# Unsupervised Single-Scene Semantic Segmentation for Earth Observation

Sudipan Saha, *Member, IEEE*, Muhammad Shahzad, Lichao Mou, Qian Song, *Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Earth observation data have huge potential to enrich our knowledge about our planet. An important step in many Earth observation tasks is semantic segmentation. Generally, a large number of pixelwise labeled images are required to train deep models for supervised semantic segmentation. On the contrary, strong intersensor and geographic variations impede the availability of annotated training data in Earth observation. In practice, most Earth observation tasks use only the target scene without assuming availability of any additional scene, labeled or unlabeled. Keeping in mind such constraints, we propose a semantic segmentation method that learns to segment from a single scene, without using any annotation. Earth observation scenes are generally larger than those encountered in typical computer vision datasets. Exploiting this, the proposed method samples smaller unlabeled patches from the scene. For each patch, an alternate view is generated by simple transformations, e.g., addition of noise. Both views are then processed through a two-stream network and weights are iteratively refined using deep clustering, spatial consistency, and contrastive learning in the pixel space. The proposed model automatically segregates the major classes present in the scene and produces the segmentation map. Extensive experiments on four Earth observation datasets collected by different sensors show the effectiveness of the proposed method. Implementation is available at https://gitlab.lrz.de/ai4eo/cd/-/tree/main/unsupContrastiveSemanticSeg.**

***Index Terms*— Deep learning, self-supervised learning, semantic segmentation, single-scene training.**

Manuscript received January 30, 2022; revised March 30, 2022; accepted April 22, 2022. Date of publication May 12, 2022; date of current version May 26, 2022. This work was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme under Grant ERC-2016-StG-714087, Acronym: *So2Sat*; in part by the Helmholtz Association through the Framework of Helmholtz AI under Grant ZT-I-PF-5-01—Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100; in part by the German Federal Ministry of Education and Research (BMBF) under the Framework of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001; and in part by the German Federal Ministry of Economics and Technology in the framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Corresponding author: Xiao Xiang Zhu.)*

Sudipan Saha is with the Department of Aerospace and Geodesy, Chair of Data Science in Earth Observation, Technical University of Munich, 85521 Munich, Germany (e-mail: sudipan.saha@tum.de).

Muhammad Shahzad is with the Department of Aerospace and Geodesy, Chair of Data Science in Earth Observation, Technical University of Munich, 85521 Munich, Germany, on leave from the School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan (e-mail: muhammad.shahzad@tum.de).

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

Qian Song is with the Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: qian.song@dlr.de).

Digital Object Identifier 10.1109/TGRS.2022.3174651

## I. Introduction

RAPID development of remote sensing technologies has drastically increased the quantity of Earth observation sensors acquiring images with different spatial, spectral, and temporal resolution [1], [2]. A large volume of unlabeled images are currently available for characterizing various objects on the Earth's surface. Automatic analysis of such images is useful to study various anthropogenic and natural factors, including urban monitoring [3], disaster management [4], [5], agricultural monitoring [6], and monitoring natural resources' exploitation [7].

An important step in understanding images is semantic segmentation that assigns each pixel in image/scene to a meaningful category or class. This is true for both computer vision and Earth observation images [8]. Research toward supervised image segmentation methods has received significant attention in the era of deep learning that has outperformed previous methods [9]–[11]. Superior performance of deep learning, especially convolutional neural networks (CNNs), for semantic segmentation can be attributed to their capability to learn spatial features from large volume of labeled data. Most computer vision problems can use crowdsourcing [12] to collect large volume of labeled data. However, collecting labeled data in Earth observation is significantly challenging due to several factors that require domain expertise, including variation among different Earth observation sensors and disparity among different applications. Moreover, active (e.g., synthetic aperture radar) and lower resolution optical images are visually unintelligible, thus making them difficult to be labeled by a volunteer in a crowdsourcing platform. Thus, the applicability of supervised segmentation has been limited on Earth observation images due to the lack of labeled data [13]. Moreover, many Earth observation applications assume presence of only the target scene and no additional scene [14], [15]. Analyzing using only the target scene can be especially useful for quick disaster mapping when there is little time to collect additional unlabeled images.

Recently, unsupervised and self-supervised learning have gained significant attention in machine learning. Such approaches have been devised for different problems, e.g., image clustering [16], video analysis [17], and change detection in Earth observation images [18]. While most

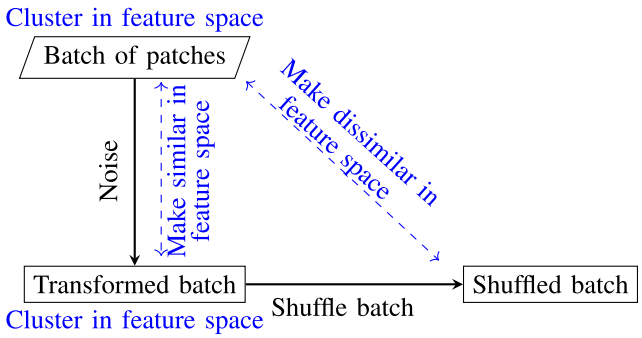



Fig. 1. Batch of patches is extracted from the training scene. Model is trained from this batch using deep clustering. Furthermore, this batch is simply transformed and shuffled, to form two other batches, first of which must be similar to the original batch in the feature space and the other must be dissimilar to the original batch in the feature space.

deep-learning-based semantic segmentation methods are supervised [19], [20], unsupervised semantic segmentation methods have been proposed in the literature exploiting deep clustering [21]. Deep-clustering-based approaches have also been extended for Earth observation bitemporal image analysis [3]. As such, self-supervised learning can be potentially used to learn from a single unlabeled scene.

Earth observation scenes generally capture a geographic area and are significantly large in comparison to images in a typical computer vision dataset. As an example, scenes in the International Society for Photogrammetry and Remote Sensing (ISPRS) semantic labeling dataset [22] are up to 6000 $\times$ 6000 pixels. Due to the repetitive nature of geographic objects, an Earth observation scene generally captures many instances of the same objects in a single scene. Based on this, we propose to sample smaller patches from a large scene. When randomly sampled, many such patches essentially represent the same object category (e.g., buildings). By taking a batch of patches, an augmented version can be conveniently obtained by data transformation, e.g., noise addition. This allows us to process the patches using a two-stream network similar to contrastive learning [23] and other multiaugmentation methods [24]. By jointly using concepts such as pixelwise deep clustering [25], similarity between multiple augmentations of the same input [24], and contrastive learning [23], we propose a self-supervised method to simultaneously train a network and assign pixelwise labels to an Earth observation scene. The conceptualization behind the proposed method is shown in Fig. 1. The key contributions of our work are as follows.

1) We propose a self-supervised segmentation method that does not require any annotated data and can be trained using single unlabeled Earth observation scene, without requiring any additional pool of unlabeled data.
2) We use the concept of pixelwise deep clustering [25] to automatically discern different classes from a single remote sensing scene. We further use multiple augmentations of same input [24] to ensure that similar inputs produce similar segmentation map. We use the concept of contrastive learning [23] to ensure that dissimilar inputs produce dissimilar output.
3) By performing a set of experiments using input of different sensors and resolutions, we show that the proposed method is able to automatically discern important Earth observation classes. This implies, irrespective of exact application, our method can be a precursor to further analysis in most such applications.

