Unsupervised Deep Joint Segmentation of Multitemporal High-Resolution Images

Sudipan Saha, *Graduate Student Member, IEEE*, Lichao Mou, Chunping Qiu, *Graduate Student Member, IEEE*, Xiao Xiang Zhu, *Senior Member, IEEE*, Francesca Bovolo, *Senior Member, IEEE*, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*— High/very-high-resolution (HR/VHR) multitemporal images are important in remote sensing to monitor the dynamics of the Earth's surface. Unsupervised object-based image analysis provides an effective solution to analyze such images. Image semantic segmentation assigns pixel labels from meaningful object groups and has been extensively studied in the context of single-image analysis, however not explored for multitemporal one. In this article, we propose to extend supervised semantic segmentation to the unsupervised joint semantic segmentation of multitemporal images. We propose a novel method that processes multitemporal images by separately feeding to a deep network comprising of trainable convolutional layers. The training process does not involve any external label, and segmentation labels are obtained from the argmax classification of the final layer. A novel loss function is used to detect object segments from individual images as well as establish a correspondence between distinct multitemporal segments. Multitemporal semantic labels and weights of the trainable layers are jointly optimized in iterations. We tested the method on three different HR/VHR data sets from Munich, Paris, and Trento, which shows the method to be effective. We further extended the proposed joint segmentation method for change detection (CD) and tested on a VHR multisensor data set from Trento.**

***Index Terms*— Deep learning, high resolution (HR), multitemporal image, segmentation.**

Manuscript received December 27, 2019; revised March 14, 2020; accepted April 20, 2020. Date of publication May 11, 2020; date of current version November 24, 2020. The work of X. X. Zhu was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Program (*So2Sat*) under Grant ERC-2016-StG-714087, in part by Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Local Unit "Munich Unit @ Aeronautics, Space and Transport (MASTr)," and in part by Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research." *(Corresponding author: Francesca Bovolo.)*

Sudipan Saha is with Fondazione Bruno Kessler, 38123 Trento, Italy, and also with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: saha@fbk.eu).

Lichao Mou is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany.

Chunping Qiu is with Signal Processing in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany.

Xiao Xiang Zhu is with Signal Processing in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany, and also with Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany.

Francesca Bovolo is with Fondazione Bruno Kessler, 38123 Trento, Italy (e-mail: bovolo@fbk.eu).

Lorenzo Bruzzone is with the Department of Information Engineering and Computer science, University of Trento, 38123 Trento, Italy.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.2990640

I. Introduction

MULTITEMPORAL information extraction and analysis have seen an increased interest in the last decade. Many new high/very-high-resolution (HR/VHR) sensors (e.g., Sentinel-2, Pleiades, Quickbird, and Gaofen-2) were launched, making it possible to analyze multitemporal images at an unprecedented scale. Different from the multitemporal low/medium resolution images, pixels in HR/VHR images show a strong spatial correlation and complexity. Thus, there is consensus in the literature that object-level information and semantic details should be extracted to effectively exploit multitemporal HR/VHR images [1]–[3]. Multitemporal image analysis methods can be both supervised and unsupervised. Though supervised methods potentially provide a better result and are more effective to extract object information, unsupervised methods are preferred in the literature [4] due to the difficulty of collecting multitemporal labeled data. Most multitemporal image analysis methods are designed for change detection (CD) [4], [5] in bitemporal images acquired over the same geographical area. Other applications include multitemporal classification and time-series trend analysis [1], [6].

Segmentation has proved to be useful for multitemporal HR/VHR image analysis. There are few methods that rely on superpixel segments for object-based multitemporal image analysis [4], [7]. In [7], a method is proposed that first detects superpixel segments from one of the bitemporal images and, subsequently, applies the same segmentation mask on the other image. Features extracted from each segment are compared each other to extract change information. Such methods utilize segmentation only as a tool to obtain coherent image parts as a spatial unit of comparison in the CD method. They do not investigate the semantic meaning of the segments or temporal relationship between segments. A step forward is semantic segmentation [8], [9] that assigns semantic label to pixels. Many works in the computer vision literature [10] have shown the superiority of semantic segmentation for image understanding tasks. It has proven to be useful in solving multitemporal image analysis tasks [5]. Saha *et al.* [5] presented an unsupervised deep-change-vector-analysis (DCVA) framework that uses a CNN pretrained for semantic segmentation [9] as multitemporal feature extractor.

In contrast to mere segmentation, semantic segmentation attempts to partition the target image into semantically

meaningful parts. Semantic segmentation assigns a semantic label to each pixel. Still, it is able to get generally nice looking segments and not salt and pepper pixel labels. Most state-of-the-art semantic segmentation methods employ deep neural network [11]–[13] treating semantic segmentation as a supervised learning problem. Such methods employ a large pixelwise labeled data set to train a deep model, and the trained model is subsequently applied on target images for pixelwise labeling. They require a substantial amount of pixelwise labeled training data and have not been explored in the literature in the context of the multitemporal image analysis. As a brute-force mechanism, it is possible to individually apply supervised semantic segmentation on each image separately for multitemporal analysis. However, such a mechanism does not exploit the temporal correlation between multitemporal images. There is a scope of further investigating semantic segmentation for unsupervised multitemporal image analysis.

Recently, a few deep unsupervised image segmentation [14]–[16] methods have been proposed in the computer vision literature. In [14], a large unlabeled data set is required for training. Exploiting similarity between deep segmentation and unsupervised deep pixel grouping [15], [17] proposed a method for unsupervised single-image segmentation. Inspired by the success of these methods [15] and the potential of semantic segmentation in multitemporal image analysis [5], we propose to extend the concept of semantic segmentation to multitemporal setting in an unsupervised way. The proposed novel unsupervised segmentation method can ingest multitemporal images as input and produces a semantic segmentation map for each of them where: 1) each label represents a semantically meaningful entity and 2) the segmentation process takes into account the temporal correlation between corresponding pixels in the multitemporal images. Different from the supervised segmentation, the exact name of the semantic label is unknown in the unsupervised scenario. The proposed method simultaneously processes multitemporal input images through a trainable deep network and jointly optimizes the feature representation and label assignment for each image in an iterative fashion to solve the multitemporal semantic segmentation problem. As an end result, we obtain coherent multitemporal semantic segmentation maps being temporally consistent. We further extend the proposed method for the CD. The key contributions of this article are summarized as follows:

1) Extension of the notion of semantic segmentation to the multitemporal scenario.
2) Definition of an unsupervised multitemporal joint segmentation method.
3) Definition of a set of novel loss functions that can simultaneously segment the multitemporal images while establishing a temporal correspondence between multitemporal segments.
4) Extension of proposed joint segmentation method for CD.

The rest of this article is organized as follows. We discuss the related works in Section II. We present the problem statement and a synopsis of the proposed solution in Section III. We detail the proposed method in Section IV. Section V describes extension of the proposed method for CD in bitemporal images. Data sets and results are presented in Section VI. Results related to CD are presented in Section VII. We conclude this article and discuss the scope of future research in Section VIII.

II. Related Works

As our work focuses on unsupervised deep learning-based multitemporal semantic segmentation, here, we briefly discuss the following topics that are related to our work: 1) deep learning-based image segmentation methods; 2) deep learning-based multitemporal image analysis; and 3) previous usage of segmentation in multitemporal image analysis.

A. Deep Segmentation

In the computer vision literature, supervised deep semantic segmentation is a well-explored topic. Compared with that, unsupervised methods for deep image segmentation are a less explored research avenue.

1) Supervised Deep Semantic Segmentation: The supervised approaches for deep image semantic segmentation can be considered as a pixel-level supervised learning task given a set of reliable training pixels. A number of methods have been proposed in the literature for semantic segmentation using deep neural networks [9], [11]–[13], [18]–[20]. In [18], region proposal is combined with CNN for semantic segmentation and object detection. Fully convolutional network (FCN) [11] that replace the fully connected layers with the convolutional layers is one of the effective approaches for supervised semantic segmentation. FCN takes as input an arbitrary spatial dimension and generates a segmentation map of the same size. Many variants of the FCN have been proposed in the literature, e.g., the one in [12] that presents a U-shaped architecture to supplement a usual contracting network by successive layers to capture context and a symmetric expanding path to improve the localization accuracy. SegNet [13] is another variation of the FCN. It upsamples encoded features by storing the max-pooling indices used in the pooling layer. As these methods require a substantial amount of training data, their application in multitemporal image analysis is limited.

