Self-Supervised Multisensor Change Detection

Sudipan Saha, *Member, IEEE*, Patrick Ebel, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Most change detection (CD) methods assume that prechange and postchange images are acquired by the same sensor. However, in many real-life scenarios, e.g., natural disasters, it is more practical to use the latest available images before and after the occurrence of incidence, which may be acquired using different sensors. In particular, we are interested in the combination of the images acquired by optical and synthetic aperture radar (SAR) sensors. SAR images appear vastly different from the optical images even when capturing the same scene. Adding to this, CD methods are often constrained to use only target image-pair, no labeled data, and no additional unlabeled data. Such constraints limit the scope of traditional supervised machine learning and unsupervised generative approaches for multisensor CD. The recent rapid development of self-supervised learning methods has shown that some of them can even work with only few images. Motivated by this, in this work, we propose a method for multisensor CD using only the unlabeled target bitemporal images that are used for training a network in a self-supervised fashion by using deep clustering and contrastive learning. The proposed method is evaluated on four multimodal bitemporal scenes showing change, and the benefits of our self-supervised approach are demonstrated. Code is available at https://gitlab.lrz.de/ai4eo/cd/-/tree/main/sarOpticalMultisensorTgrs2021.**

***Index Terms*— Change detection (CD), deep learning, multisensor analysis, self-supervised learning.**

I. Introduction

OUR earth is rapidly changing, both due to natural and man-made causes. Satellite image-based change detection (CD) is generally used to monitor the temporal evolution of the dynamic earth [1]–[7]. CD ingests bitemporal images as input and segregates all pixels as changed/unchanged. CD is a crucial step for several applications, including disaster management, urban monitoring, forestry, glacier monitoring, and precision agriculture. Considering the variation of applications, rarity of occurrences of some change-inducing incidents (e.g., natural disasters), and large geographic variation, it is imprudent to assume that large-scale training datasets corresponding to all such tasks can be ever collected. Thus, there is a significant inclination in the CD literature toward methods that can process the target bitemporal region-of-interest without using any training label or any additional pool of unlabeled images. Motivated by its excellent performance in computer vision, researchers have applied deep learning to satellite image CD [8]. To exploit the potential of deep learning while not using any training label or additional unlabeled images, transfer learning-based CD methods are popular, which reuse a pretrained network for bitemporal feature extraction and comparison [1].

A striking feature of satellite data is its variability, in terms of different sensors. Images captured using a passive optical sensor are quite similar to the natural images studied in computer vision. However, images captured by the active sensors, e.g., synthetic aperture radar (SAR), are remarkably different from the optical images [9]–[11]. While optical sensors use wavelengths near visible light (approx. 1 μm), SAR uses a wavelength of 1 cm to 1 m. Moreover, optical sensors rely upon the natural illumination (e.g., sun) to create the brightness observed by the sensor, while the SAR sensors carry their own illumination source, in the form of radio waves transmitted by an antenna. Moreover, satellite images are captured with a different number of spectral bands (one to a few hundred), different spatial resolutions (few cm/pixel to Km/pixel), and different polarizations. While this vast variation provides an opportunity for detailed earth observation, it is not trivial to use the same set of methods for images from different sensors. Due to this reason, most existing CD methods assume that the prechange and postchange images are acquired using the same sensor. The temporal frequency at which the same sensor can image the same place depends on the revisit period of the satellite on which the sensor is mounted. However, the better the spatial resolution, the more close the satellite is to the earth, and the more time it takes to revisit the same place. This is a hindrance in the use of same-sensor CD in time-bound applications, e.g., fast response for disaster management and precision agriculture. Using different sensors may allow us to obtain temporal sequences with better temporal frequency without sacrificing spatial resolution. However, it is not trivial to process multisensor bitemporal images as they are affected by the spectral characteristics of the sensors. Moreover, different sensors capture a different type of information, making

Manuscript received May 10, 2021; revised June 23, 2021 and July 21, 2021; accepted July 25, 2021. This work was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. [ERC-2016-StG-714087], Project acronym: So2Sat), in part by the Helmholtz Association through the Framework of Helmholtz Artificial Intelligence (AI)—Local Unit "Munich Unit at Aeronautics, Space and Transport (MASTr)" under Grant ZT-I-PF-5-01, in part by the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, and in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001. *(Corresponding author: Xiao Xiang Zhu.)*

Sudipan Saha and Patrick Ebel are with the Department of Aerospace and Geodesy, Data Science in Earth Observation, Technical University of Munich, 85521 Ottobrunn, Germany (e-mail: sudipan.saha@tum.de; patrick.ebel@tum.de).

Xiao Xiang Zhu is with the Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Weßling, Germany, and also with the Department of Aerospace and Geodesy, Data Science in Earth Observation, Technical University of Munich, 85521 Ottobrunn, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Digital Object Identifier 10.1109/TGRS.2021.3109957

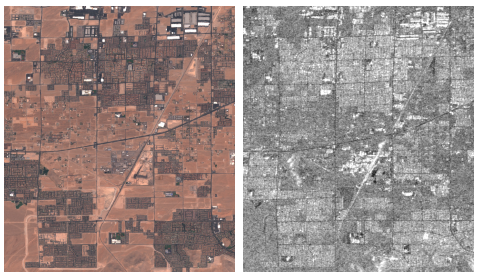
their comparison often challenging [12]. The difficulty of this problem is further accentuated by the fact that we are interested to detect change without using any labeled training data or any abundant pool of unlabeled data.

The emergence of deep learning has seen many such problems solved that were thought to be very challenging in the past [13], [14]. Self-supervised learning has shown remarkable success recently, even when only few images are available [15]. Intrigued by this, in this article, we explore the challenging problem of CD between optical and SAR images, the disparity between which is evident in Fig. 1. We exploit recent developments in the self-supervised learning and deep clustering to propose a method for challenging SAR-optical CD where one of the bitemporal images is acquired by an optical sensor, while the other is acquired by an SAR sensor.

The proposed method requires only the bitemporal target scene (where change is to be detected), no training label, and no additional unlabeled data. The target bitemporal scene is typically large, few hundred pixels by few hundred pixels. Smaller bitemporal patches (e.g., 64×64) are extracted from it to train a two-branch network, similar to the Siamese network [16]. Each branch of the network has a projection module and a predictor. Projection modules learn features unique to optical and SAR data without sharing weights, while predictors share the weight. The output of the predictors is used to estimate deep clustering loss for both images separately. Moreover, considering that the prior probability of changed pixels is much less than the unchanged ones, a temporal consistency loss is proposed, which ensures that pixels in the same location at two different times tend to get the same label. To ensure that this does not lead the network to learn a trivial solution, a contrastive loss is used. By the combination of these losses, the proposed method learns useful semantic features from the multisensor (SAR-optical) bitemporal target scene, and after training, the network predictions can be compared for CD.