## II. Related Work

### A. Deep Segmentation for Earth Observation

Popular deep-learning-based segmentation architectures include fully convolutional networks (FCNs) [19], U-Net [26], SegNet [27], and dilated convolutional models including DeepLab [28]. For Earth observation images, several supervised segmentation algorithms have been proposed using these architectures [8], [29]–[35]. However, these methods necessitate a large amount of training data for supervised learning. To deal with the lack of training data, Hua *et al.* [13] proposed a semantic segmentation approach that uses spatially sparse annotations to train the model. In [3], an unsupervised deep clustering algorithm is introduced for the problem of multitemporal Earth observation segmentation. To effectively capture the domain knowledge, Li *et al.* [36] combine the deep learning module and knowledge-guided ontology reasoning.

Compared with optical images, SAR image segmentation is more challenging due to the sensitivity to noise [37]. The traditional SAR segmentation methods rely on superpixel merging [38], [39]. There are very few methods using deep learning for SAR image segmentation [40]. Wang *et al.* [40] noted that to train an effective deep model for SAR semantic segmentation, it is important to have high-quality ground-truth data that are not always available.

### B. Unsupervised and Self-Supervised Learning

Practicality of supervised methods is limited due to difficulty in acquiring labeled data. Unsupervised learning focuses on alleviating these limitations by learning semantic representations from unlabeled images without relying on predefined annotations. Clustering is an extensively studied unsupervised learning topic. Extending this, deep clustering [16] jointly optimizes the parameters of a deep network and the cluster assignments of the data in feature space. Deep clustering and its variants [41]–[44] divide a set of unlabeled training inputs into groups in terms of inherent latent semantics. Some self-supervised approaches use pretext tasks for learning semantic features [45], [46]. Popular pretext tasks include image rotation [45], jigsaw transformation [47], and rearranging of time-series [48]. Capitalizing on the availability of positive and negative pairs, contrastive methods aim to spread the representations of negative pairs apart while bringing closer the representations of the positive pairs [23], [49]. Bootstrap Your Own Latent (BYOL) [24] further eliminates the necessity of negative pairs using augmented instances of the input. Several works have shown that self-supervised learning can produce good representation even when available data are scarce [50]. Weakly supervised [48], [51], [52], unsupervised [53], and self-supervised learning [54], [55] have also been used in many remote sensing applications, e.g., cloud detection [52], change detection [53], and scene classification [54]. Tao *et al.* [54] used self-supervised learning for classification using limited label. Yue *et al.* [56] used self-supervised learning for hyperspectral scene classification.

TABLE I

KERNEL NUMBER AND RELEVANT DETAILS OF FIVE-CHANNEL NETWORK ($L = 5$), CONSIDERING A THREE-CHANNEL INPUT. ONLY ONE OF THE TWO STREAMS IS SHOWN. BATCH NORMALIZATION AND ACTIVATION FUNCTIONS ARE NOT SHOWN

| Layer type | Kernel number | Kernel size | Stride |
|---|---|---|---|
| Conv. | 64 | (3,3) | 1 |
| Conv. | 128 | (3,3) | 1 |
| Conv. | 128 | (3,3) | 1 |
| Conv. | 64 | (3,3) | 1 |
| Conv. | $K$ | (1,1) | 1 |

### C. Unsupervised Deep Segmentation

Aligned with the increased interest in unsupervised methods, efforts toward reducing supervision have gained traction in semantic segmentation [21], [57]. A simple yet effective approach toward this is using deep clustering in the pixel space [21], [58]. In [21], a lightweight architecture is used for single-image segmentation and output/label is obtained by arg-max classification of the final layer. Predicted pixel labels and network representation are adjusted in iterations. Pixel-level feature clustering using invariance and equivariance (PiCIE) [25] further exploits geometric consistency in addition to deep clustering for unsupervised segmentation.

Our work is closely related to the above-mentioned unsupervised methods. Like [21] and [25], it exploits pixelwise deep clustering. Our method relies on multiple augmentations of the same input, similar to BYOL [24]. Similar to [23], the method uses contrastive learning. The method focuses on single scene, thus further showing potential of deep self-supervised learning in data-constrained situation, similar to [50]. While works on self-supervised remote sensing classification [54] or self-supervised hyperspectral scene classification [56] still use some labeled samples, our method does not use any labeled sample.

## III. METHODOLOGY

To describe the proposed idea, let us denote the available unlabeled scene/image as $X$ and its transformed version as $\hat{X}$ having the same spatial dimensions of $R \times C$. Although any transformation $T_X$ could be useful, we use simple transformations such as addition of Gaussian noise. The transformed version can be taken as an alternative view of the same scene. This allows us to formulate the task of semantic segmentation at hand as a self-supervised problem which typically exploits the idea of reducing the gap among feature representations of multiple views of the same image in an iterative manner without using any labeled data. Both $X$ and $\hat{X}$ are then processed through a two-stream network and weights are iteratively refined using pseudo labels generated via deep clustering [16] and a contrastive learning strategy [23] in the pixel space to automatically segregate the major classes present in the scene. The proposed method produces segmentation map without using any explicit labels as detailed in Sections III-A–III-F.

### A. Proposed Network Architecture

To enable feature learning, a Siamese-like two-stream network architecture is proposed that takes as input the patches

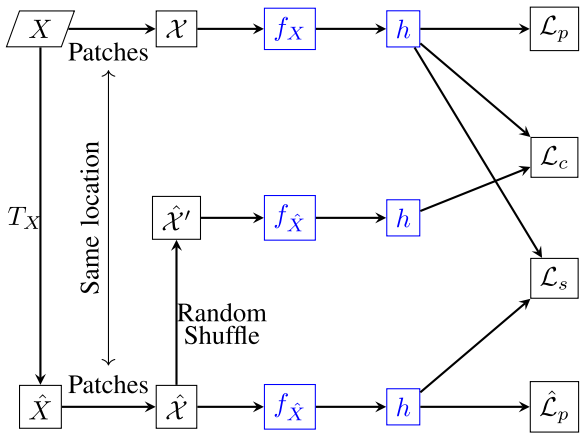



Fig. 2. Computation of losses for training proposed unsupervised segmentation framework. Network components are shown in blue outlined nodes to distinguish them from inputs, intermediate tensors, and losses.

of size $R' \times C'$ ($R' < R$ and $C' < C$) extracted from $X$ and $\hat{X}$. Each training batch is formed by drawing $\mathcal{B}$ patches from $X$ denoted as $\mathcal{X} = \{x^1, \ldots, x^{\mathcal{B}}\}$ and the spatially corresponding patches from $\hat{X}$, symbolized as $\hat{\mathcal{X}}$. Since the bispatial patches can be seen as multiple views of the same location, the semantic information can be inferred from them using a proposed Siamese-like architecture [59]. Both the branches have the projection modules $f_X$ and $f_{\hat{X}}$ to obtain learned feature representations for the original and transformed images, respectively. These learned feature representations are then fed to the subsequent prediction modules $h_X$ and $h_{\hat{X}}$ to obtain the respective activation volumes. It is important to note that the projection modules do not share weights (hence, two-stream), while the prediction module does share the weights (i.e., $h_X = h_{\hat{X}}$), therefore denoted using $h$ only. The projection and prediction modules consist of $L_1$ and $L_2$ (in our case $L_2 = 1$) convolutional layers, respectively, where the total layer $L$ is the sum of $L_1$ and $L_2$. Convolution layers are followed by activation function [rectified linear unit (ReLU)] and batch normalization layer. The input size is preserved in the output as pooling or stride is not used. The projection module uses convolution filters of size $3 \times 3$, whereas the prediction module is formed with $1 \times 1$ filter. The $K$ kernels in the final layer groups or clusters input data (pixels) into $K$ groups or classes.