2) Unsupervised Deep Segmentation: Recently, an increased interest can be seen in the exploration of unsupervised deep learning techniques, especially based on transfer learning methods [21] and generative adversarial network (GAN)-based methods that still require ample amount of unlabeled data [22]. Aligned with this trend, we observe that few works have been proposed in the deep learning literature to address the image segmentation problem in an unsupervised way [14], [15]. In [14], a W-shaped network is proposed by modifying the U-network. However, the network is complex consisting of 46 convolutional layers. Though the method in [14] does not use label information for training, it still requires a substantially big data set for training (trained on Pascal VOC 2012 data set [23]). In [15], an unsupervised single-image segmentation method is proposed based on deep clustering that uses fewer layers compared with [14].

The output label is obtained at the final layer using argmax classification that is further regulated by superpixel-based refinement. Weights of the convolutional layers and predicted labels are jointly optimized in an iterative fashion by using a loss function that does not require any labeled data.

B. Deep Multitemporal Image Analysis

Recently, deep learning has shown excellent performance in many image understanding and computer vision tasks [24], [25]. Deep learning-based methods are suitable to extract high-level semantically rich visual features [26], and thus, they have been shown to be effective for remote sensing image analysis too [8], [9], [27]–[29]. Since deep learning methods are data-intensive, they have seen comparatively fewer applications in the multitemporal image analysis [30]. Most of these methods are supervised and deal with specific applications. Deep Siamese network is a popular technique in supervised deep multitemporal image analysis, e.g., the supervised CD method proposed by Zhan *et al.* [31]. Some methods use preclassification schema to obtain a coarse initial change map that is used subsequently to further train the CD model to reduce labeled data dependence [32], [33]. Zhang *et al.* [32] proposed a method for CD where a coarse change map is first detected to identify most likely unchanged pairs that are used to learn a mapping neural network. Gao *et al.* [33] proposed a similar approach using wavelet features for multitemporal synthetic aperture radar (SAR) image analysis. The recurrent neural network has been used in several works related to supervised CD [34], [35]. Geng *et al.* [36] proposed a supervised binary CD method based on contractive autoencoders. Xu *et al.* [37] proposed an autoecnoder-based method to learn the correspondence between prechange image and postchange image. In an attempt to design deep unsupervised CD, pretrained CNN networks are used without labeled training data, e.g., deep change vector analysis (DCVA) [5], [38]. Such methods rely on transferring the pretrained deep features for multitemporal image analysis, but they do not take into account the distribution of target scenes. Their accuracy is affected by the similarity between training data and target scenes.

C. Segmentation for Multitemporal Image Analysis

Segmentation is often used in the object-based multitemporal image analysis as a preprocessing step [4]. Such methods do not focus on obtaining a semantic segmentation map, rather they merely use a superpixel segmentation that defines a perceptually similar continuous region in the image. In [7], a method is proposed that first detects a segmentation mask from one of the bitemporal images and, subsequently, applies the same segmentation mask on the other image. Features extracted from each segment are compared each other to extract change information. Zhang *et. al.* [39] proposed a method that first detects multitemporal changed objects based on separate segmentation of the multitemporal images and then establishes a spatial correspondence between them.

DCVA [5] partially illustrates the usefulness of semantic segmentation for multitemporal image analysis. In [5], a deep network trained for semantic segmentation is used to obtain coherent multitemporal deep features that are pixelwise compared each other to obtain change information. Peng *et. al.* [40] proposed a semantic segmentation inspired architecture for CD.

III. Problem Formulation and Synopsis of the Proposed Solution

A. Problem

We aim to design a multitemporal image segmentation framework that addresses the segmentation problem jointly in the multitemporal images. Let $\mathcal{X} = \{X_1, X_2, \ldots, X_T\}$ be a set of T HR/VHR optical images taken over the same geographical region at different time ($t = 1, 2, \ldots, T$). The images consist of N pixels $x_{n,t}$ ($n = 1, \ldots, N$ and $t = 1, 2, \ldots, T$) with M channels. We aim to assign a label $c_{n,t}$ to each pixel $x_{n,t}$ such that each distinct label bears some meaningful semantic notion. After segmentation, pixels $x_{n,t}$ and $x_{n,t+1}$ at the same geographical position but acquired in two successive time steps t and $t+1$ tend to have the same label if they are unchanged. If they have different labels, it would imply a possible presence of change between considered time steps.

B. Synopsis of the Proposed Method

The proposed deep convolutional network-based joint segmentation method addresses the abovementioned aspect in an effective manner by: 1) obtaining deep representations from multitemporal images; 2) clustering pixels (i.e., assigning labels) based on the property that pixels belonging to the same segment are likely to obtain similar deep feature representation; 3) iteratively adjusting the deep representation and the labels to converge in a spatially and temporally consistent way.

Let us assume that multitemporal images are coregistered as per standard procedure [41]. Multitemporal images are separately processed through a deep network consisting of $L-1$ convolutional layers and one linear projection layer (convolutional layer with filter size 1×1). The network includes also other postprocessing layers/functions, e.g., rectified linear unit (ReLU) and batch normalization. Pooling layer or stride are omitted in the convolutional layers, and hence, the spatial size is maintained at the final layer. The deep network is trainable in an unsupervised way. The training process involves a set of loss functions that do not require labeled data. Toward this goal, the same network (i.e., with the same architecture and same weights) is separately applied to each image in $\mathcal{X}$, and the output of the final layer is regarded as deep feature representation of the input images. Using the property that semantically similar pixels are probable to obtain the highest activation in the same deep feature, the segmentation labels are obtained as argmax classification of the deep feature representation. The multitemporal labels (obtained as argmax classification) and the multitemporal deep representations are used to compute a set of novel loss functions that account for the spatial and temporal homogeneities of the deep feature representations. The loss functions do not require any external label and are designed in a way that the network converges (in iterations) toward a more coherent semantic segmentation map from each image in $\mathcal{X}$ and also learns to assign the

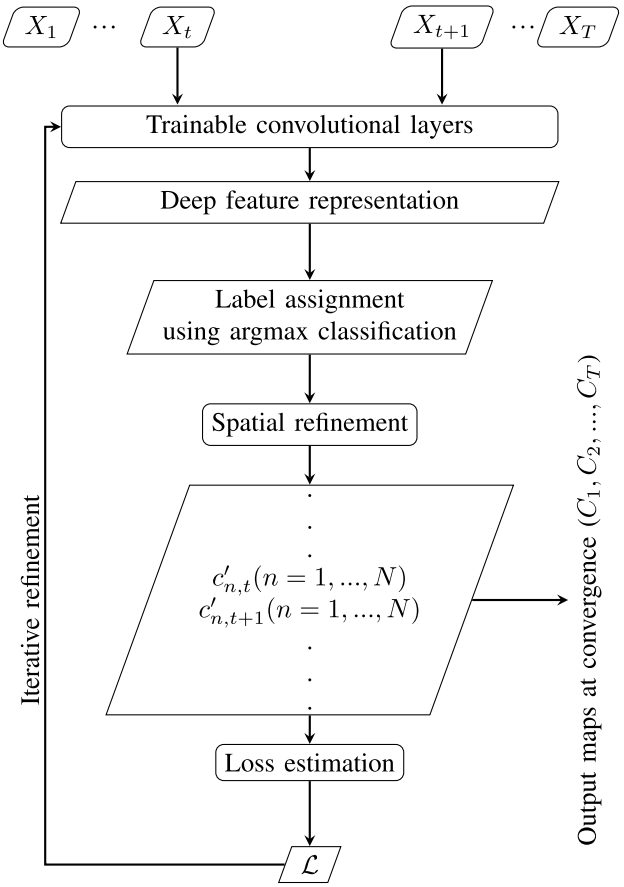


Fig. 1. Proposed multitemporal image joint segmentation mechanism.

same semantic labels to pixels whenever the values are time consistent. The network is modulated in an iterative fashion using the estimated loss. The iterative process is stopped using a stopping criterion. At convergence, the network is used to obtain multitemporal segmentation maps. The proposed framework is shown in Fig. 1.