The contributions of this article are given as follows.

1) We propose a self-supervised learning method for CD in a bitemporal scene where one image is captured by the optical sensor and the other by the SAR sensor. The proposed method, only exploiting the available target unlabeled scene, effectively absorbs several concepts from the recent self-supervised learning literature, e.g., deep clustering, augmented view, Siamese network, and contrastive learning. By effectively exploiting these concepts and modifying them appropriately for the target multisensor bitemporal data, the proposed method is able to train a network that is further used for bitemporal comparison and CD.
2) We show the versatility of self-supervised learning on spatiotemporal satellite data that are very different from typical computer vision images. Even though some form of aerial images (e.g., drone images) is often studied in computer vision, we stress that our satellite data (both optical and SAR) are significantly different from the typical aerial images.



(a) (b)

Fig. 1. Visual contrast for Las Vegas between (a) optical image (prechange) and (b) SAR image (postchange). Optical and SAR images emphasize different properties of the target area, thus performing CD on them is challenging.

3) We experimentally show the efficacy of the proposed method on four different bitemporal multisensor scenes.

The rest of this article is organized as follows. Related works are briefly discussed in Section II. Section III outlines the proposed method. Datasets and experimental results are detailed in Section IV. Finally, we conclude this article in Section V.

II. Related Work

In this section, we briefly discuss existing works on unsupervised CD (with a focus on the multisensor CD) and self-supervised learning.

A. Change Detection

Prior to the emergence of deep learning, most unsupervised CD methods used the concept of pixelwise image differencing, i.e., change vector analysis (CVA) [17]. A number of superpixels and spatial neighborhood-based variants of CVA have been proposed, e.g., parcel change vector analysis (PCVA) [18] and robust change vector analysis (RCVA) [19]. Most deep learning-based unsupervised CD methods use transfer learning. Reference [1] proposed deep change vector analysis (DCVA), a CD framework that combines ideas from CVA with feature extraction based on pretrained neural networks. In nutshell, a deep model that has been trained for some other task is reused to obtain pixelwise bitemporal deep features from the target scene. Bitemporal deep features are then compared to obtain deep change hypervectors for each pixel in the scene, which is analyzed based on magnitude (ℓ_2 norm) to identify the changed pixels. While [20] shows that sensor-specific pretrained network is more suitable for transfer learning, [5] advocates models trained on ImageNet [21] for transfer learning in CD. There is another class of unsupervised CD methods that preclassifies some pixels with high confidence as changed/unchanged using some traditional approach and further uses those confident samples for training a CD model [22].

It is not trivial to process multisensor bitemporal images as they are affected by differences in spatial resolution and differences in the spectral characteristics of the sensors. Due to this, there are very few works that can work in the setting where prechange and postchange images have different spatial resolution [23], [24] or bands with different spectral characteristics [25]. Moreover, those works deal

with only minor variations in spatial or spectral characteristics. Saha *et al.* [23] proposed a cycle-consistent generative adversarial network-based method to learn transcoding between multisensor multitemporal domain. However, their work assumes that a large (unlabeled) area corresponding to both sensors is available as training data. Liu *et al.* [26] used a symmetric convolutional coupling network (SCCN), and [27] used denoising autoencoder (DAE) for CD in multisensor images. Though those works considered optical-SAR images, they applied their methods to scenes with limited spatial complexity. While our work is strongly motivated by the existing works on multisensor CD [23], [24], it takes them a step further by considering the challenging scenario of optical-SAR CD in complex urban scenes and, furthermore, by integrating recent developments in self-supervised learning.

B. Self-Supervised Learning

Considering the difficulty of collecting labeled data and the abundance of unlabeled data, machine learning researchers have focused on developing unsupervised and self-supervised deep learning methods in the recent past. Gidaris *et al.* [28] used image rotation as a pretext task to learn unsupervised semantic feature. Several other pretext tasks have been explored in the literature, e.g., relative patch prediction [29] and image inpainting [30]. Deep clustering, i.e., joint learning of the parameters of the deep network and the cluster assignment of the resulting features, has also been shown to be effective for unsupervised representation learning [31]. Remarkably, [15] has shown that the abovementioned unsupervised methods learn useful semantic features even with a single-image input. Contrastive methods function by bringing the representation of different views of the same image ("positive pairs") closer while spreading representations of different images ("negative pairs") apart [32]–[34]. Boostrap your own latent [35] and its variant SiamSiam [16] eliminate the requirement of negative pair by using multiple views of the same image. In more detail, SiamSiam [16] ingests as input two randomly augmented views of an image and processes it through a Siamese architecture. Each Siamese branch consists of an encoder and a prediction head. The encoders share weight between two views.

The proposed method is strongly inspired from the above self-supervised methods. Like deep clustering [31], the proposed method uses the concept of simultaneous representation learning and cluster/label assignment. The bitemporal images can be considered to be views of the same scene, such as SiamSiam [16]. Like the contrastive methods, the proposed method uses the idea of bringing closer the representation of positive pairs and spreading apart the negative pairs. Like [15], the proposed method works on a single scene (a pair of images capturing the same location at two different times).

Multitemporal satellite image processing researchers have also proposed self-supervised representation learning methods, e.g., deep clustering for multitemporal segmentation [36] and learning by rearranging randomly shuffled time-series images [37]. The proposed method is related to them, using the concept of deep clustering as in [36].

III. Proposed Method

Let X_1 and Z_2 be two images of size $R \times C$ taken over the same geographical region at times t_1 and t_2, respectively. Without loss of generality, we assume that the prechange image X_1 is acquired by an optical sensor (RGB), and the postchange image Z_2 is acquired by the SAR sensor. Since SAR image is grayscale, the same channel is replicated thrice to make it three-channel like the optical input. We aim to detect changes from the images X_1 and Z_2 in an unsupervised manner, i.e., without using any training labels and any additional unlabeled data pool. Our goal is to divide the set of all pixels Ω into two subsets Ω_c and ω_{nc} corresponding to changed and unchanged pixels, respectively. Like most existing unsupervised CD methods [1], we assume that the prior probability of occurrence of change is less compared to no change [38].