The simplified network architecture for a five-channel network ($L = 5$) is shown in Table I. The reasoning behind using such an *ad hoc* lightweight architecture can be explained by the following.

1) Given that training mechanism is unsupervised and training patches are sampled from a single scene, we have limited number of patches. Thus, using a large network is ineffective in such case. This is further supported by previous works on single-scene segmentation [21] that also used such lightweight network.
2) Given that most Earth observation images have much coarser resolution compared with those in computer vision, small networks using only few convolution layers can still capture required spatial context.

### B. Pseudo Label Activations

The patches $x^b$ and $\hat{x}^b$ refer to patch extracted from the same location in $\mathcal{X}$ and $\hat{\mathcal{X}}$, respectively. The outcome of the network for $x^b$ and $\hat{x}^b$ can be represented as $y^b = h(f_X(x^b))$ and $\hat{y}^b = h(f_{\hat{X}}(\hat{x}^b))$ where $y^b$ and $\hat{y}^b$ tensors have the spatial dimensions of $R' \times C' \times K$. Here, each pixel in this tensor can be viewed as a $K$-dimensional vector of activations. If we denote any generic $i$th pixel in $y^b$ as $y_i^b$, then we can obtain the prediction of the semantic label by simply selecting the kernel in $y_i^b$ that has maximum value. Based on this simple intuition, we formulate the pseudo label assignment as the process of computing $c_i^b$ by finding the feature having the highest value in $K$-dimensional pixel activation vector $y_i^b$.

### C. Pseudo Label Loss Objective

The computed pseudo label $c_i^b$ is thus considered as the label of prediction $y_i^b$. This enables us to quantify per-pixel cross-entropy loss $\ell_i^b$ between $y_i^b$ and $c_i^b$. $\ell_i^b$ is aggregated (by computing mean) over pixels in $x^b$ and patches in the batch to obtain the loss $\mathcal{L}_p$. $\mathcal{L}_p$ is used to adjust the weights of $h$ and $f_X$. Similarly, $\hat{\mathcal{L}}_p$ is computed from $\hat{x}^b$ $(b = 1, \ldots, \mathcal{B})$ and used to adjust the weights of $h$ and $f_{\hat{X}}$. Ising $\mathcal{L}_p$ and $\hat{\mathcal{L}}_p$ to iteratively adjust the weights of the network, the proposed method simulates deep clustering in the pixel space.

### D. Spatial Consistency

The bispatial patches $x^b$ and $\hat{x}^b$ refer to the same location and hence to same objects, and therefore the features computed for such a bispatial pair patch should be similar. To ensure this, we compute per-pixel absolute error loss $\ell_i^b$ as absolute difference between $y_i^b$ and $\hat{y}_i^b$. The mean of $\ell_i^b$ over all pixels for all the patches in the batch gives the loss term $\mathcal{L}_s$ that ensures that the pixels in the bispatial patches $x^b$ and $\hat{x}^b$ tend to have the same label. We note that spatial consistency criterion is conceptually similar to bringing closer the multiple views of input as in some self-supervised learning methods [24]. However, differently from them, spatial consistency loss aims to reduce the representation gap at pixel level instead of image level.

A pitfall of the spatial consistency loss is that merely trying to reduce the representation gap of $x^b$ and $\hat{x}^b$ may generate trivial solution, simply producing the same output for all pixels.

### E. Representation Learning From Disparity

The spatial consistency loss encourages the features computed for a paired bispatial patch to be similar. To balance the overall training procedure, we also use a strategy similar to contrastive learning to ensure that the network should also learn different feature representations for dissimilar patches. To create dissimilar pair patches, we randomly shuffle the batch of patches $\hat{\mathcal{X}}$ to produce $\hat{\mathcal{X}}'$. This ensures that the paired patches in $\mathcal{X}$ and $\hat{\mathcal{X}}'$ are indeed dissimilar. These dissimilar bispatial patches are then used to enable the model to learn disparate features computed from $x^b$ and $\hat{x}'^b$. Specifically, $\ell_i'^b$ is computed as (negative) absolute error loss between $y_i^b$ and $\hat{y}_i'^b$. The proposed loss $\mathcal{L}_c$ is the mean of $\ell_i'^b$ over all patches in batch and all pixels in each patch.

**Algorithm 1** Self-Supervised Training for Semantic Segmentation in Earth Observation Data

```
Initialize the weights of the network
for i ← 1 to 𝓘 do
    Sample 𝒳 = {x^1, ..., x^ℬ} from X
    Obtain spatially corresponding ℬ patches from X̂, denoted as X̂ = {x̂^1, ..., x̂^ℬ}
    Obtain X̂' by randomly shuffling X̂
    for j ← 1 to 𝓙 do
        for b ∈ ℬ do
            y^b = h(f_X(x^b))

            ŷ^b = h(f_X̂(x̂^b))

            ŷ'^b = h(f_X̂(x̂'^b))

        end for
        Estimate pseudo label losses ℒ_p, ℒ̂_p
        Estimate spatial consistency loss - ℒ_s
        Estimate loss similar to contrastive learning - ℒ_c
        Use the losses to train the network
    end for
end for
```

### F. Progressive Network Training

The proposed mechanism for network training is shown in Fig. 2 and Algorithm 1. Initially, all the trainable weights $\mathbb{W}^1, \ldots, \mathbb{W}^L$ corresponding to all $L$ layers in the network are initialized using He initialization strategy proposed in [60]. Instead, a pretrained network could have been used to initialize weights. However, we note that Earth observation deals with a variety of sensors with different specifics, and suitable pretrained network is not always available. This motivates us to exclude importing weights from pretrained networks.