C. Significance Relative to the Related Works

Comparing to the related works discussed in Section II, the proposed method is novel and is built upon the methods discussed in Section II-A, especially the ones that use FCNs [11] and unsupervised deep clustering [15]. The proposed method tackles a task unexplored before in deep multitemporal image analysis (see Section II-B). The proposed task can be significantly useful, as Section II-C illustrates that segmentation and semantic segmentation methods are used in different fields of multitemporal image analysis.

IV. Proposed Method

We describe the architecture of the deep network and the mechanism of obtaining deep feature representation from input image pixels in Section IV-A. The mechanism of assigning labels based on deep feature representation is described in Section IV-B. Section IV-C describes the method to further refine the labels based on spatial homogeneity. The deep network described in Section IV-A is trainable without using any external labels, and this is achieved by using a set of loss functions detailed in Section IV-D. Section IV-E describes the method of iteratively refining the network weights and segmentation labels, and Section IV-F details the method of obtaining final segmentation maps.

A. Deep Feature Representation

In this step, we obtain deep feature representation that captures pixelwise semantics for each of the multitemporal images in $\mathcal{X}$. Inspired by the capability of convolutional layers to learn high-level semantic features [42], a set of learnable convolutional layers are used toward this goal. Filters from the first convolutional layer of a deep network are convolved with the input images to compute a feature representation from the input image. Filters from successive convolutional layers are convolved with the output from the preceding layer. All images are processed separately through identical networks (i.e., the same set of convolutional layers with the same weights). In an ideal scenario where the multitemporal images do not show any change or radiometric differences, the output produced for each image is expected to be identical. In a practical scenario, multitemporal images show many differences even in the absence of changes on the ground, and thus, the deep feature representation obtained for images in $\mathcal{X}$ differs.

For effectively capturing the spatial and semantic information, $L-1$ convolutional layers are used. The first convolutional layer ingests input images with M channels and produces an output of M^1 features. In doing so, the layer uses filters of spatial size 3×3, i.e., learnable weight $\mathbb{W}^1$ has dimension $\mathbb{R}^{3\times3\times M\times M^1}$. The convolution operation does not use any stride, i.e., filters are moved one pixel at a time. Pooling layer is not used, and hence, the output of the first layer has the same spatial dimension as input. The convolutional layer is followed by a ReLU activation function that introduces nonlinearity to the output of the convolutional layer. Output $H_t^{1'}$ from the first layer for X_t can be represented as

$$H_t^{1'} = \text{ReLU}(X_t \circledast \mathbb{W}^1). \tag{1}$$

The ReLU activation function is followed by a batch normalization layer where batch normalization process involves all the pixels in the image. The T sets of batch normalization parameters $\{\mu_t^1, \sigma_t^1\}$ $(t = 1, 2, \ldots, T)$ are obtained separately for each image in $\mathcal{X}$ and are used to normalize $H_t^{1'}$ to obtain H_t^1 $(t = 1, \ldots, T)$.

The following convolutional layers $(l = 2, \ldots, L-1)$ take input of feature size M^{l-1} and generate output of the feature size M^l. They are further processed through ReLU activation functions and batch normalization layers. The output obtained for X_t from the last convolutional layer is H_t^{L-1}.

After the convolutional layers, a linear projection layer (i.e., a convolutional layer with filters of spatial size 1×1) is used to change the dimensionality in the kernel space to $M^L = K$, where K (set as 100) is a number much larger than the maximum number of segmentation labels. In this layer, $\mathbb{W}^L$ of size $\mathbb{R}^{1\times1\times M^{L-1}\times K}$ is used to produce an output of K features each of the N pixels. Each of the K

feature maps is normalized separately for each image X_t ($t = 1, \ldots, T$) (similar to instance normalization [43]) to have zero mean and unit variance. Featurewise normalization produces feature maps on a similar scale, and thus, they have a similar chance to be selected via the argmax classification [15]. All the weights of the network $\mathbb{W}^1, \ldots, \mathbb{W}^L$ are learnable (using a set of loss functions described later which does not need any external label). After processing images in $\mathcal{X}$ through the convolutional layers, for each input pixel $x_{n,t}$ ($n = 1, \ldots, N; t = 1, 2, \ldots, T$), we obtain deep features $y_{n,t}$ of dimension K.

B. Pixel Clustering/Label Assignment

For an input image processed through a series of convolutional layers, it can be assumed that semantically similar pixels produce high activation in the same deep feature. Based on this hypothesis, $y_{n,t}$ is further processed in this step to assign the labels $c_{n,t}$. In detail, $c_{n,t}$ for a specific pixel $x_{n,t}$ is obtained by argmax classification, i.e., by choosing the feature in $y_{n,t}$ that has maximum value [15]. If the kth feature of $y_{n,t}$ is represented by $y_{n,t}(k)$, then $c_{n,t}$ is obtained as follows:

$$c_{n,t} = \arg\max_{k \in K} y_{n,t}(k). \tag{2}$$

The abovementioned processing corresponds to the clustering of feature vectors into K clusters. Using the argmax classification, we are able to determine the feature that corresponds to the highest activation of an input pixel. The pixels that obtain the highest activation in the same feature are likely to have similar semantics, and hence, they are grouped together. The label assignment is separately conducted for each image in $\mathcal{X}$.

C. Spatial Label Refinement

In semantic segmentation, it is expected that there is a spatial continuity in the labels of the image pixels, i.e., pixels lying in spatial vicinity are likely to have the same semantic label. Though this property is partly ensured by usage of convolutional layers that captures spatial context, we add an additional constraint that favors similar labels in a neighborhood. $c_{n,t}$ are further refined through a spatial mode-based statistical filtering to obtain $c'_{n,t}$ for each pixel $x_{n,t}$ ($t = 1, 2, \ldots, T$). This helps in preserving spatial consistency and to reduce a very small group of pixels having a distinct label from other pixels in the neighborhood. This process helps in merging labels with a smaller number of pixels with the ones with a larger number of pixels, thus reducing the number of distinct labels. Instead of the chosen mode-based refinement, any other spatial refinement process can be used, e.g., superpixel-based refinement [15].

D. Spatial and Temporal Homogeneities Measurement

Spatial and temporal homogeneities are estimated from the multitemporal deep feature representation and multitemporal labels using a set of loss functions that are designed toward two objectives.

1) Label assignment of single-time image gets refined in a meaningful way such that semantic information of the image is captured and label assignment converge as training iterations progress.
2) Pixels $x_{n,t}$ for each image in $\mathcal{X}$ tend to get the same label, thus capturing the temporal correlation, however accounting for a possible change between two time instants. This is based on the assumption that in a coregistered time series, generally two pixels in the consecutive time belong to same object (since changes have a low prior probability [1]).

To achieve the first goal, following [15], we use cross-entropy loss between the continuous-valued deep feature representation $y_{n,t}$ computed by the linear projection layer and the discrete-valued label assignment after spatial refinement $c'_{n,t}$. We obtain T distinct loss values $\ell_{n,t}$ ($t = 1, \ldots, T$) for the pixel $x_{n,t}$ corresponding to T images in $\mathcal{X}$

$$\ell_{n,t} = \text{crossentropy}\left(y_{n,t}, c'_{n,t}\right). \tag{3}$$

The loss term $\mathcal{L}_t$ for all N pixels is computed as

$$\mathcal{L}_t = \frac{1}{N} \sum_{n=1}^{N} \ell_{n,t} \tag{4}$$

$\mathcal{L}_t$ captures the spatial homogeneity of a single-time image X_t.