We can extract a set of bitemporal patches of size $R' \times C'$ ($R' < R$ and $C' < C$) from the images X_1 and Z_2. In practice, one training iteration involves only a batch of $\mathcal{B}$ patches from X_1, denoted as $\mathcal{X} = \{x_1^1, \ldots, x_1^{\mathcal{B}}\}$, and corresponding patches from Z_2, denoted as $\mathcal{Z} = \{z_2^1, \ldots, z_2^{\mathcal{B}}\}$. x_1^b and z_2^b are processed separately with deep clustering loss, as detailed in Section III-C. Furthermore, considering that x_1^b and z_2^b represent same location at two different times and prior probability of change is less, a temporal consistency loss (see Section III-D) is formulated using each such pair. Furthermore, $\mathcal{Z}$ is shuffled to form negative samples $\mathcal{Z}'$, and a contrastive loss is used between pairs from $\mathcal{X}$ and $\mathcal{Z}'$, as outlined in Section III-E. The proposed method is outlined in Fig. 2.

A. Bitemporal Patches are Multiple Views of the Same Location

We recall from Section II-B that many self-supervised learning approaches build upon the concept of bringing closer the representation of the multiple views of the same image. Different views of the same image are generally obtained by different augmentation techniques, e.g., random crops. We argue that multisensor bitemporal patches x_1^b and z_2^b can be similarly thought to be multiple views of the same location. They represent augmentation of the same place, where the augmentation transformation is naturally caused by multisensor differences and other factors, including weather conditions. Considering that the prior probability of change is less [38], most of the time, such a pair of patches x_1^b and z_2^b represent the same information but from the eyes of two different viewers (sensors).

B. Siamese Representation

Since bitemporal patches can be seen as multiple views of the same location, we argue that semantic information can be captured from them by using a Siamese-like architecture. Similar to [16], both branches of the two-branch network have projection modules f_{opt} and f_{sar} for the optical and SAR branch, respectively. In addition, both branches have prediction modules h_{opt} and h_{sar} for the optical and SAR branches, respectively. However, unlike [16], the projection

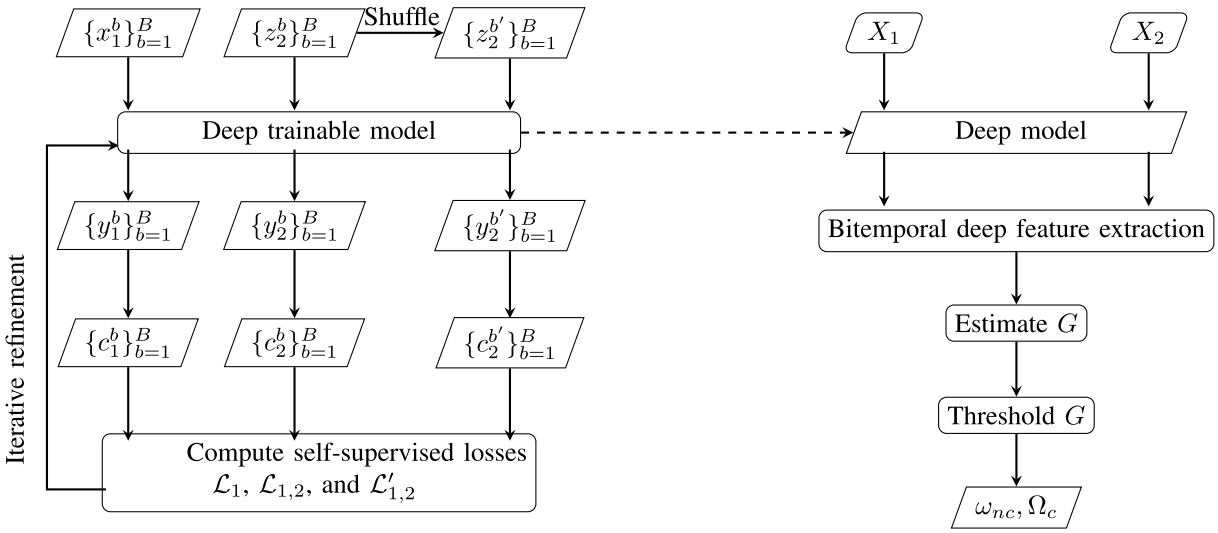


Fig. 2. Proposed unsupervised multisensor (optical-SAR) CD framework. The left-hand side denotes the self-supervised training process, while the right-hand side shows the CD process using already trained model.

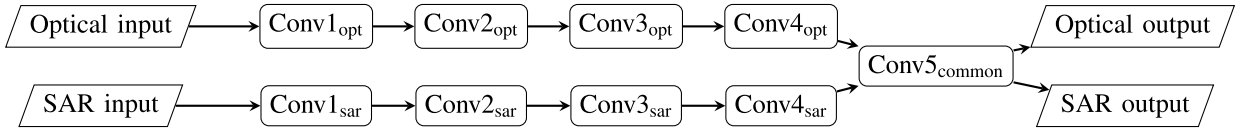


Fig. 3. Network simplified architecture with $L_1 = 4$ and $L_2 = 1$. Optical and SAR inputs are processed separately and subsequently fed to a common prediction layer.

modules f_{opt} and f_{sar} do not share weight. This is because SAR and optical images are significantly different processed by two different projection modules using different sets of weights. However, the prediction modules h_{opt} and h_{sar} share weights and, henceforth, simply denoted as h.

The projection and the prediction networks consist of L_1 and L_2 (generally $L_2 = 1$) convolutional layers, respectively, where $L = L_1 + L_2$. The two projections compute a projected representation from the optical and SAR images and project them to a common domain. In the ideal scenario, where the projectors have perfectly learned to project optical and SAR images into a common domain and the bitemporal images do not show any change, the output generated for an input pair is expected to be identical. However, practically even in absence of any change, there are differences caused by multisensor acquisition and other factors that are not trivial for projection modules to mitigate.

All but the last convolution layers are followed by the ReLU activation function. They are further followed by the batch normalization layer. We do not use any pooling layer; hence, the size of the input is preserved in the output. While filters of spatial size 3×3 are used for all convolution layers for projection, the prediction module uses 1×1 filter. The kernel number of the final layer is K and can be thought of as K different clusters/classes. Each pixel can be assigned to one of these K clusters (as detailed in Section III-C). The network architecture is shown in Fig. 3.

