNN-based classification," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 4, 2019, Art. no. 189.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[15] C. Pelletier, S. Valero, J. Inglada, N. Champion, and G. Dedieu, "Assessing the robustness of random forests to map land cover with high-resolution satellite image time series over large areas," *Remote Sens. Environ.*, vol. 187, pp. 156–168, 2016.

[16] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes, "Operational high-resolution land cover map production at the country scale using satellite image time series," *Remote Sens.*, vol. 9, no. 1, 2017, Art. no. 95.

[17] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. Marais Sicre, and G. Dedieu, "Effect of training class label noise on classification performances for land cover mapping with satellite image time series," *Remote Sens.*, vol. 9, no. 2, 2017, Art. no. 173.
[18] V. Maus, G. Câmara, R. Cartaxo, A. Sanchez, F. M. Ramos, and G. R. De Queiroz, "A time-weighted dynamic time warping method for land-use and land-cover mapping," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3729–3739, Aug. 2016.
[19] Y.-L. Kong, Q. Huang, C. Wang, J. Chen, J. Chen, and D. He, "Long short-term memory neural networks for online disturbance detection in satellite image time series," *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 452.
[20] D. S. Reddy and P. R. C. Prasad, "Prediction of vegetation dynamics using NDVI time series data and LSTM," *Model. Earth Syst. Environ.*, vol. 4, no. 1, pp. 409–419, 2018.
[21] H. Crisóstomo de Castro Filho et al., "Rice crop detection using LSTM, Bi-LSTM, and machine learning models from sentinel-1 time series," *Remote Sens.*, vol. 12, no. 16, 2020, Art. no. 2655.
[22] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1685–1689, Oct. 2017.
[23] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 523.
[24] D. Ienco, Y. J. E. Gbodjo, R. Gaetano, and R. Interdonato, "Weakly supervised learning for land cover mapping of satellite image time series via attention-based cnn," *IEEE Access*, vol. 8, pp. 179547–179560, 2020.
[25] D. Ienco, R. Gaetano, R. Interdonato, K. Ose, and D. H. T. Minh, "Combining sentinel-1 and sentinel-2 time series via RNN for object-based land cover classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 4881–4884.
[26] Y. J. E. Gbodjo, D. Ienco, L. Leroux, R. Interdonato, R. Gaetano, and B. Ndao, "Object-based multi-temporal and multi-source land cover mapping leveraging hierarchical class relationships," *Remote Sens.*, vol. 12, no. 17, 2020, Art. no. 2814.
[27] V. S. F. Garnot, L. Landrieu, and N. Chehata, "Multi-modal temporal attention models for crop mapping from satellite time series," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 294–305, 2022.
[28] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "M3Fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4939–4949, Dec. 2018.
[29] T. Di Martino, M. Lenormand, and E. C. Koeniguer, "Multi-branch deep learning model for detection of settlements without electricity," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 1847–1850.
[30] D. Ienco, R. Interdonato, R. Gaetano, and D. H. T. Minh, "Combining sentinel-1 and sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture," *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 11–22, 2019.
[31] C. Robinson et al., "Large scale high-resolution land cover mapping with multi-resolution data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12726–12735.
[32] X. Liu et al., "Deep multiple instance learning-based spatial–spectral classification for pan and ms imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 461–473, Jan. 2018.
[33] R. Gaetano, D. Ienco, K. Ose, and R. Cresson, "MRFusion: A deep learning architecture to fuse PAN and MS imagery for land cover mapping," Jun. 2018, *arXiv:1806.11452*.
[34] R. M. Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke, and D. Lobell, "Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 75–82.
[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[36] J. A. C. Martinez, L. E. C. La Rosa, R. Q. Feitosa, I. D. Sanches, and P. N. Happ, "Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences," *ISPRS J. Photogrammetry Remote Sens.*, vol. 171, pp. 188–201, 2021.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[38] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial–contextual information in very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013.
[39] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised multiple-change detection in VHR multisensor images via deep-learning based adaptation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5033–5036.
[40] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, "3D convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 75.
[41] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[42] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.
[43] P. Zhang, Y. Ke, Z. Zhang, M. Wang, P. Li, and S. Zhang, "Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery," *Sensors*, vol. 18, no. 11, 2018, Art. no. 3717.
[44] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Fully convolutional neural networks for remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 5071–5074.
[45] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 1–9.
[46] B. K. Horn, "Hill shading and the reflectance map," *Proc. IEEE*, vol. 69, no. 1, pp. 14–47, Jan. 1981.
[47] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*.
[48] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, 1992, pp. 950–957.

Luca Bergamasco (Member, IEEE) received the "Laurea" (B.Sc.) degree in electronics and telecommunication engineering, the "LaureaMagistrale" (M.Sc.) degree in information and communication engineering, and the Ph.D. degree in information and communication technologies (cum laude) from the University of Trento, Trento, Italy, in 2016, 2018, and 2022, respectively.