For each batch of data, training is performed for $\mathcal{J}$ iterations when the weights are iteratively optimized using stochastic gradient descent with momentum [61]. Sampling all possible patches from the training scene is equivalent to one epoch, and the training process is performed for a total $\mathcal{I}$ epochs. Since pseudo label losses ($\mathcal{L}_p$ and $\hat{\mathcal{L}}_p$) and other two losses ($\mathcal{L}_s$ and $\mathcal{L}_c$) have values in different range, the first epoch is optimized with the sum of $\mathcal{L}_p$ and $\hat{\mathcal{L}}_p$, while from the second epoch onward the sum of all four losses ($\mathcal{L}_p$, $\hat{\mathcal{L}}_p$, $\mathcal{L}_s$, and $\mathcal{L}_c$) is used, which yields a balanced training process taking into account coherent cluster formation, spatial feature consistency, and feature dissimilarity for unpaired patches.

## IV. Experiments

### A. Dataset

We use the following datasets for experimental validation.

1) Vaihingen dataset, an urban semantic segmentation benchmark [22], [62] acquired over Vaihingen, Germany, with 9 cm/pixel resolution. The images in

**Algorithm 2** Matching Unsupervised Segmented Map to the Reference Map

**Input**: Reference image and set of its constituent $N$ classes $\{\mathcal{S}_j\}_{j=1}^{N}$, ordered by number of pixels
**Input**: Obtained segmented image and set of its constituent $K$ classes $\{\mathcal{T}_i\}_{i=1}^{K}$
$\hat{\mathcal{T}}^1 \leftarrow \{\mathcal{T}_i\}_{i=1}^{K}$

**for** $j \leftarrow 1$ to $N$ **do**
    Find $\mathcal{T}_{\mathbb{D}^j}$ as the class in $\hat{\mathcal{T}}^j$ having highest intersection/overlap with $\mathcal{S}_j$
    Assign $\mathcal{T}_{\mathbb{D}^j}$ as match for $\mathcal{S}_j$
    $\hat{\mathcal{T}}^{j+1} = \hat{\mathcal{T}}^j \setminus \mathcal{T}_{\mathbb{D}^j}$

**end for**
Assign any remaining class in $\hat{\mathcal{T}}^{N+1}$ to background.

the dataset are composed of three bands—near infrared (NIR), red (R), and green (G), and each image covers approximately 1.38 km$^2$. The images show six land-cover classes: building, impervious surface, low vegetation, tree, car, and background. Following previous works [13], for test we use the image IDs 11, 15, 28, 30, and 34, i.e., total five test scenes. We train our unsupervised model on a single scene, image ID 1.

2) Zurich summer dataset [63] acquired using Quickbird sensor over Zurich, Switzerland. The images show a spatial resolution of 0.62 m/pixel. Following previous works [13] we use NIR, R, and G in our experiments. Eight different urban classes are present: roads, buildings, trees, grass, bare soil, water, railways and swimming pools. Image IDs 16–20 (i.e., total five test scenes) are used for test, while we train our unsupervised single-scene model on image ID 1.
3) A polarimetric synthetic aperture radar (PolSAR) [64] scene showing an area in Germany comprising four classes [65]. Being characterized by speckle noise and complex backscattering mechanism at the junction of different landcovers, PolSAR images are significantly different from optical images. Thus, the experiment on this dataset illustrates the application of the proposed method beyond typical optical images. Furthermore, due to less visual saliency, PolSAR scenes are challenging to label and there are not many labeled PolSAR datasets. This further proves the application of the proposed single scene unsupervised method on a case where label is actually scarce. This dataset [65] is acquired by ESAR L-band sensor. ESAR is an airbone SAR system of German Aerospace Center (DLR). It captures a semi-urban area in Germany (Oberperfaffenhofen, Bavaria province). The scene shows an area of 1300 × 1200 pixels. The reference information for the area is obtained using manually labeling based on the aerial images over the same area in Google Earth. The entire image is classified into four categories: built-up areas (in blue), wood land (in green), open areas (in yellow), and others (in dark blue). The classes are unbalanced, with much more open areas than others. Besides, there are some similarities between the built-up areas and wood land in terms of PolSAR image. Thus, segmentation of this scene is a challenging task for unsupervised methods.
4) Fire disturbance is recognized as an essential climate variables (ECVs) and burned area is its primary descriptive variable [66]. Here, we show the segmentation result produced by the proposed method on a burned area in an Alpine area in north Italy [67]. The fire event took place on February 27, 2019. We applied our segmentation method on postfire image acquired on March 3, 2019, using Sentinel-2 sensor (10 m/pixel spatial resolution and 13 spectral bands), part of Copernicus program of European Space Agency. The goal of this study is to investigate whether the proposed method can identify burned area as a separate cluster from the postevent image.

The proposed unsupervised training can be performed either on a different scene from the test scenes (as in the first two cases above) or on the same scene as the test scene (as in the third and fourth cases above).

### *B. Compared Methods*

Our work is one of the first attempts toward obtaining multiclass segmentation in unsupervised way by training on single-scene Earth observation image. Thus, we exclude entirely supervised methods from compared methods and choose following unsupervised/weakly supervised methods for comparison.

1) FEature and Spatial relaTional regulArization (FESTA) [13] is a weakly supervised method proposed in the context of semantic segmentation of high-resolution Earth observation images. The same training scene is used for training FESTA as our method; however, our method assumes no annotated point, while FESTA assumes the presence of some annotated points. We design two variants of FESTA, "FESTA 5 points" by considering five labeled point in the training scene and similarly "FESTA 10 points."
2) An unsupervised deep-clustering-based approach by adopting [16] in pixel space. The same training scene is used as the proposed method, and this method assumes no annotated data as in the proposed approach.
3) Combining deep clustering with image reconstruction as an additional pretext task. This model uses two outputs, one output is optimized for clustering and the other is optimized to reconstruct the input image [68].
4) Online deep clustering (ODC), derived from [41].
5) An unsupervised method by simply extracting pixel-wise features from the second convolutional layer of VGG16 [69] and applying $k$-means clustering on the extracted features. This particular layer is chosen since beyond this layer, the spatial size reduces, and thus pixelwise feature extraction is not possible.

Since FESTA assumes the presence of labeled pixels in the training scene and we use the same scene for training/testing in

TABLE II

PERFORMANCE VARIATION IN THE PROPOSED METHOD ON THE VAIHINGEN DATASET WITH RESPECT TO EPOCH

| Epoch ($\mathcal{I}$) | Mean F1 | Mean IOU | Acc. |
|---|---|---|---|
| 1 | 0.43 | 0.30 | 46.33 |
| 2 | 0.45 | 0.32 | 48.54 |
| 3 | 0.46 | 0.31 | 48.36 |
| 4 | 0.43 | 0.30 | 47.28 |

TABLE III

PERFORMANCE VARIATION IN THE PROPOSED METHOD ON THE VAIHINGEN DATASET WITH RESPECT TO $K$

| $K$ | Mean F1 | Mean IOU | Acc. |
|---|---|---|---|
| 6 | 0.40 | 0.28 | 46.85 |
| 8 | 0.45 | 0.32 | 48.54 |
| 12 | 0.40 | 0.29 | 42.96 |

case of the PolSAR scene, we exclude comparison to FESTA for that scene.

The burned area scene is evaluated for change detection and hence compared with the relevant method in [67].