C. Deep Clustering

The deep clustering process involves the joint learning of the parameters of the deep network and the cluster assignment of the resulting features [31]. Deep clustering helps the network to learn discriminative features that can identify different classes/clusters in the images. Considering the processing of the two images as an independent process, deep clustering can be performed for each of them. The output obtained by the network for a paired input patches x_1^b and z_2^b is

$$y_1^b = h\big(f_{\text{opt}}\big(x_1^b\big)\big) \tag{1}$$

$$y_2^b = h\big(f_{\text{sar}}\big(z_2^b\big)\big). \tag{2}$$

y_1^b has same spatial dimension $R' \times C'$ as x_1^b and has kernel number (or, feature dimension) K. The deep clustering process is performed over the pixels, i.e., each pixel is assigned to a cluster. Without loss of generality, we, henceforth, explain the deep clustering process in reference to a generic pixel $y_{1,n}^b$ from y_1^b. The dimension of $y_{1,n}^b$ is K that can be converted to 1-D label $c_{1,n}^b$ by argmax classification. This is achieved by selecting the kernel/feature in $y_{1,n}^b$ that has maximum value. If the kth feature of $y_{1,n}^b$ is represented by $y_{1,n}^b(k)$, then label $c_{1,n}^b$ is obtained as follows:

$$c_{1,n}^b = \arg\max_{k \in K} y_{1,n}^b(k). \tag{3}$$

The rationale behind finding the highest activation of an input pixel is that the pixels that obtain the highest activation in the same feature are likely to have similar semantics, thus belonging to the same group. While there are several possible ways to define the pseudolabel, our approach more closely follows the ones based on argmax classification of the final layer [39], [40]. Once the pixels are assigned to the K clusters, parameters of the deep network can be updated by using a loss between the feature $y_{1,n}^b$ and the cluster $c_{1,n}^b$. We use

cross-entropy loss as

$$\ell_{1,n}^{b} = \text{crossentropy}\big(y_{1,n}^{b}, c_{1,n}^{b}\big). \tag{4}$$

In practice, the loss term $\mathcal{L}_1$ is computed by taking mean of $\ell_{1,n}^{b}$ over all pixels in x_1^b and all patches in the batch ($b = 1, \ldots, \mathcal{B}$). $\mathcal{L}_1$ is used to adjust the weights of h and f_{opt}. Similarly, $\mathcal{L}_2$ is computed from z_2^b ($b = 1, \ldots, \mathcal{B}$) and used to modulate the weights of h and f_{sar}.

While deep clustering helps to learn representation for each sensor separately, they do not ensure that the independently learned features are aligned with each other.

D. Temporal Consistency

Recalling from Section III-B, multisensor bitemporal patches x_1^b and z_2^b are multiple views of the same location in the absence of any change. In other words, in coregistered bitemporal images, pixels in the same spatial location generally tend to belong to the same object as changes have a low prior probability than the unchanged class. Thus, the features computed for the bitemporal paired patches x_1^b and z_2^b should be similar in most cases. For each input pixel $x_{1,n}^b$ and $z_{2,n}^b$, we compute absolute error (AE) loss as

$$\ell_{12,n}^{b} = \big\|y_{1,n}^{b} - y_{2,n}^{b}\big\|_1. \tag{5}$$

A loss term $\mathcal{L}_{1,2}$ is computed by taking the mean of $\ell_{12,n}^{b}$ over all considered pixels for all patches in the batch. The proposed temporal consistency only ensures that the pixels at the same location, however, at two different times, tend to have the same label. This may lead to a degenerate solution where all pixels simply have the same prediction for both times. Moreover, some bitemporal pairs x_1^b and z_2^b may be indeed changed and, however, penalized for producing dissimilar output in this step.

E. Contrastive Learning

While Section III-D encourages the features computed for paired patch x_1^b and z_2^b to be similar, in this section, we encourage the network to produce a dissimilar feature for different inputs by employing concepts inspired by contrastive learning. While we do not have negative samples under the unsupervised setting in which our work is based on, we simply shuffle the batch of patches $\mathcal{Z}$ to $\mathcal{Z}'$. Recall that $\mathcal{X}$ and $\mathcal{Z}$ have location-wise paired patches. This implies that $\mathcal{X}$ and $\mathcal{Z}'$ have unpaired patches. Thus, there should be more dissimilar in comparison to the paired patches in Section III-D. We encourage features computed for x_1^b and $z_2^{b'}$ to be dissimilar. This is achieved by computing (negative) AE loss for each input pixel $x_{1,n}^b$ and $z_{2,n}^{b'}$

$$\ell_{12,n}^{b'} = -\big\|y_{1,n}^{b} - y_{2,n}^{b'}\big\|_1. \tag{6}$$

$\ell_{12,n}^{b'}$ has negative value. Ideally, $\ell_{12,n}^{b'}$ should be encouraged to be more and more negative. However, in practice, we note that simply shuffling $\mathcal{Z}$ to $\mathcal{Z}'$ does not always ensure that $\mathcal{X}$ and $\mathcal{Z}'$ have semantically different patches. Even after shuffling, they may have the semantically paired patches, however penalized in this step for producing similar features. Thus, to control its impact, we penalize the network with $\ell_{12,n}^{b'}$ only when it approaches 0, i.e., $y_{1,n}^{b}$ and $y_{2,n}^{b'}$ become too similar. This is achieved by computing the loss term $\mathcal{L}'_{1,2}$ as mean of exponentials of $\ell_{12,n}^{b'}$ over all considered pixels for all patches in the batch.

Algorithm 1 Self-Supervised Training for Multisensor CD

```
Initialize 𝕎^1, ..., 𝕎^L
for i ← 1 to 𝓘 do
    Sample ℬ patches from X_1, denoted as 𝒳 = {x_1^1, ..., x_1^ℬ}
    Obtain corresponding ℬ patches from Z_2, denoted as 𝒵 = {z_2^1, ..., z_2^ℬ}
    Obtain 𝒵' as random shuffling of 𝒵
    for j ← 1 to 𝓙 do
        for b ∈ ℬ do
            y_1^b = h(f_opt(x_1^b))
            y_2^b = h(f_sar(z_2^b))
            y_2^b' = h(f_sar(z_2^b'))
        end for
        Calculate deep clustering losses ℒ_1, ℒ_2
        Calculate temporal consistency loss ℒ_1,2
        Calculate contrastive loss ℒ'_1,2
        if i ≤ 𝓘_1 then
            Use loss (ℒ_1 + ℒ_2)/2 to modulate 𝕎^1, ..., 𝕎^L
        else
            For each 3 consecutive iterations j, use ℒ_1, ℒ_1,2, and ℒ'_1,2, respectively, to modulate 𝕎^1, ..., 𝕎^L
        end if
    end for
end for
```

F. Overall Loss and Network Refinement

The initialization process [41] is used to initialize all the trainable weights of the network $\mathbb{W}^1, \ldots, \mathbb{W}^L$, corresponding to L layers. For updating of weights, we exploit stochastic gradient descent (SGD) mechanism with momentum [42]. The training process is executed in two different steps of $\mathcal{I}_1$ and $\mathcal{I}_2$ epochs (summing to $\mathcal{I}$). For each batch of data, $\mathcal{J}$ iterations are performed. For the first $\mathcal{I}_1$ epochs, only the sum of deep clustering losses ($\mathcal{L}_1 + \mathcal{L}_2$) is used to modulate the network weights. For subsequent $\mathcal{I}_2$ epochs, in one training iteration, $\mathcal{L}_1$ is used as loss function; in the following iteration, $\mathcal{L}_{1,2}$ is used; and in the following iteration, $\mathcal{L}'_{1,2}$ is used. The combination of three loss functions yields a balanced training process taking into account coherent cluster formation, temporal feature consistency, and feature dissimilarity for unpaired patches. Alternatively, sum of $\mathcal{L}_1$, $\mathcal{L}_{1,2}$, and $\mathcal{L}'_{1,2}$ can also be used as aggregated loss function. The self-supervised mechanism for network training is shown in Algorithm 1.