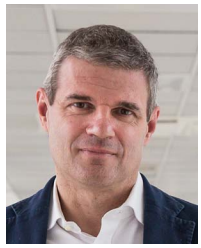
He is currently a Postdoctoral Researcher with the Remote Sensing for Digital Earth Unit (RSDE), Fondazione Bruno Kessler (FBK), Trento. His research interests include remote sensing data analysis through machine learning, pattern recognition, and deep learning techniques. Some research topics involve deep learning methods for multitemporal image analysis, change detection in multispectral, hyperspectral, synthetic-aperture-radar (SAR), multimodal data, time-series analysis, and domain adaptation.

Dr. Bergamasco was the recipient of the prize for the 2019 best three Italian Master's thesis in geoscience and remote sensing awarded by the Italian Chapter of the IEEE Geoscience and Remote Sensing Society. He is a referee for several journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Francesca Bovolo (Senior Member, IEEE) received the Laurea (B.S.) and the Laurea Specialistica (M.S.) degrees (summa cum laude) in telecommunication engineering, and the Ph.D. degree in communication and information technologies, all from the University of Trento, Trento, Italy, in 2001, 2003, and 2006, respectively.

She was a Research Fellow with the University of Trento, until 2013. She is currently the Founder and the Head of Remote Sensing for Digital Earth Unit, Fondazione Bruno Kessler, Trento, and a member of the Remote Sensing Laboratory, Trento. She is one of the co-investigators of the Radar for Icy Moon Exploration instrument of the European Space Agency Jupiter Icy Moons Explorer and member of the science study team of the EnVision mission to Venus. She conducts research on the topics of her research interests within the context of several national and international projects, which include remote-sensing image processing, multitemporal remote sensing image analysis, change detection in multispectral, hyperspectral, and synthetic aperture radar images, and very high-resolution images, time-series analysis, content-based time-series retrieval, domain adaptation, and light detection and ranging (LiDAR) and radar sounders.

Dr. Bovolo is a Member of the program and scientific committee of several international conferences and workshops. She was the recipient of the First Place in the Student Prize Paper Competition of the 2006 IEEE International Geoscience and Remote Sensing Symposium (Denver, 2006). She was the Technical Chair of the Sixth International Workshop on the Analysis of Multitemporal Remote-Sensing Images (MultiTemp 2011 and 2019). She has been a Co-Chair of the SPIE International Conference on Signal and Image Processing for Remote Sensing since 2014. She is the Publication Chair of the International Geoscience and Remote Sensing Symposium in 2015. She has been an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING since 2011 and the Guest Editor of the Special Issue on Analysis of Multitemporal Remote Sensing Data for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She is a referee for several international journals.



Lorenzo Bruzzone (Fellow, IEEE) received the Laurea (M.S.) degree in electronic engineering (summa cum laude) and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications with the University of Trento, Trento, Italy, where he teaches remote sensing, radar, and digital communications. He is also the Founder and the Director of the Remote Sensing Laboratory (https://rslab.disi.unitn.it/), Department of Information Engineering and Computer Science, University of Trento. He is the author (or co-author) of 294 scientific publications in referred international journals (221 in IEEE journals), more than 340 papers in conference proceedings, and 22 book chapters. His current research interests include the areas of remote sensing, radar and synthetic-aperture-radar (SAR), signal processing, machine learning, and pattern recognition.

Dr. Bruzzone is the Principal Investigator of many research projects. He promotes and supervises research within the frameworks of many national and international projects. Among the others, he is currently the Principal Investigator of the Radar for icy Moon exploration (RIME) instrument in the framework of the JUpiterICy moons Explorer (JUICE) mission of the European Space Agency (ESA) and the Science Lead for the High Resolution Land Cover project in the framework of the Climate Change Initiative of ESA. He is Editor/Co-Editor of 18 books/conference proceedings and 1 scientific book. His papers are highly cited, as proven from the total number of citations (more than 40000) and the value of the h-index (92) (source: Google Scholar). He was invited as a keynote Speaker in more than 40 international conferences and workshops. Since 2009, he has been a Member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS), where since 2019, he has been the Vice President for Professional Activities. He ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, WA, USA, July 1998. He was the recipient of many international and national honors and awards, including the recent IEEE GRSS 2015 Outstanding Service Award, the 2017 and 2018 IEEE IGARSS Symposium Prize Paper Awards and the 2019 WHISPER Outstanding Paper Award. He was a Guest Co-Editor of many Special Issues of international journals. He is the co-founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images series and is currently a Member of the Permanent Steering Committee of this series of workshops. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the founder of the IIEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE for which he has been Editor-in-Chief between 2013 and 2017. He is currently an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He has been a Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society between 2012 and 2016.