### C. Settings

The training process of the proposed method is performed using $\mathcal{I} = 2$, $\mathcal{J} = 50$. The number of kernels in the final layer ($K$) is set as slightly larger than the number of target classes in dataset, e.g., $K = 8$ for the Vaihingen dataset and $K = 12$ for the Zurich dataset. $R' = C' = 224$ is used to sample patches from the training scene. A learning rate of 0.001 is used for training.

In an unsupervised clustering setting, it is not possible to automatically discern the name of classes. Hence, each class in obtained segmentation is assigned to the class with most overlap in the reference map. This procedure is further shown in Algorithm 2.

The results are shown as F1 score and intersection over union (IoU). The indices are computed for each target class and the mean is computed over all the classes. We also show accuracy; however, note that accuracy may be misleading as constituent classes are imbalanced and merely learning a single class can lead to seemingly good accuracy.

For the proposed method, the segmentation results are shown as an average of ten runs.

### D. Result on Vaihingen Dataset

*1) Result Variation With Respect to Parameters:* Table II shows the performance variation in the proposed method as the number of epochs $\mathcal{I}$ is varied by fixing the other parameters. We observe that the performance improvement beyond $\mathcal{I} = 2$ is not significant. Hence, we used $\mathcal{I} = 2$ in our subsequent experiments. To further understand this, we visualize evolution of losses in Fig. 3. $\mathcal{L}_s + \mathcal{L}_c$ keeps decreasing slightly beyond $\mathcal{I} = 2$; however, it shows an oscillatory behavior beyond that, which provides further indication toward why optimum result is already reached by $\mathcal{I} = 2$.

Table III shows the performance variation in the proposed method as the number of kernels in the final layer ($K$) is varied. We recall that the value of $K$ implies the number of classes that we want to cluster the data. The best performance is obtained for $K = 8$ which is slightly larger than the actual number of classes in the Vaihingen dataset (six classes).

TABLE IV

PERFORMANCE VARIATION IN THE PROPOSED METHOD ON THE VAIHINGEN DATASET WITH RESPECT TO THE NUMBER OF LAYERS IN THE MODEL

| Layers ($L$) | Mean F1 | Mean IOU | Acc. |
|---|---|---|---|
| 4 | 0.41 | 0.28 | 44.87 |
| 5 | 0.45 | 0.32 | 48.54 |
| 6 | 0.40 | 0.28 | 45.07 |

TABLE V

PERFORMANCE VARIATION IN THE PROPOSED METHOD ON THE VAIHINGEN DATASET BY VARYING COMPONENTS OF LOSS FUNCTIONS

| Loss function) | Mean F1 | Mean IOU | Acc. |
|---|---|---|---|
| $\{\mathcal{L}_p\}$ | 0.40 | 0.28 | 45.09 |
| $\{\mathcal{L}_p, \hat{\mathcal{L}}_p\}$ | 0.44 | 0.31 | 47.62 |
| $\{\mathcal{L}_s, \mathcal{L}_c\}$ | 0.30 | 0.20 | 38.65 |
| $\{\mathcal{L}_p, \hat{\mathcal{L}}_p, \mathcal{L}_s\}$ | 0.44 | 0.31 | 47.28 |
| $\{\mathcal{L}_p, \hat{\mathcal{L}}_p, \mathcal{L}_c\}$ | 0.43 | 0.30 | 46.78 |
| $\{\mathcal{L}_p, \hat{\mathcal{L}}_p, \mathcal{L}_s, \mathcal{L}_c\}$ | 0.45 | 0.32 | 48.54 |

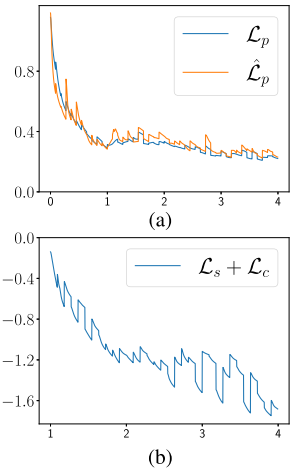



Fig. 3. Visualization of loss functions on Vaihingen scene. (a) $\mathcal{L}_p$ and $\hat{\mathcal{L}}_p$. (b) $\mathcal{L}_s + \mathcal{L}_c$. $x$-axis represents epochs and $y$-axis represents loss values.

Table IV shows the performance variation in the proposed method as the number of layers ($L$) is varied. The result confirms that only few layers are sufficient for the proposed method, and further increasing the number of layers may not improve the performance.

*2) Ablation Study of Loss Function:* Table V tabulates the performance of the proposed method with different combinations of losses: only $\{\mathcal{L}_p\}$ (i.e., only pseudo label loss on target scene); $\{\mathcal{L}_p, \hat{\mathcal{L}}_p\}$ (i.e., both pseudo-label losses); $\{\mathcal{L}_s, \mathcal{L}_c\}$ (i.e., excluding pseudo label losses): $\{\mathcal{L}_p, \hat{\mathcal{L}}_p, \mathcal{L}_s\}$ (i.e., excluding contrastive loss $\mathcal{L}_c$); $\{\mathcal{L}_p, \hat{\mathcal{L}}_p, \mathcal{L}_c\}$ (i.e., excluding spatial consistency loss $\mathcal{L}_s$); and all $\{\mathcal{L}_p, \hat{\mathcal{L}}_p, \mathcal{L}_s, \mathcal{L}_c\}$.

Pseudo label loss plays more crucial role than other two losses which is evident from superior performance of $\{\mathcal{L}_p\}$ in comparison to $\{\mathcal{L}_s, \mathcal{L}_c\}$. $\{\mathcal{L}_p, \hat{\mathcal{L}}_p\}$ significantly outperforms

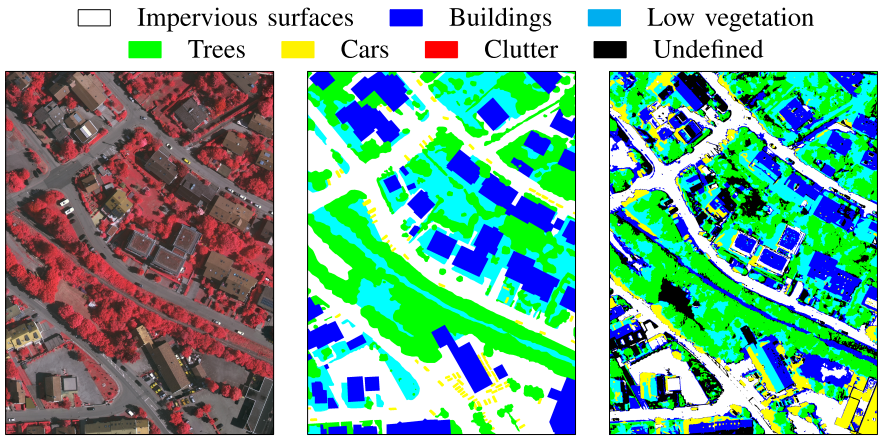



Fig. 4. Visualization of segmentation on the Vaihingen dataset. (a) Input image 11 (false color composition). (b) Corresponding reference segmentation. (c) Segmentation produced by the proposed unsupervised method.