G. Change Detection

Once the network is trained, it can be used to detect change between X_1 and Z_2. Since the network is fully convolutional, it enables us to obtain a pixelwise feature vector of dimension K from X_1 and Z_2. Similar to [1], the pixelwise change

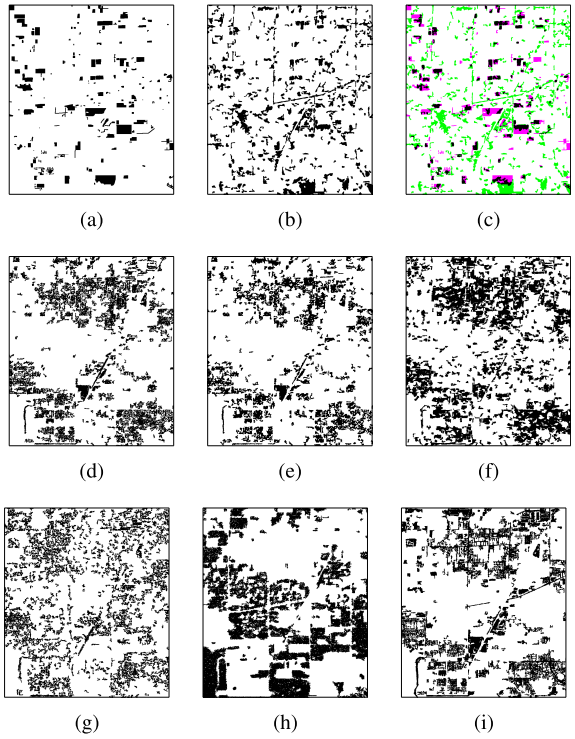


Fig. 4. CD results for Las Vegas. CD maps: (a) reference, (b) proposed, (c) false color composite (FCC) between reference and proposed (the correctly detected region are in black, false alarms are in green, and missed alarms are in pink), (d) CVA, (e) RCVA, (f) PCVA, (g) DCVA, (h) encoder–decoder, and (i) SCCN.

information is captured by taking the magnitude (ℓ_2 norm) of difference of the feature vectors computed from prechange and postchange pixels. Changed pixels (Ω_c) generate a higher difference magnitude in comparison to the unchanged ones ω_{nc}, and they can be distinguished by using any suitable threshold determination scheme [43].

IV. Experimental Validation

A. Datasets

We use four paired optical (prechange)–SAR (postchange) images to validate the proposed method. Optical images are acquired by the Sentinel-2 sensor and are taken from the Onera Satellite Change Detection (OSCD) dataset [44]. They show 10-m/pixel spatial resolution. The OSCD dataset is originally a single-sensor dataset consisting of only Sentinel-2 images. Recalling the importance of multisensor CD (see Section I), we extend this dataset by collecting the postchange SAR Sentinel-1 images for the nearest available date as the postchange image in the original OSCD dataset. Both Sentinel-2 and Sentinel-1 sensors are part of the European Space Agency's Copernicus program.

The four scenes are collected over Las Vegas in United States (824 × 716 pixels) (see Fig. 4), Chongqing in China (730 × 544 pixels) (see Fig. 5), Abu Dhabi (799 × 785 pixels) (see Fig. 6), and Montpellier in France (426 × 451 pixels) (see Fig. 7). Thus, this provides us an opportunity to validate the proposed method on geographically distributed complex urban scenes with large variation.

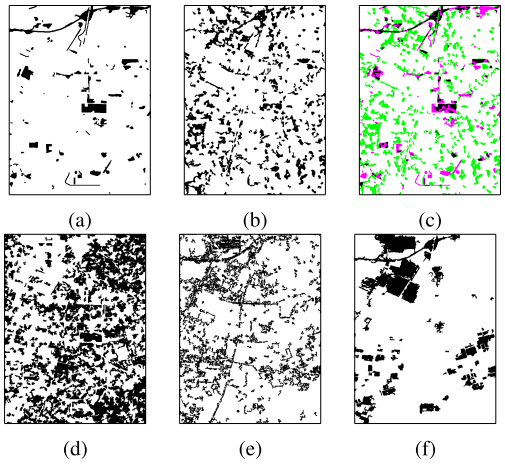


Fig. 5. CD results for the Chongqing. CD maps: (a) reference, (b) proposed, (c) FCC between reference and proposed (the correctly detected region are in black, false alarms are in green, and missed alarms are in pink), (d) PCVA, (e) DCVA, and (f) SCCN.

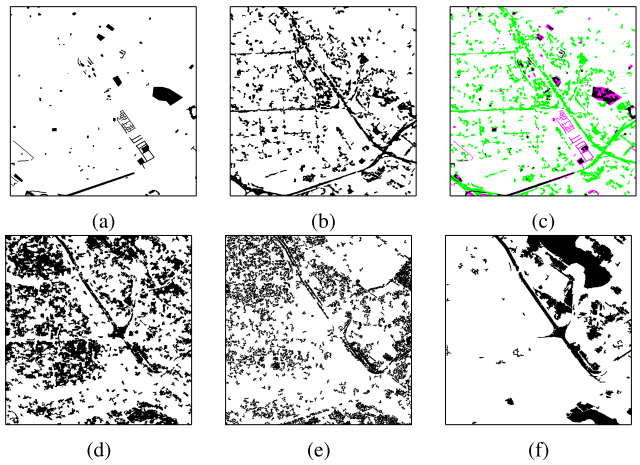


Fig. 6. CD results for the Abu Dhabi. CD maps: (a) reference, (b) proposed, (c) FCC between reference and proposed (the correctly detected region are in black, false alarms are in green, and missed alarms are in pink), (d) PCVA, (e) DCVA, and (f) SCCN.