To achieve the second goal, we recall that since the same network is applied to the multitemporal images, the same spatial location in images in $\mathcal{X}$ are supposed to produce similar deep feature representation and, hence, similar label assignment. Thus, cross-entropy loss is computed between each consecutive pair of images X_t and X_{t+1} in $\mathcal{X}$. For each input pixel $x_{n,t}$ and $x_{n,t+1}$, we compute the cross-entropy loss between features of X_t ($y_{n,t}$) with the predicted refined labels for X_{t+1} ($c_{n,t+1}$)

$$\ell_{n,t,t+1} = \text{crossentropy}\left(y_{n,t}, c'_{n,t+1}\right). \tag{5}$$

This loss function ensures that feature produced for X_t are consistent with the labels generated for X_{t+1}. Similarly, $\ell_{n,t+1,t}$ is obtained as

$$\ell_{n,t+1,t} = \text{crossentropy}\left(y_{n,t+1}, c'_{n,t}\right). \tag{6}$$

It is possible that a certain pixel n experienced a change between t and $t+1$. In that case, it is expected that they produce high values for computed losses $\ell_{n,t,t+1}$ and $\ell_{n,t+1,t}$ and can contribute in undesired way to the overall loss. To tackle this issue, we exploit the trimmed cross-entropy loss inspired by [44]. We sort $\ell_{n,t,t+1}$ for the N pixels in ascending order to obtain rearranged values $\ell'_{n,t,t+1}$ ($n = 1, \ldots, N$). We trim the last N' pixels while computing the loss capturing temporal consistency between X_t and X_{t+1}

$$\mathcal{L}_{t,t+1} = \frac{1}{2(N - N')} \left(\sum_{n=1}^{N-N'} \ell'_{n,t,t+1} + \sum_{n=1}^{N-N'} \ell'_{n,t+1,t} \right). \tag{7}$$

The total loss $\mathcal{L}$ is obtained as sum of loss capturing single-time image spatial consistency (computed for all T images) and cross-time temporal consistency (computed for all $T-1$ adjacent pairs from T images)

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t + \sum_{t=1}^{T-1} \mathcal{L}_{t,t+1}. \tag{8}$$

The loss $\mathcal{L}$ captures both the semantic information from X_t $t = 1, \ldots, T$ and the temporal consistency between each of them. Since this loss function computation does not require any external label, the proposed joint segmentation method is completely unsupervised.

E. Iterative Segmentation Refinement

All the trainable weights $\mathbb{W}^1, \ldots, \mathbb{W}^L$ of the network are initialized by the Xavier initialization process [45]. The training can be thought of as two different interrelated processes. The first process predicts the cluster labels $c_{n,t}$ ($t = 1, 2, \ldots, T$ and $n = 1, \ldots, N$) given fixed network weights. The predicted cluster labels are treated as target cluster labels, and the latter process estimates loss $\mathcal{L}$ and updates the network weights $\mathbb{W}^1, \ldots, \mathbb{W}^L$. For updating of weights, we exploit stochastic gradient descent mechanism with momentum. To accomplish a reasonable training process, this training process is executed for $\mathcal{I}$ iterations. However, if the total number of clusters (distinct labels) prematurely (i.e., before $\mathcal{I}$ iterations) reaches $\mathcal{K}$ ($\mathcal{K} < K$), then the training process is stopped.

F. Obtaining Final Cluster Labels

After the completion of the training process using a stopping criterion, the trained network is separately applied on each of the images in $\mathcal{X} = \{X_1, X_2, \ldots, X_T\}$ to obtain final segmentation maps $\mathcal{C} = \{C_1, C_2, \ldots, C_T\}$. The proposed method is summarized in algorithmic form in Algorithm 1.

Algorithm 1 Unsupervised Multitemporal Joint Semantic Segmentation

Input: $x_{n,t} \in \mathbb{R}^M$ ($n = 1, \ldots, N$, $t = 1, 2, \ldots, T$, and M is number of bands)
Output: $c_{n,t}$ ($n = 1, \ldots, N$, $t = 1, 2, \ldots, T$)
Initialize $\mathbb{W}^1, \ldots, \mathbb{W}^L$ [45]
for $i \leftarrow 1$ to $\mathcal{I}$ **do**
  for $t \leftarrow 1$ to T **do**
    Extract feature $y_{n,t} \in \mathbb{R}^K$ ($n = 1, \ldots, N$)
    $c_{n,t} \leftarrow \arg\max_{k \in K} y_{n,t}(k)$
    $c'_{n,t} \leftarrow \text{spatialRefinement}(c_{n,t})$
    if number of distinct $c'_{n,t} \leq \mathcal{K}$ **then**
      goto *FinalLabel*
    Estimate $\mathcal{L}_t$
  for $t \leftarrow 1$ to $T - 1$ **do**
    Estimate $\mathcal{L}_{t,t+1}$
  $\mathcal{L} \leftarrow \sum_{t=1}^{T} \mathcal{L}_t + \sum_{t=1}^{T-1} \mathcal{L}_{t,t+1}$
  $\mathbb{W}^1, \ldots, \mathbb{W}^L \leftarrow \text{update}(\mathcal{L})$
FinalLabel: Estimate $c_{n,t}$ ($n = 1, \ldots, N$, $t = 1, 2, \ldots, T$)

V. Use of Deep Joint Segmentation for CD

The network trained for multitemporal joint segmentation can be used as a bitemporal deep feature extractor in a DCVA framework [5] to distinguish the changed pixels (Ω_c) from the unchanged ones (ω_{nc}). This is based on the assumption that the network has learned semantic attributes from images while iteratively refining the multitemporal segmentation maps. For bitemporal CD, we focus, here, on two images X_1 and X_2, and the same methodology can be applied to any other pair of images from $\mathcal{X}$. X_1 and X_2 are separately processed through the trained network that is used as a deep feature extractor in this step. Deep features are extracted from $L-1$ convolutional layers of the trained network to form a deep feature hypervector that captures the semantic in a hierarchical fashion for CD. An automatic variance-based layerwise feature selection strategy is applied [5], and the deep change hypervector (G) is obtained as a concatenation of the deep-feature-differences of all considered layers. A deep magnitude ρ is obtained as the Euclidean norm of deep change hypervector G. As unchanged pixels (ω_{nc}) generate similar deep features, while changed pixels (Ω_c) generate dissimilar deep features, we distinguish changed pixels (Ω_c) from the unchanged ones (ω_{nc}) by using a threshold applied to ρ [5], [46].

VI. Experimental Validation of Multitemporal Semantic Segmentation

A. Evaluation Criteria

Due to the possible multiple satisfactory results, unsupervised segmentation can be considered as a subjective task. To objectively evaluate the performance of the proposed method for a particular data set, we fix a target class and evaluate the performance of the proposed joint segmentation method in terms of how well the target class is detected in a single cluster on each of the multitemporal images. Considering target class pixels as positive and all other pixels as negative, we compute sensitivity (over the positive pixels) and accuracy (over all pixels). This is based on the assumption that a good semantic segmentation approach will assign all pixels belonging to the target class the same label and different labels to the others. Considering that there are N_+ and N_- number of positive and negative pixels ($N_+ + N_- = N$) and N_+^* ($N_+^* \leq N_+$) and N_-^* ($N_-^* \leq N_-$) are identified correctly, the sensitivity is defined as

$$\text{Sensitivity} = \frac{N_+^*}{N_+}. \tag{9}$$

The accuracy is defined as

$$\text{Accuracy} = \frac{N_+^* + N_-^*}{N}. \tag{10}$$

B. Baseline Methods

To the best of our knowledge, the problem statement is novel, and hence, we cannot directly compare it with existing methods. We have designed a baseline method using the single-image deep segmentation method in [15] and transfer learning. In more detail, the transfer learning-based approach applies the single-image deep segmentation method individually on one of the images in time series. The method learns a deep model. Considering that other images in the time series are similar (since they are coregistered images from different times and changes have a low prior probability), the transfer-learning-based method applies the same deep model on the