TABLE VI
QUANTITATIVE COMPARISON FOR THE VAIHINGEN DATASET

| Method | Mean F1 | Mean IOU | Acc. |
|---|---|---|---|
| Proposed | 0.45 | 0.32 | 48.54 |
| FESTA 5 points | 0.26 | 0.16 | 33.64 |
| FESTA 10 points | 0.32 | 0.23 | 49.44 |
| Deep clustering | 0.25 | 0.14 | 25.03 |
| Clustering with reconstruction | 0.25 | 0.15 | 29.54 |
| ODC | 0.22 | 0.13 | 29.24 |
| VGG16+kMeans | 0.23 | 0.15 | 34.11 |

TABLE VII
QUANTITATIVE COMPARISON FOR THE ZURICH DATASET

| Method | Mean F1 | Mean IOU | Acc. |
|---|---|---|---|
| Proposed | 0.39 | 0.33 | 68.70 |
| FESTA 5 points | 0.15 | 0.09 | 29.29 |
| FESTA 10 points | 0.13 | 0.07 | 30.83 |
| Deep clustering | 0.33 | 0.26 | 67.23 |
| Clustering with reconstruction | 0.29 | 0.22 | 61.89 |
| ODC | 0.36 | 0.29 | 67.70 |
| VGG16+kMeans | 0.34 | 0.30 | 75.45 |

$\{\mathcal{L}_p\}$, showing that introduction of $\hat{\mathcal{L}}_p$ benefits segmentation. Both $\{\mathcal{L}_p, \hat{\mathcal{L}}_p, \mathcal{L}_s\}$ and $\{\mathcal{L}_p, \hat{\mathcal{L}}_p, \mathcal{L}_c\}$ are outperformed by $\{\mathcal{L}_p, \hat{\mathcal{L}}_p\}$. However, a combination of all losses $\{\mathcal{L}_p, \hat{\mathcal{L}}_p, \mathcal{L}_s, \mathcal{L}_c\}$ outperforms all other combinations. This shows that losses $\mathcal{L}_s$ and $\mathcal{L}_c$ cannot work on their own; however, they work when both of them are used together.

*3) Comparison to Existing Methods:* The quantitative result is shown in Table VI. The proposed method outperforms FESTA 5 points, deep clustering, deep clustering with image reconstruction, ODC, and VGG16 + kMeans with respect to all three indices and outperforms FESTA 10 points with respect to two out of three indices. We recall that FESTA is a semisupervised method that uses few annotated points. The proposed method still outperforms it, which shows the efficacy of the proposed method. Segmentation map corresponding to image ID 11 is visualized in Fig. 4. The three columns show input image, reference segmentation, and obtained segmentation, in that order. We observe that dominant classes like buildings (blue) and impervious surfaces (white) are clearly detected by the proposed method. However, it identifies spectrally similar low vegetation and trees in the same cluster. The classwise F1 score is 0.66, 0.48, 0.40, 0.64, and 0.08, for impervious surface, buildings, low vegetation, trees, and cars, respectively. This shows that the proposed unsupervised method is capable of identifying the major classes while its scope is limited for visually inconspicuous classes like cars.

TABLE VIII
QUANTITATIVE COMPARISON FOR POLSAR SCENE

| Method | Mean F1 | Mean IOU | Acc. |
|---|---|---|---|
| Proposed | 0.52 | 0.39 | 61.23 |
| Deep clustering | 0.37 | 0.26 | 52.14 |
| Clustering with reconstruction | 0.37 | 0.28 | 58.16 |
| ODC | 0.31 | 0.22 | 52.89 |
| VGG16+kMeans | 0.24 | 0.17 | 50.25 |

### *E. Result on Zurich Dataset*

The quantitative result of the proposed method versus the compared methods is shown in Table VII. The proposed method outperforms all the compared methods in terms of mean F1, and mean IOU, showing again its superiority even against semisupervised FESTA. Segmentation map for image ID 17 is visualized in Fig. 5. Similar to the observation for Vaihingen, we observe that the dominant classes are clearly detected by the proposed method. However, the performance deteriorates for the nondominant classes.

### *F. Result on PolSAR Scene*

Pauli-color-coded input, reference segmentation map, and the segmentation produced by the proposed method are visualized in Fig. 6. Despite different nature of PolSAR data, the proposed method is able to identify the major classes from the target scene. The quantitative result is tabulated in Table VIII

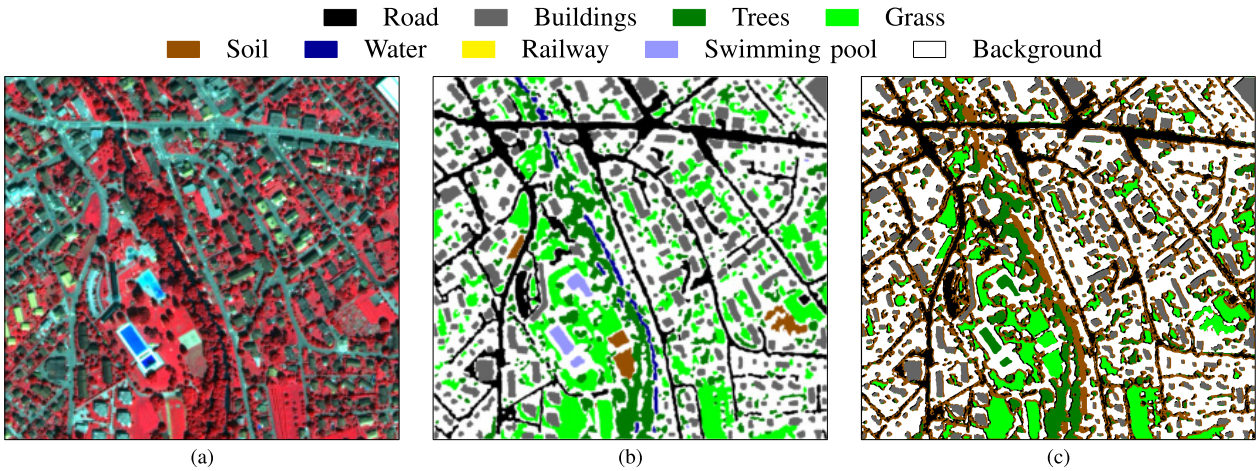



Fig. 5. Visualization of segmentation on the Zurich dataset. (a) Input images 17 (false color composition). (b) Corresponding reference segmentation. (c) Segmentation produced by the proposed unsupervised method.

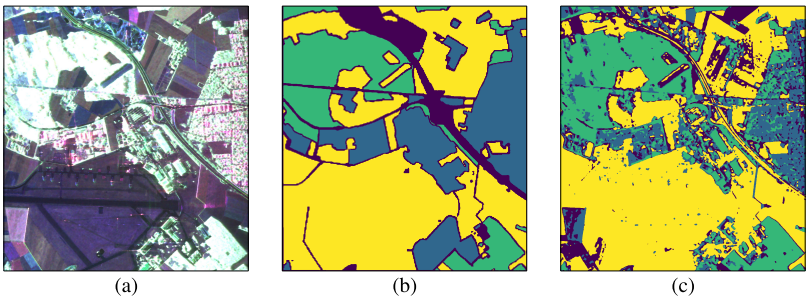



Fig. 6. Visualization of segmentation on PolSAR scene. (a) Pauli-color-coded scene. (b) Reference image. (c) Segmentation produced by the proposed method.