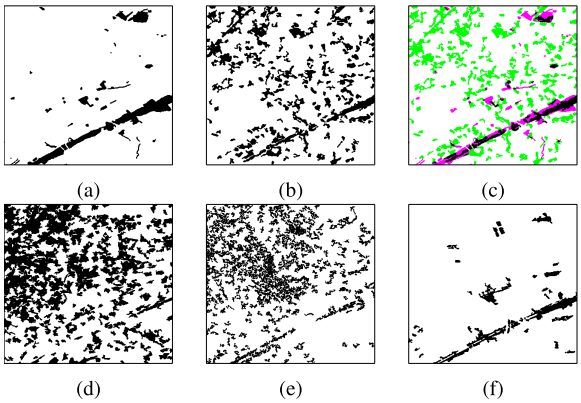


Fig. 7. Qualitative CD results for the Montpellier. CD maps: (a) reference, (b) proposed, (c) FCC between reference and proposed (the correctly detected region are in black, false alarms are in green, and missed alarms are in pink), (d) PCVA, (e) DCVA, and (f) SCCN.

B. Compared Methods

To verify the effectiveness of the proposed method, we compare it to related unsupervised CD methods.

TABLE I
STRUCTURE OF THE NETWORK FOR PROCESSING ONE OF THE TWO INPUTS

Layer	Kernel number	Kernel size	Stride
convolution	64	(3,3)	1
convolution	64	(3,3)	1
convolution	64	(3,3)	1
convolution	64	(3,3)	1
convolution	K	(1,1)	1

1) CVA [17], [45], a classical difference-based unsupervised model for CD.
2) RCVA [19] that modifies CVA by taking into account pixel neighborhood effects.
3) PCVA [18] that incoporates notion of the object (superpixels) in CVA.
4) DCVA [1] that detects change by comparing bitemporal deep features extracted using a pretrained network. We used the second convolution layer of pretrained VGGNet [46] for feature extraction.
5) Image-to-image transfer model based on an encoder–decoder network architecture that projects prechange optical images into postchange SAR image [47]. The CD map can be obtained by the difference between the simulated prechange SAR image (obtained as the projection of prechange optical image) and the original postchange SAR image.
6) DAE-based joint feature extraction [27].
7) SCCN [26] that first identifies some unchanged pixels and uses them to learn a coupled network.

While methods 1–3 are not deep learning-based, the following ones are deep learning-based. Methods 1–4 do not have any explicit adaptation for multisensor input, while methods 5–7 have.

C. Experimental Settings

The proposed method and compared methods are fed with preprocessed images and postprocessed similarly. For the proposed method, we use $\mathcal{I} = 5$ ($\mathcal{I}_1 = 1$, $\mathcal{I}_2 = 4$), $\mathcal{J} = 50$ $K = 4$, $L_1 = 4$, and $L_2 = 1$. We show the architecture of the network in Table I. A relatively simple architecture is used considering that the number of patches available to us is very few compared to the images in typical computer vision datasets. Moreover, our target image has a coarse resolution (10 m/pixel) compared to natural images in computer vision. Spatial complexity in such coarse images can be handled by simpler architecture compared to those in computer vision. 64×64 patches are used to train the model, and patches are extracted from the bitemporal scene with a stride of 32. The actual number of training patches for a scene depends on the size of the particular scene. For example, for the Las Vegas scene (824×716 pixels), the number of patches extracted is 504. For optimization, the SGD method is used with a learning rate set to 0.001.

We show the result in terms of sensitivity (accuracy in percentage computed over reference changed pixels) and specificity (computed over reference unchanged pixels). In more detail, given true positive (TP), true negative (TN), false positive (FP), and false negative (FN), sensitivity is TP/(TP + FN), and specificity is TN/(TN + FP).

TABLE II
COMPARISON OF DIFFERENT METHODS ON LAS VEGAS

Method	Sensitivity	Specificity
Proposed	50.28	88.06
CVA	9.64	77.13
RCVA	8.65	78.97
PCVA	23.60	67.92
DCVA	20.67	75.40
Encoder-decoder	46.30	68.58
DAE	39.07	83.72
SCCN	24.58	75.30

TABLE III
VARIATION OF RESULT FOR LAS VEGAS AS $\mathcal{I}$ IS VARIED

$\mathcal{I}$	Sensitivity	Specificity
0 (Just initialized)	32.21	85.30
1	61.70	53.40
2	52.20	79.82
3	46.30	86.52
5	50.28	88.06
10	41.79	88.80

TABLE IV
VARIATION OF RESULT FOR LAS VEGAS AS K IS VARIED

K	Sensitivity	Specificity
2	41.47	87.48
4	50.28	88.06
8	42.75	84.70
16	60.01	77.86

D. Results

1) Las Vegas: The reference CD map (ground truth) for Las Vegas is shown in Fig. 4(a). Fig. 4(b) shows the result obtained by the proposed method. For better visualization, a false color composition between the reference map and the obtained result is shown in Fig. 4(c). The proposed method can detect most of the changed objects with fewer false alarms in comparison to the compared methods. In many cases, the proposed method partly detects the changed object, thus missing some objects only partially [shown in pink in Fig. 4(c)]. CVA [see Fig. 4(d)] performs poorly and incorrectly detects most urban areas as changed. The result obtained by RCVA [see Fig. 4(e)] is similar to CVA. While PCVA [see Fig. 4(f)], DCVA [see Fig. 4(g)], encoder–decoder [see Fig. 4(h)], DAE, and SCCN [see Fig. 4(i)] improve the result over CVA, the proposed method still outperforms them by large margin. Quantitative evaluation (see Table II) clearly shows the superiority of the proposed method over state-of-the-art unsupervised methods. This can be attributed to the superior capability of the proposed method to ingest multisensor multitemporal images.

Further studies are conducted by varying different parameters on the Las Vegas image pair.

Training epochs $\mathcal{I}$ are varied with different values, as tabulated in Table III, while setting $K = 4$. We observe clear improvement in performance from $\mathcal{I} = 1$ to 2. Recalling

TABLE V
VARIATION OF RESULT FOR LAS VEGAS AS THRESHOLD DETERMINATION SCHEME IS VARIED

Thresholding	Sensitivity	Specificity
Otsu	50.28	88.06
ISODATA	50.48	87.95
Adaptive	50.19	83.65

TABLE VI
COMPARISON OF DIFFERENT METHODS ON CHONGQING

Method	Sensitivity	Specificity
Proposed	36.17	83.26
CVA	40.70	48.82
RCVA	41.96	44.28
PCVA	35.15	56.76
DCVA	32.67	79.18
Encoder-decoder	19.59	82.23
DAE	17.58	82.60
SCCN	30.67	85.73

TABLE VII
COMPARISON OF DIFFERENT METHODS ON ABU DHABI

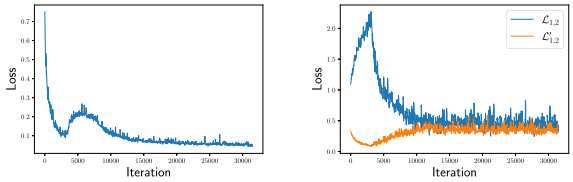
Method	Sensitivity	Specificity
Proposed	48.92	84.38
CVA	4.18	73.09
RCVA	5.75	73.47
PCVA	13.30	65.24
DCVA	20.35	76.74
Encoder-decoder	36.52	74.22
DAE	46.29	72.29
SCCN	9.79	83.88

from Section III-F that, for first $\mathcal{I}_1 = 1$ iterations, only deep clustering loss is used, this shows that bitemporal deep clustering itself is not sufficient to learn the correspondence between two images, and the other losses ($\mathcal{L}_{1,2}$ and $\mathcal{L}'_{1,2}$) are required. From $\mathcal{I} = 2$ onward, we observe an increment in performance initially followed by performance getting saturated/dropping. Despite variation in performance, the proposed method outperforms all compared methods for $\mathcal{I} = 3, 5, 10$.