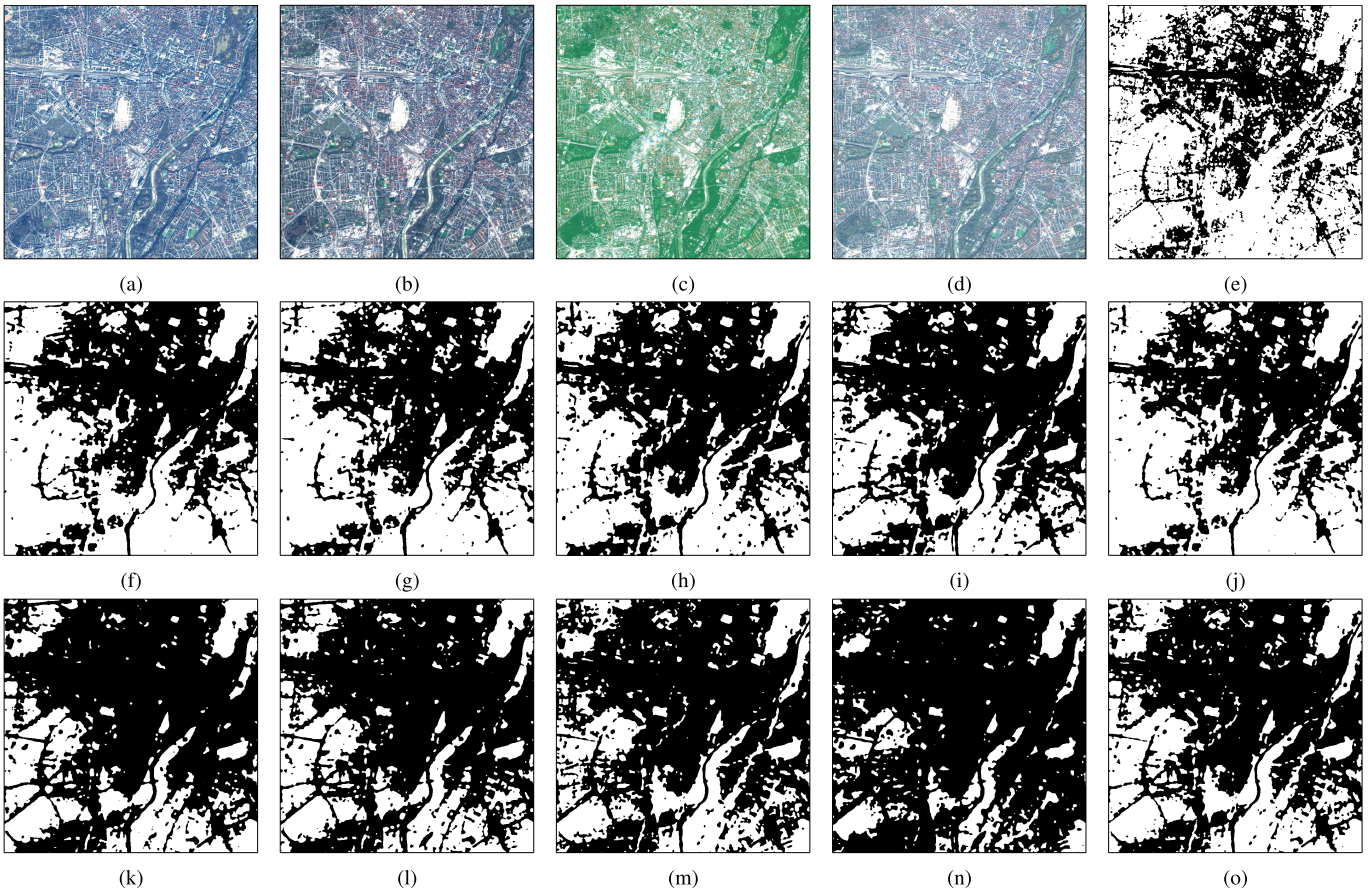


Fig. 2. Munich Sentinel2 Images—input images (RGB). (a) Winter. (b) Spring. (c) Summer. (d) Autumn. (e) Reference urban mask. Segmentation using proposed method: (f) winter, (g) spring, (h) summer, (i) autumn, and (j) four seasons combined. Segmentation using transfer learning: (k) winter, (l) spring, (m) summer, (n) autumn, and (o) four seasons combined. Black: urban pixels.

other images of the time series (without training on them) to obtain segmentation maps. However, this approach does not exploit the temporal correlation in its training phase. Since this transfer learning-based approach learns the model individually on one of the images in time series, it has T times less time requirement than the proposed joint segmentation-based method.

C. Test Data Sets

We used three quasi-urban data sets comprising of HR (Sentinel-2) and VHR (Pleiades) images for multitemporal segmentation showing increasing complexity to test the proposed method in different working conditions. The two HR data sets show comparable complexity in terms of structural and geometrical information of the quasi-urban areas: Munich and Paris. However, the prominent presence of fog/cloud in the Paris data set makes the latter one slightly more complex because of the temporal variations in atmospheric conditions. The VHR (Pleiades) data set is acquired over Trento, Italy, and is more complex due to higher spatial correlation in VHR images.

1) *Munich Data Set*: It is built using HR urban images from the Sentinel-2 sensor (10m/pixel) consisting of four images acquired in winter, spring, summer, and autumn of 2017 [47]. We used only R, G, B, and NIR channels (channel numbers: 4, 3, 2, and 8). The images capture the central area of Munich and its surroundings and show an area of 718 $\times$ 718 pixels. Winter, spring, summer, and autumn images are shown in Fig. 2(a)–(d), respectively. They show some seasonal change despite no substantial change on the ground. The summer image shows little cloud/fog, whereas the other three do not show any substantial amount of cloud. A reference urban mask for the area is shown in Fig. 2(e).
2) *Paris Data Set*: It is built using HR urban images from the S2 sensor (10 m/pixel) consisting of two images acquired in spring and summer 2017 [47]. The images capture a dominantly urban area of Paris also showing some vegetation and the Seine river and cover an area of 718×718 pixels. Summer and spring images are shown in Fig. 3(a) and (b). They show significant seasonal change but no significant change on the ground. The spring image shows a substantial amount of cloud/fog. A reference urban mask for the area is shown in Fig. 3(c). We used only R, G, B, and NIR channels (channel numbers: 4, 3, 2, and 8).
3) *Trento Data Set*: It is built using VHR urban images acquired using the Pleiades sensor (0.5 m/pixel) on August 2012 [see Fig. 4(a)] and September 2013

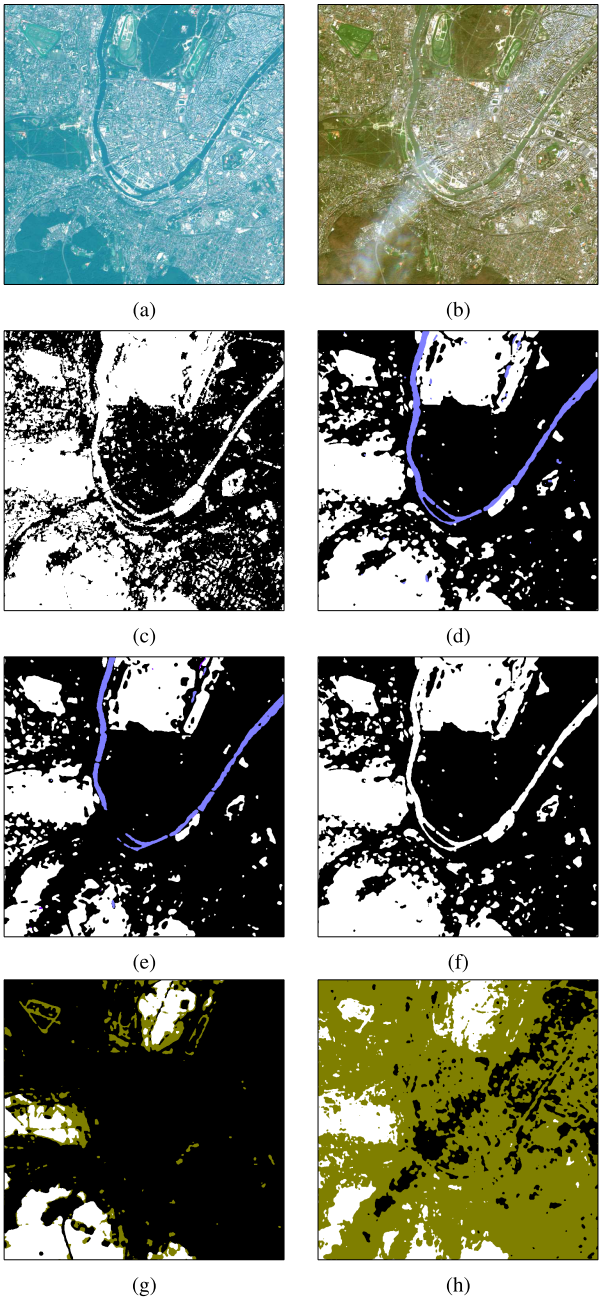


Fig. 3. Paris Sentinel2 images—input images (RGB). (a) Summer. (b) Spring. (c) Reference urban mask. Segmentation using proposed method: (d) summer, (e) spring, and (f) two seasons combined. Segmentation using transfer learning approach: (g) summer and (h) spring. Black: urban pixels.