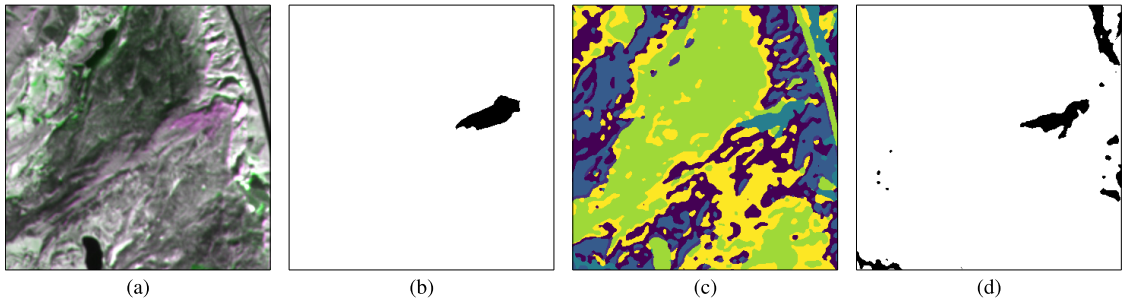



Fig. 7. Visualization of segmentation on Alpine burned area scene. (a) False color composition between prechange and postchange SWIR bands. (b) Reference burned area. (c) Obtained multiclass segmentation. (d) Obtained segmentation projected to two classes (black corresponds to burned pixels).

which shows the superiority of the proposed method against other unsupervised methods.

### *G. Result on Sentinel-2 Burned Area Scene*

Our segmentation method is applied on the postchange image (acquired on March 3, 2019). The target area is significantly complex, showing mountain, some snow, forest, in addition to the burned area. Showing the cluster that has the best match to the burned area as positive class and rest as negative class, we obtain a binary segmentation map, as visualized in Fig. 7. It is evident that the proposed method can segregate the target burned area as one class with little false alarm. The method obtains an accuracy of 97.19%. The result obtained by the proposed method is superior to or comparable to the change detection methods compared in [67] (worst accuracy: 76.16%, best accuracy 99.0%), though the change detection methods use both pre/postchange images, while the proposed method uses only the postchange image.

### *H. Comments on Computation Time*

The proposed unsupervised training on a single scene can be achieved in reasonable time, e.g., it takes approximately 195 s for training on Vaihingen image ID 1 using a machine equipped with GeForce RTX 3090. Using the same hardware and for the same scene, deep clustering [16] takes 280 s and ODC takes [41] 295 s. VGG + kMeans does not involve a

training phase. FESTA takes considerably more time than the proposed method (approximately 10 min).

### *I. Summary of Observations*

The proposed method is an inexpensive method, both in terms of annotation (not needed) and computation time. In addition to clustering in pixel space, the proposed method effectively exploits spatial consistency and contrastive loss, which is evident from the fact that the proposed method outperforms deep clustering. While the proposed method's effectiveness to automatically segment small classes is limited, it can effectively segregate the major classes, seen in all the datasets. However, this suits most Earth observation applications where the task is to quickly find one or two classes of interest, e.g., building during Earthquake disaster management and burned area during postfire operations.

## V. Conclusion

We proposed an unsupervised single-scene segmentation method that combines different recently popular topics from unsupervised and self-supervised learning, e.g., deep clustering in pixel space, different view/augmentation, and contrastive learning. The experimental results on four different Earth observation datasets show that the method can effectively learn dominant classes, e.g., buildings in the Vaihingen dataset. On the other hand, the effectiveness of the method is limited for classes that are inconspicuous. However, given the strong constraints under which the method works (only a single unlabeled scene for training), learning such classes is certainly challenging. A potential direction of extension of this work is training weakly supervised model given few labeled pixels from only such inconspicuous classes. The proposed method complements the supervised models by providing a quick unsupervised way of creating reasonable segmentation map. In future, we will experiment on the images acquired by other popular sensors in Earth observation, e.g., light detection and ranging (LiDAR).

## References


[1] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12325–12334.

[2] J. Lee, S. Seo, and M. Kim, "SIPSA-Net: Shift-invariant pan sharpening with moving object alignment for satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10166–10174.

[3] 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.

[4] S. Yang, M. Lupascu, and K. S. Meel, "Predicting forest fire using remote sensing data and machine learning," 2021, *arXiv:2101.01975*.

[5] 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. 2020.

[6] T.-X. Zhang, J.-Y. Su, C.-J. Liu, and W.-H. Chen, "Potential bands of sentinel-2A satellite for classification problems in precision agriculture," *Int. J. Autom. Comput.*, vol. 16, no. 1, pp. 16–26, Feb. 2019.

[7] J. Gallwey, C. Robiati, J. Coggan, D. Vogt, and M. Eyre, "A Sentinel-2 based multispectral convolutional neural network for detecting artisanal small-scale mining in ghana: Applying deep learning to shallow mining," *Remote Sens. Environ.*, vol. 248, Oct. 2020, Art. no. 111970.

[8] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[9] 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.

[10] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.

[11] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[12] A. Kovashka, O. Russakovsky, L. Fei-Fei, and K. Grauman, "Crowdsourcing in computer vision," 2016, *arXiv:1611.02145*.

[13] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," 2021, *arXiv:2101.03492*.

[14] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2017.

[15] B. Huang, Z. Li, C. Yang, F. Sun, and Y. Song, "Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1806–1813.

[16] 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.

[17] J. Wang, J. Jiao, and Y.-H. Liu, "Self-supervised video representation learning by pace prediction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 504–521. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-58520-4_30

[18] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2021.

[19] 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.

[20] M. Zhen, J. Wang, L. Zhou, T. Fang, and L. Quan, "Learning fully dense neural networks for image semantic segmentation," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 9283–9290, Jul. 2019.

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

[22] F. Rottensteiner *et al.*, "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Ann. Photogramm., Remote Sensing Spatial Inf. Sci.*, vol. 1, no. 3, pp. 293–298, 2012.

[23] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[24] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.

[25] J. H. Cho, U. Mall, K. Bala, and B. Hariharan, "PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 16794–16804.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image 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

[27] 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.

[28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[29] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*.

[30] 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.

[31] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7092–7103, Dec. 2017.