The kernel number of the last layer (K) is varied from 2 to 16 in multiplicative steps of 2 while fixing the $\mathcal{I} = 5$. The variation in performance is shown in Table IV. While performance improves from $K = 2$ to $K = 4$, a gradual fall in performance is observed henceforth. The increasing value of K is equivalent to allowing the scene to be partitioned into more classes. Since the spatial area of the scene is fixed and not too large (only few hundred pixels by few hundred pixels), a large number of classes potentially leads the model to learn irrelevant classes, impacting CD performance.

Thresholding is done using Otsu's method [43], as it is popular in unsupervised CD methods [19], [48]. However, any other suitable method can be used, as shown in Table V. Results obtained by the ISODATA method [49], [50] and the adaptive method [1] are similar to Otsu's method [43].

Loss plot visualization in Fig. 8 shows the interplay between different components of loss. $\mathcal{L}_1$ consistently decreases [see Fig. 8(a)] except that it rises for a while after epoch 1 when



(a) (b)

Fig. 8. Evolution of the loss over training iterations for Las Vegas: (a) deep clustering loss $\mathcal{L}_1$ and (b) temporal consistency loss $\mathcal{L}_{1,2}$ and contrastive loss $\mathcal{L}'_{1,2}$.

TABLE VIII
COMPARISON OF DIFFERENT METHODS ON MONTPELLIER

Method	Sensitivity	Specificity
Proposed	43.05	81.88
CVA	9.09	74.60
RCVA	8.76	72.31
PCVA	25.74	58.59
DCVA	32.37	76.49
Encoder-decoder	46.10	75.52
DAE	24.46	73.97
SCCN	50.57	97.18

$\mathcal{L}_{1,2}$ and $\mathcal{L}'_{1,2}$ are introduced to the training process. $\mathcal{L}_{1,2}$ and $\mathcal{L}'_{1,2}$ balance each other, as shown in Fig. 8(b).

Projection layers f_{opt} and f_{sar} need to be modeled independently by not sharing weights between them to capture the different semantic properties of optical and SAR patches, as hypothesized in Section III-B. Here, we test this hypothesis by instead sharing the weights between f_{opt} and f_{sar}. For $\mathcal{I} = 5$ and $K = 4$, the proposed method fails to detect most of the changes. This shows that it is crucial to model the optical and SAR patches differently.

The computation time requirement is not high. We tested our code on a machine equipped with a Quadro T2000 GPU, which is a low-end GPU. For processing the Las Vegas dataset (training process over five epochs), it takes approx. 460 s. The Las Vegas scene is 824×716 pixels with the 10-m/pixel resolution, and thus, processing it is equivalent to processing an approximate area of $8 * 7 = 56$ km^2 in terms of geography.

The same sensor bitemporal input can be ingested by the proposed method, though designed for multisensor CD. For Las Vegas prechange optical–postchange optical input, the proposed method can obtain a sensitivity of 64.74% and specificity of 97.89%. However, we note that some characteristics of the proposed method (e.g., temporal consistency loss) are designed to reduce the representation gap of multisensor input, which is less relevant in single-sensor input. Thus, the proposed method may not be the most suitable choice for single-sensor scenarios as there are numerous existing CD techniques particularly designed for the same-sensor scenario [1].

2) Chongqing and Abu Dhabi: Reference CD map (ground truth) for Chongqing is shown in Fig. 5(a). Fig. 5(b) and (c) shows the result obtained by the proposed method and false color composition between the reference map and the obtained result, respectively. The proposed method outperforms all compared methods, as can be observed in

quantitative results in Table VI). Similar result is obtained for Abu Dhabi (see Fig. 6 and Table VII).

3) Montpellier: Reference CD map (ground truth) for Montpellier is shown in Fig. 7(a). The proposed method [see Fig. 7(b)] outperforms most of the state-of-the-art methods, including PCVA [see Fig. 7(d)] and DCVA [see Fig. 7(e)], as shown in Table VIII. However, SCCN [see Fig. 7(f)] outperforms the proposed method. The performance of the proposed method is relatively poor for Montpellier, which can be possibly explained by: 1) smaller size of Montpellier scene, which implies fewer data to learn proposed self-supervised network and 2) uniform (showing mostly urban areas) geospatial characteristics of Montpellier scene in comparison to Las Vegas and Chongqing that show complex distribution consisting of both urban and nonurban areas.

V. Conclusion

This article proposed a self-supervised learning-based method for CD in multisensor bitemporal images where one of the images is acquired by an optical sensor and the other one is captured by an SAR sensor. The proposed method effectively utilizes several concepts from self-supervised learning, e.g., deep clustering, Siamese network, multiple views, and contrastive learning, and operates under severe constraints, i.e., nothing except that the target scene is used, and no labeled data or additional unlabeled image is used. Despite the strong difference in the input modalities and operating under stringent constraints, it can identify a large fraction of the changed pixels. Comparisons with the existing methods working under unsupervised scenarios show that the proposed method brings significant improvement, especially when the target scene is large. Potential improvement of the proposed method may be achieved by prior learning of clusters on the unrelated domains/sensors and transferring them to target sensors on the fly [51]. In addition, our future work will focus on extending the method to other application domains, e.g., the comparison of biomedical images.

References

[1] 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.

[2] A. Appice, N. Di Mauro, F. Lomuscio, and D. Malerba, "Empowering change vector analysis with autoencoding in bi-temporal hyperspectral images," in *Proc. CEUR Workshop*, vol. 2466, 2019, pp. 1–10.

[3] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep Siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.

[4] F. Rahman, B. Vasu, J. V. Cor, J. Kerekes, and A. Savakis, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 958–962.

[5] A. Pomente, M. Picchiani, and F. Del Frate, "Sentinel-2 change detection based on deep features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 6859–6862.