[see Fig. 4(b)]. The images show an area of 512 × 512 pixels and capture a residential area in Ravina locality of Trento composed of buildings, roads, and small vegetation patches. The building mask for the area is obtained using photointerpretation and direct knowledge of the area and is shown in Fig. 4(c).

D. Munich Data Set

The segmentation map obtained by the proposed method in winter, spring, summer, and autumn images are shown in Fig. 2(f)–(i), respectively. They were trained with the following parameters: $L = 4$, $\mathcal{I} = 500$, and $\mathcal{K} = 2$. The black class corresponds to the urban mask. The winter segmentation map achieves a sensitivity of 90.91% and an accuracy of 81.45%. The spring, summer, and autumn segmentation maps show similar results (see Table I). All segmentation maps agree on most of the analyzed scene, and we illustrate a common urban mask obtained by their intersection [see Fig. 2(j)].

TABLE I
SEGMENTATION RESULTS FOR MUNICH DATA SET

Method	Category	Sensitivity	Accuracy
Proposed	**Winter**	90.91%	81.45%
	Spring	91.05%	82.22%
	Summer	97.03%	81.54%
	Autumn	98.12%	75.64%
	Combined	88.59%	84.23%
Transfer learning	**Winter**	99.30%	67.22%
	Spring	99.34%	69.02%
	Summer	99.81%	69.92%
	Autumn	99.94%	55.87%
	Combined	98.96%	74.87%

TABLE II
SEGMENTATION RESULTS FOR PARIS DATA SET

Method	Category	Sensitivity	Accuracy
Proposed	**Summer**	97.19%	84.91%
	Spring	99.03%	77.70%
	Combined	96.99%	85.62%
Transfer learning	**Summer**	99.98%	68.75%
	Spring	24.21%	59.68%
	Combined	24.21%	59.75%

We compare the proposed method to the transfer learning-based method. For this approach, we apply single-time segmentation on winter image and, subsequently, the learned network on the other three images. The results obtained by this approach are shown in Fig. 2(k) (winter), (l) (spring), (m) (summer), and (n) (autumn). The sensitivity and the accuracy obtained by them are given in Table I. The transfer learning approach obtains slightly oversegmented results; however, the results are still comparable to those obtained by the proposed method.

E. Paris Data Set

The segmentation map obtained by the proposed method on summer and spring images are shown in Fig. 3(d) and (e), respectively. They were trained with the following parameters: $L = 4$, $\mathcal{I} = 500$, and $\mathcal{K} = 2$. As evident from the figures, the black class corresponds to the urban mask. The quantitative results are shown in Table II. The summer segmentation map achieves a sensitivity of 97.19% and an accuracy of 84.91%. The spring segmentation map achieves a sensitivity of 99.03% and an accuracy of 77.70%. We observe some noticeable difference in the segmentation map between summer and spring. However, it can be clearly observed that this difference is due to strong fog/cloud in the spring image [see Fig. 3(b)]. We illustrate a common urban mask obtained by their intersection [see Fig. 3(f)]. The segmentation maps obtained by the transfer learning approach on summer and spring images are shown in Fig. 3(g) and (h), respectively. For transfer learning

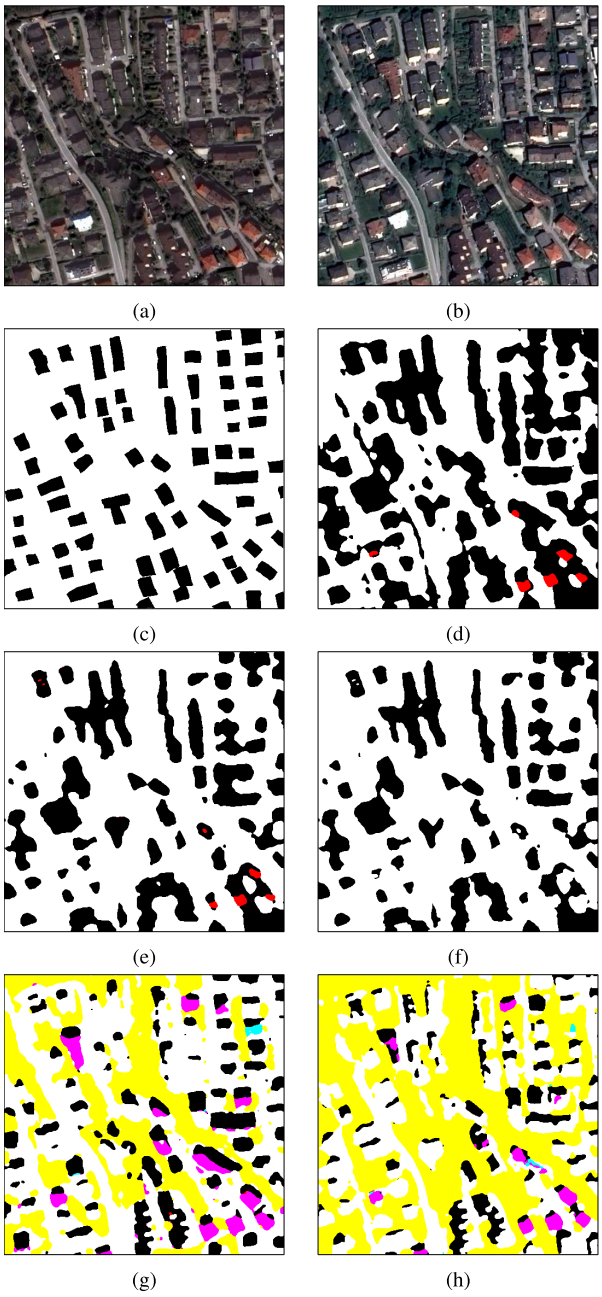


Fig. 4. Trento Pleiades Images—input images (RGB): (a) 2012 and (b) 2013. (c) Reference building mask. Segmentation using proposed method: (d) 2012, (e) 2013, and (f) 2012 and 2013 combined. Segmentation using transfer learning approach: (g) 2012 and (h) 2013. Black: building pixels.

approach, we learn the segmentation network on summer image and, subsequently, use the trained network on spring image. As described previously in Section VI-B, this process can be described as single-time segmentation being applied on the summer image, and the learned network is used to extract segmentation map from the spring image. Using this approach, the summer segmentation map achieves a sensitivity of 99.98% and an accuracy of 68.75%, and the spring segmentation map achieves a sensitivity of 24.21% and an accuracy of 59.68%. We observe that single-time segmentation on the summer image obtains comparable results to the proposed joint segmentation method; however, the transfer learning approach fails to effectively propagate the segmentation mask to the spring image. This clearly shows the advantage of temporal learning. While the proposed joint segmentation approach clearly learns to assign the same class to most pixels in summer and winter images, the reference method (that does not exploit temporal correlation) fails to do so.

TABLE III
SEGMENTATION RESULTS FOR TRENTO DATA SET

Method	Category	Sensitivity	Accuracy
Proposed	**2012**	89.01%	73.69%
	2013	80.12%	84.40%
	Combined	76.68%	84.85%
Transfer learning	**2012**	41.02%	80.02%
	2013	28.75%	80.58%
	Combined	18.97%	80.07%

TABLE IV
CD RESULTS FOR THE TRENTO CD DATA SET

Method	Sensitivity	Specificity	Overall Accuracy
PCVA	0.66	0.90	89.56%
DCVA	0.51	0.96	93.57%
Proposed	0.69	0.98	97.15%

F. Trento Data Set

The segmentation maps obtained by the proposed method in 2012 and 2013 images are shown in Fig. 4(d) and (e), respectively. We set the parameters as $L = 5$, $\mathcal{I} = 500$, and $\mathcal{K} = 2$. The black class corresponds to the building mask. As shown in Table III, the 2012 segmentation map achieves a sensitivity of 89.01% and an accuracy of 73.69%. The 2013 segmentation map achieves a sensitivity of 80.12% and an accuracy of 84.40%. Combining building class in 2012 and 2013 segmentation maps by their intersection, we obtain a common building mask [see Fig. 4(f)]. It obtains a sensitivity of 76.68% and an accuracy of 84.85%. The segmentation maps obtained by the transfer learning approach on 2012 and 2013 images are shown in Fig. 4(g) and (h), respectively. For transfer learning approach, we learn the segmentation network on 2012 image and, subsequently, use the trained network on 2013 image. Using the transfer learning approach, the 2012 segmentation map achieves a sensitivity of 41.02% and an accuracy of 80.02%, and the 2013 segmentation map achieves a sensitivity of 28.75% and an accuracy of 80.58%. We observe that not only the segmentation process could not be propagated to the 2013 image from the 2012 image but also the segmentation is not satisfactory for the 2012 image. This clearly shows the advantage of temporal learning. The joint segmentation-based approach exploits multitemporal information, and it is useful to identify the dominant class in the target multitemporal images.