[32] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[33] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5367–5376, Aug. 2020.
[34] Y. Su, J. Cheng, W. Wang, H. Bai, and H. Liu, "Semantic segmentation for high-resolution remote-sensing images via dynamic graph context reasoning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[35] B. Ma and C.-Y. Chang, "Semantic segmentation of high-resolution remote sensing images using multiscale skip connection network," *IEEE Sensors J.*, vol. 22, no. 4, pp. 3745–3755, Feb. 2022.
[36] Y. Li, S. Ouyang, and Y. Zhang, "Combining deep learning and ontology reasoning for remote sensing image semantic segmentation," *Knowl.-Based Syst.*, vol. 243, May 2022, Art. no. 108469.
[37] D. Xiang, T. Tang, C. Hu, Y. Li, and Y. Su, "A kernel clustering algorithm with fuzzy factor: Application to SAR image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 7, pp. 1290–1294, Jul. 2013.
[38] D. Xiang *et al.*, "Adaptive statistical superpixel merging with edge penalty for PolSAR image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2412–2429, Apr. 2020.
[39] D. Xiang *et al.*, "Fast pixel-superpixel region merging for SAR image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9319–9335, Nov. 2021.
[40] X. Wang, L. Cavigelli, M. Eggimann, M. Magno, and L. Benini, "HR-SAR-Net: A deep neural network for urban scene segmentation from high-resolution SAR data," in *Proc. IEEE Sensors Appl. Symp. (SAS)*, Mar. 2020, pp. 1–6.
[41] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6688–6697.
[42] M. Jabi, M. Pedersoli, A. Mitiche, and I. B. Ayed, "Deep clustering: On the link between discriminative models and K-means," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1887–1896, Jun. 2021.
[43] S. E. Chazan, S. Gannot, and J. Goldberger, "Deep clustering based on a mixture of autoencoders," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.
[44] X. Guo *et al.*, "Adaptive self-paced deep clustering with data augmentation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 9, pp. 1680–1693, Sep. 2020.
[45] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.
[46] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.
[47] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 69–84. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46466-4_5
[48] S. Saha, F. Bovolo, and L. Bruzzone, "Change detection in image time-series using unsupervised LSTM," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2020.
[49] Y. Tian *et al.*, "What makes for good views for contrastive learning?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6827–6839.
[50] 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*.
[51] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 182–196, Dec. 2018.
[52] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112045.
[53] H. Dong, W. Ma, L. Jiao, F. Liu, and L. Li, "A multiscale self-attention deep clustering for change detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021.
[54] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2020.
[55] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1182–1191.
[56] J. Yue, L. Fang, H. Rahmani, and P. Ghamisi, "Self-supervised learning with adaptive distillation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.
[57] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9865–9874.
[58] 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/10.1007/978-3-030-30645-8_46
[59] X. Chen and K. He, "Exploring simple Siamese representation learning," 2020, *arXiv:2011.10566*.
[60] 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.
[61] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.
[62] F. Rottensteiner, G. Sohn, M. Gerke, J. D. Wegner, U. Breitkopf, and J. Jung, "Results of the ISPRS benchmark on urban object detection and 3D building reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 93, pp. 256–271, Jul. 2014.
[63] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 1–9.
[64] Y. Yamaguchi, *Polarimetric SAR Imaging: Theory and Applications*. Boca Raton, FL, USA: CRC Press, 2020.
[65] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7177–7188, Dec. 2017.
[66] M. K. Vanderhoof, N. Fairaux, Y.-J.-G. Beal, and T. J. Hawbaker, "Validation of the USGS Landsat burned area essential climate variable (BAECV) across the conterminous United States," *Remote Sens. Environ.*, vol. 198, pp. 393–406, Sep. 2017.
[67] 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.
[68] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.
[69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

**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 was an Engineer with TSMC Limited, Hsinchu, Taiwan, from 2015 to 2016. In 2019, he was a Guest Researcher with the Technical University of Munich (TUM), Munich, Germany, where he has been a Post-Doctoral Researcher since 2020. 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 was a recipient of the Fondazione Bruno Kessler Best Student Award 2020. He is a reviewer for several international journals. He served as a Guest Editor at *Remote Sensing* (MDPI) special issue on "Advanced Artificial Intelligence for Remote Sensing: Methodology and Application" and *Frontiers In Remote Sensing* Research Topic on "Learning with Limited Label."

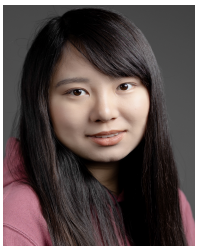
**Muhammad Shahzad** received the B.E. degree in electrical engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 2004, the M.Sc. degree in autonomous systems (robotics) from the Bonn Rhein Sieg University of Applied Sciences, Sankt Augustin, Germany, in 2011, and the Ph.D. degree in radar remote sensing and image analysis from the Department of Signal Processing in Earth Observation (SiPEO), Technische Universität München (TUM), Munich, Germany, in 2016. His Ph.D. topic was on automatic 3-D reconstruction of objects from point clouds retrieved from spaceborne synthetic aperture radar (SAR) image stacks.

He possesses rich experience in multimodal remote sensing data processing. Besides, he also attended twice two weeks of professional thermography training course at the Infrared Training Center (ITC), North Billerica, MA, USA, and Portland, OR, USA, in 2005 and 2007, respectively. He was a Visiting Scientist with the Institute for Computer Graphics and Vision, Technical University of Graz, Graz, Austria, in 2015 and 2016. From 2016 to 2021 (currently on leave), he was an Assistant Professor with the School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan. Furthermore, he also remained as a Co-Principal Investigator of the established Deep Learning Laboratory (DLL) under the umbrella of the National Center of Artificial Intelligence (NCAI), Islamabad. He is currently a Core Project Scientist and a Guest Professor with the AI4EO Future Laboratory, Chair of Data Science in Earth Observation, TUM. His research interests include application of deep learning for processing unstructured/structured 3-D point clouds, optical RGBD data, and very high-resolution radar images.

**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, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. Since 2019, he has been a Research Scientist with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany, and an AI Consultant with the Helmholtz Artificial Intelligence Cooperation Unit (HAICU). He is currently a Guest Professor with the Munich AI Future Laboratory "Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (AI4EO), TUM, and the Head of the Visual Learning and Reasoning Team, Department "EO Data Science," IMF, DLR.

Dr. Mou was a recipient of the First Prize in the 2016 IEEE GRSS Data Fusion Contest and the finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.



**Qian Song** (Member, IEEE) received the B.E. degree (Hons.) from the School of Information Science and Technology, East China Normal University, Shanghai, China, in 2015, and the Ph.D. degree (Hons.) from Fudan University, Shanghai, China, in 2020.

She is currently a Post-Doctoral Fellow with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. Her research interests include advanced deep learning technologies and their applications in synthetic aperture radar image interpretation.

Dr. Song received the International Union of Radio Science (URSI) Young Scientist Award in 2020.

**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 the Munich Data Science Research School, Munich (www.mu-ds.de). Since 2019, she has been the Head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has 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 also been serving as a Co-Director of the Munich Data Science Institute (MDSI), TUM. She is currently a Professor at the Chair of Data Science in Earth Observation (former: Signal Processing in Earth Observation), 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 at ESA's Phi-Lab, Frascati, Italy. Her main research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, e.g., global urbanization, United Nations Sustainable Development Goals (UN's SDG), and climate change.

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 serves in 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 an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and serves as the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.