[6] S. T. Seydi and M. Hasanlou, "A new land-cover match-based change detection for hyperspectral imagery," *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 517–533, Jan. 2017.

[7] M. Puhm, J. Deutscher, M. Hirschmugl, A. Wimmer, U. Schmitt, and M. Schardt, "A near real-time method for forest change detection based on a structural time series model and the Kalman filter," *Remote Sens.*, vol. 12, no. 19, p. 3135, Sep. 2020.

[8] 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, 2017, Art. no. 042609.

[9] M. Hirschmugl, J. Deutscher, C. Sobe, A. Bouvet, S. Mermoz, and M. Schardt, "Use of SAR and optical time series for tropical forest disturbance mapping," *Remote Sens.*, vol. 12, no. 4, p. 727, Feb. 2020.

[10] N. Zhou, X. Li, Z. Shen, T. Wu, and J. Luo, "Geo-parcel-based change detection using optical and SAR images in cloudy and rainy areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1326–1332, 2021.

[11] U. I. Ahmed, B. Rabus, and M. F. Beg, "SAR and optical image fusion for urban infrastructure detection and monitoring," *Proc. SPIE*, vol. 11535, Sep. 2020, Art. no. 115350M.

[12] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 4, pp. 6–23, Dec. 2016.

[13] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[14] G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein, *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences*. Hoboken, NJ, USA: Wiley, 2021.

[15] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "A critical analysis of self-supervision, or what we can learn from a single image," 2019, *arXiv:1904.13132*. [Online]. Available: http://arxiv.org/abs/1904.13132

[16] X. Chen and K. He, "Exploring simple Siamese representation learning," 2020, *arXiv:2011.10566*. [Online]. Available: http://arxiv.org/abs/2011.10566

[17] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with Landsat," in *Proc. LARS Symp.*, 1980, p. 385.

[18] 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.

[19] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, "Robust change vector analysis (RCVA) for multi-sensor very high resolution optical satellite data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 50, pp. 131–140, Aug. 2016.

[20] S. Saha, F. Bovolo, and L. Bruzzone, "Building change detection in VHR SAR images via unsupervised deep transcoding," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 1917–1929, Mar. 2021.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[22] H. M. Keshk and X.-C. Yin, "Change detection in SAR images based on deep learning," *Int. J. Aeronaut. Space Sci.*, vol. 21, no. 2, pp. 549–559, 2020.

[23] 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.

[24] 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.

[25] M. Volpi, G. Camps-Valls, and D. Tuia, "Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 107, pp. 50–63, Sep. 2015.

[26] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Mar. 2018.

[27] T. Zhan, M. Gong, X. Jiang, and S. Li, "Log-based transformation feature learning for change detection in heterogeneous images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1352–1356, Sep. 2018.

[28] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*. [Online]. Available: http://arxiv.org/abs/1803.07728

[29] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1422–1430.

[30] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2536–2544.

[31] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 132–149.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*. [Online]. Available: http://arxiv.org/abs/2002.05709

[33] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*. [Online]. Available: http://arxiv.org/abs/2003.04297

[34] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" 2020, *arXiv:2005.10243*. [Online]. Available: http://arxiv.org/abs/2005.10243

[35] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*. [Online]. Available: http://arxiv.org/abs/2006.07733

[36] S. Saha, L. Mou, C. Qiu, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Unsupervised deep joint segmentation of multitemporal high-resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8780–8792, Dec. 2020.

[37] S. Saha, F. Bovolo, and L. Bruzzone, "Change detection in image time-series using unsupervised LSTM," *IEEE Geosci. Remote Sens. Lett.*, early access, Dec. 24, 2020, doi: 10.1109/LGRS.2020.3043822.

[38] 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.

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

[40] 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.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[42] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*. [Online]. Available: http://arxiv.org/abs/1609.04747

[43] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[44] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2115–2118.

[45] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[47] 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.

[48] S. Saha, Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "Unsupervised deep transfer learning-based change detection for HR multispectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 856–860, May 2021.

[49] T. W. Ridler and S. Calvard, "Picture thresholding using an iterative selection method," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 8, pp. 630–632, Aug. 1978.

[50] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imag.*, vol. 13, no. 1, pp. 146–165, 2004.

[51] W. Menapace, S. Lathuilière, and E. Ricci, "Learning to cluster under domain shift," 2020, *arXiv:2008.04646*. [Online]. Available: http://arxiv.org/abs/2008.04646

Sudipan Saha (Member, IEEE) received the M.Tech. degree in electrical engineering from IIT Bombay, Mumbai, India, in 2014, and the Ph.D. degree in information and communication technologies from the University of Trento, Trento, Italy, and Fondazione Bruno Kessler, Trento, in 2020.

He is currently a Post-Doctoral Researcher with the Technical University of Munich (TUM), Munich, Germany. Previously, he worked as an Engineer with TSMC Ltd., Hsinchu, Taiwan, from 2015 to 2016. In 2019, he was a Guest Researcher with TUM.

His research interests include multitemporal remote sensing image analysis, domain adaptation, time-series analysis, image segmentation, deep learning, image processing, and pattern recognition.

Dr. Saha is also a reviewer for several international journals. He was a recipient of the Fondazione Bruno Kessler Best Student Award 2020. He has served as a Guest Editor for *Remote Sensing* (MDPI) Special Issue on "Advanced Artificial Intelligence for Remote Sensing: Methodology and Application."

Patrick Ebel received the B.Sc. degree in cognitive science from the University of Osnabrück, Osnabrück, Germany, in 2015, and the dual M.Sc. degree in cognitive neuroscience and artificial intelligence from Radboud University, Nijmegen, The Netherlands, in 2018. He is currently pursuing the Ph.D. degree with the Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany.

His research interests include deep learning and its applications in computer vision and to remote sensing data.

Xiao Xiang Zhu (Fellow, 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 was a Guest Scientist or a 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 at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Since 2019, she has been a Co-Coordinator of Munich Data Science Research School (www.mu-ds.de). Since 2019, she has also been the Head of the Helmholtz Artificial Intelligence–Research Field "Aeronautics, Space and Transport." Since May 2020, she has also been the Director of the International Future AI Laboratory "AI4EO–Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has been serving as the Co-Director for Munich Data Science Institute (MDSI), TUM. She is currently a Professor of data science in earth observation (former: signal processing in earth observation) with TUM and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. She is also a Visiting AI Professor with ESA's Phi-Lab. Her research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is also a member of Young Academy (Junge Akademie/Junges Kolleg) with the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves on the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ) and the Potsdam Institute for Climate Impact Research (PIK). She is also an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She also serves as the Area Editor responsible for Special Issue of *IEEE Signal Processing Magazine*.