To further demonstrate the capability of the proposed method to learn useful semantic features, we visualize two features from the last convolutional layer in Fig. 5. It can be seen that they show useful visual concepts to distinguish buildings from the background.

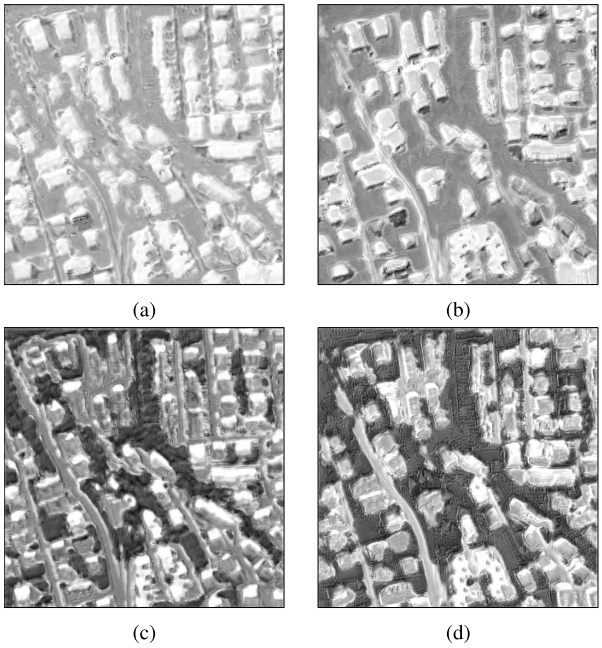


Fig. 5. Visualization of two randomly chosen features from the last convolutional layer on Trento Pleiades Images—feature 1: (a) 2012 and (b) 2013; feature 2: (c) 2012 and (d) 2013.

VII. Experimental Validation of CD

We show the results for CD on a VHR multisensor data set acquired over Trento, Italy. The Trento CD data set [48] is built using VHR urban images acquired using two different optical sensors—Quickbird (acquired in July 2006 with 14.1° off-nadir angle) and Pleiades (acquired in September 2013 with 20.9° off-nadir angle). Thus, they show a temporal difference of seven years, and they are acquired in different seasons. The prechange Quickbird image [see Fig. 6(a)] originally shows a 0.6-m/pixel resolution, and the postchange Pleiades image [see Fig. 6(b)] shows a 0.5-m/pixel resolution. Images are projected to the same spatial resolution of 0.5 m/pixel. The reference CD map obtained using photointerpretation along with knowledge of the analyzed area is shown in Fig. 6(c). This data set shows similar characteristics as the third data set in Section VI in terms of structural and geometrical information but higher complexity because of being multisensor (Quickbird-Pleiades) and showing multitemporal differences due to changes on the ground.

For CD, the sensitivity is defined as the accuracy of correctly detecting the changed pixels, and the specificity is defined as the accuracy of correctly detecting the unchanged pixels [5]. Considering the proposed method, the CD result is compared with the state-of-the-art unsupervised deep learning-based DCVA method [5] and segmentation-based Parcel CVA (PCVA) method [4].

We show the CD map obtained by the proposed method in Fig. 6(d). The proposed method detects most of the changed pixels and few false alarms. As detailed in Table IV, the proposed method outperforms the DCVA method [5] [see Fig. 6(e)] and the PCVA method [4] [see Fig. 6(f)].

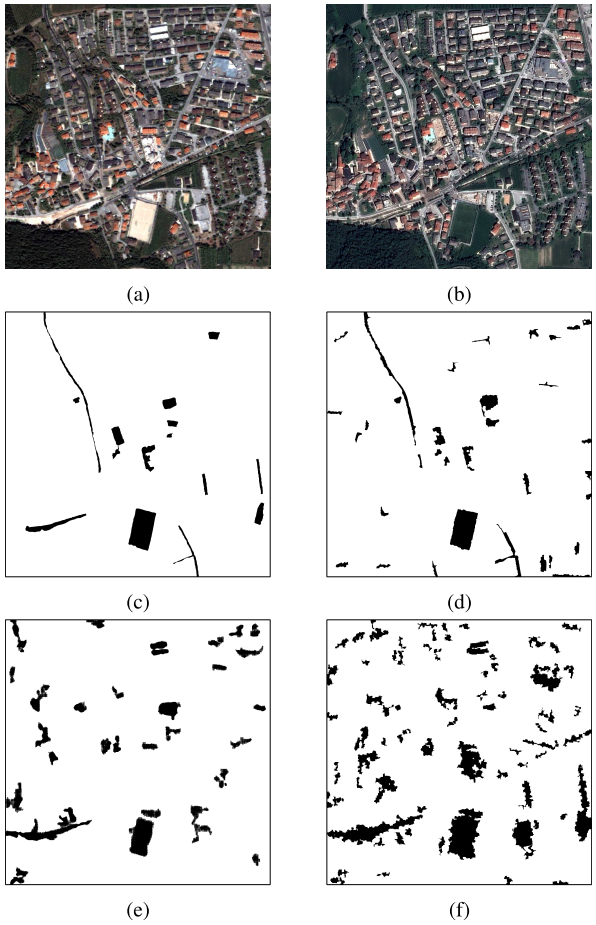


Fig. 6. Trento CD data set. (a) Prechange Quickbird image (RGB). (b) Postchange Pleiades image (RGB). CD maps: (c) reference and (d) proposed, (e) DCVA, and (f) PCVA methods.

VIII. Conclusion

In this article, we proposed a deep unsupervised multitemporal semantic segmentation method. The proposed method works directly on multitemporal images and does not require any labeled training pixel or the availability of abundant unlabeled multitemporal images for training. The proposed method represents the multitemporal images using multitemporal deep features obtained by a trainable deep network. The weights of the trainable network are adjusted in iterations along with the predicted multitemporal labels. Thus, the proposed method is able to jointly optimize the deep feature representation and the multitemporal label assignment ensuring consistency between multitemporal labels. The results show that the obtained labels are semantically meaningful and temporally consistent. Moreover, the proposed method can work on images from different seasons that show significant seasonal differences. To the best of our knowledge, this is the first work to explore multitemporal joint segmentation. We also detailed the extension of the method for CD by following the DCVA approach.

In the future, we plan to extend the proposed method for other sensors, e.g., passive (SAR) sensors. Contrary to the multispectral images, SAR images follow the multiplicative noise model [38] and emphasize the physical properties of the

target surfaces, while the optical images highlight the structural details. Thus, the adaptation of the proposed method for SAR images is expected to be challenging, e.g., modification may be needed in terms of spatial homogeneity measurement and loss function. We also plan to extend the proposed method to multisensor time series comprising of images acquired by active and passive sensors.

REFERENCES

[1] F. Bovolo and L. Bruzzone, "The time variable in data fusion: A change detection perspective," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 8–26, Sep. 2015.

[2] M. Volpi, D. Tuia, F. Bovolo, M. Kanevski, and L. Bruzzone, "Supervised change detection in VHR images using contextual information and support vector machines," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 20, pp. 77–85, Feb. 2013.

[3] 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.

[4] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 33–37, Jan. 2009.

[5] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[6] Y. T. Solano-Correa, F. Bovolo, L. Bruzzone, and D. Fernandez-Prieto, "Spatio-temporal evolution of crop fields in Sentinel-2 satellite image time series," in *Proc. 9th Int. Workshop Anal. Multitemporal Remote Sens. Images (MultiTemp)*, Jun. 2017, pp. 1–4.

[7] L. Wu, Z. Zhang, Y. Wang, and Q. Liu, "A segmentation based change detection method for high resolution remote sensing image," in *Proc. Chin. Conf. Pattern Recognit.* Berlin, Germany: Springer, 2014, pp. 314–324. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-662-45646-0_32

[8] O. Tasar, Y. Tarabalka, and P. Alliez, "Incremental learning for semantic segmentation of large-scale remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3524–3537, Sep. 2019.

[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] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Appl. Soft Comput.*, vol. 70, pp. 41–65, Sep. 2018.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[14] X. Xia and B. Kulis, "W-Net: A deep model for fully unsupervised image segmentation," 2017, *arXiv:1711.08506*. [Online]. Available: http://arxiv.org/abs/1711.08506

[15] A. Kanezaki, "Unsupervised image segmentation by backpropagation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1543–1547.

[16] S. Saha, S. Sudhakaran, B. Banerjee, and S. Pendurkar, "Semantic guided deep unsupervised image segmentation," in *Proc. Int. Conf. Image Anal. Process.*, Cham, Switzerland: Springer, 2019, pp. 499–510. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-030-30645-8_46#citeas

[17] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[19] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*. [Online]. Available: http://arxiv.org/abs/1606.02147

[20] 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.

[21] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features Off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.

[22] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2107–2116.

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2012). *The PASCAL Visual Object Classes Challenge 2012 (VOC) Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[26] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[27] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[28] X. Ma, J. Geng, and H. Wang, "Hyperspectral image classification via contextual deep learning," *EURASIP J. Image Video Process.*, vol. 2015, no. 1, pp. 1–12, Dec. 2015.

[29] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution semantic labeling with convolutional neural networks," 2016, *arXiv:1611.01962*. [Online]. Available: http://arxiv.org/abs/1611.01962

[30] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, no. 4, Sep. 2017, Art. no. 042609.

[31] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.

[32] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 24–41, Jun. 2016.

[33] F. Gao, J. Dong, B. Li, and Q. Xu, "Automatic change detection in synthetic aperture radar images based on PCANet," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1792–1796, Dec. 2016.

[34] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 6, p. 506, 2016.

[35] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning Spectral–Spatial–Temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[36] J. Geng, H. Wang, J. Fan, and X. Ma, "Change detection of SAR images based on supervised contractive autoencoders and fuzzy clustering," in *Proc. Int. Workshop Remote Sens. Intell. Process. (RSIP)*, May 2017, pp. 1–3.

[37] Y. Xu, S. Xiang, C. Huo, and C. Pan, "Change detection based on auto-encoder model for VHR images," *Proc. SPIE*, vol. 8919, Oct. 2013, Art. no. 891902.

[38] S. Saha, F. Bovolo, and L. Bruzzone, "Destroyed-buildings detection from VHR SAR images using deep features," *Proc. SPIE*, vol. 10789, Oct. 2018, Art. no. 107890Z.

[39] X. Zhang, P. Xiao, X. Feng, and M. Yuan, "Separate segmentation of multi-temporal high-resolution remote sensing images for object-based change detection in urban area," *Remote Sens. Environ.*, vol. 201, pp. 243–255, Nov. 2017.

[40] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, p. 1382, 2019.

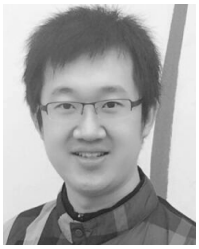
[41] Y. T. S. Correa, F. Bovolo, and L. Bruzzone, "Change detection in very high resolution multisensor optical images," *Proc. SPIE*, vol. 9244, Oct. 2014, Art. no. 924410.

[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[43] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022
[44] A. Rusiecki, "Trimmed categorical cross-entropy for deep learning with label noise," *Electron. Lett.*, vol. 55, no. 6, pp. 319–320, Mar. 2019.
[45] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, pp. 249–256, 2010.
[46] S. Saha, F. Bovolo, and L. Brurzone, "Unsupervised multiple-change detection in VHR optical images using deep features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, Jul. 2018, pp. 1902–1905.
[47] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Aggregating cloud-free Sentinel-2 images with Google Earth engine," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 145–152, Sep. 2019.
[48] 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. IGARSS*, Jul. 2019, pp. 5033–5036.

Sudipan Saha (Graduate Student Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the Institute of Engineering and Management, Kolkata, India, in 2011, and the M.Tech. degree in electrical engineering from IIT Bombay, Mumbai, India, in 2014. He is pursuing the Ph.D. degree in information and communication technologies with the University of Trento, Trento, Italy, and Fondazione Bruno Kessler, Trento.

He worked as an Engineer with TSMC Limited, Taiwan, from 2015 to 2016. In 2019, he was a Guest Researcher with the Technical University of Munich (TUM), Munich, Germany. His research interests are related to multitemporal remote sensing image analysis, domain adaptation, time-series analysis, image segmentation, deep learning, image processing, and pattern recognition.

Mr. Saha is a Reviewer for several international journals.



Lichao Mou received the Bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the Master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.-Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

In 2015 he spent six months at the Computer Vision Group at the University of Freiburg in Germany. In 2019 he was a Guest Researcher with the University of Cambridge, UK. From 2016 to 2020, he worked toward his Ph.D. at the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany, where he is currently a Research Scientist. He is also an AI consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU), Munich. His research interests are algorithms for remote sensing data analysis and visual learning and reasoning tasks. His work explores topics in machine/deep learning, remote sensing, and computer vision.

He was the recipient of the first place in the 2016 IEEE GRSS Data Fusion Contest and finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and 2019 Joint Urban Remote Sensing Event.

Chunping Qiu (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, China, in 2013 and 2016, respectively. She is pursuing the Ph.D. degree with the Technical University of Munich (TUM), Munich, Germany.

In 2019, she was a Guest Researcher with the Telecommunications and Remote Sensing Laboratory, University of Pavia, Pavia, Italy. Her research interest is focused on deep learning and remote sensing data fusion with an application on urban land cover classification and analysis.

Xiao Xiang Zhu (Senior Member, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is a Professor of signal processing in Earth observation with the Technical University of Munich (TUM), Munich, Germany, and the Head of the Department EO Data Science, Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Since 2019, she has been co-coordinating the Munich Data Science Research School. She is also leading the Helmholtz Artificial Intelligence Cooperation Unit (HAICU)-Research Field Aeronautics, Space and Transport. She was a Guest Scientist or Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan, and the University of California, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of the young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

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

Until 2013, she was a Research Fellow with the University of Trento. She is the Founder and the Head of the 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. Her research interests include remote sensing image processing, multitemporal remote sensing image analysis, change detection in multispectral, hyperspectral, synthetic aperture radar images, very high-resolution images, time-series analysis, content-based time-series retrieval, domain adaptation, and light detection and ranging (LiDAR) and radar sounders. She conducts research on these research topics within the context of several national and international projects.

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

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

He is a Full Professor of telecommunications with the University of Trento, Trento, Italy, where he teaches remote sensing, radar, and digital communications. He is the Founder and the Director of the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. His research interests are in the areas of remote sensing, radar and SAR, signal processing, machine learning, and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the Principal Investigator of many research projects. Among the others, he is the Principal Investigator of the Radar for icy Moon exploration (RIME) instrument in the framework of the JUpiter ICy 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 the author (or a coauthor) of 259 scientific publications in refereed international journals (193 in IEEE journals), more than 330 papers in conference proceedings, and 22 book chapters. He is the Editor/Coeditor of 18 books/conference proceedings and one scientific book. His papers are highly cited, as proven from the total number of citations (more than 31 600) and the value of the h-index (83) (source: Google Scholar). He was invited as a keynote speaker in more than 40 international conferences and workshops.

Dr. Bruzzone has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS), where since 2019, he has been a Vice-President for Professional Activities. He has ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, July 1998. Since that he was a 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 Coeditor 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 (MultiTemp) series and is 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 *IEEE Geoscience and Remote Sensing Magazine* for which he has been Editor-in-Chief from 2013 to 2017. He is 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 from 2012 to 2016.