Deep Unsupervised Embedding for Remotely Sensed Images Based on Spatially Augmented Momentum Contrast

Jian Kang, *Member, IEEE*, Ruben Fernandez-Beltran, *Senior Member, IEEE*, Puhong Duan, *Member, IEEE*, Sicong Liu, *Member, IEEE*, and Antonio J. Plaza, *Fellow, IEEE*

***Abstract*— Convolutional neural networks (CNNs) have achieved great success when characterizing remote sensing (RS) images. However, the lack of sufficient annotated data (together with the high complexity of the RS image domain) often makes supervised and transfer learning schemes limited from an operational perspective. Despite the fact that unsupervised methods can potentially relieve these limitations, they are frequently unable to effectively exploit relevant prior knowledge about the RS domain, which may eventually constrain their final performance. In order to address these challenges, this article presents a new unsupervised deep metric learning model, called spatially augmented momentum contrast (SauMoCo), which has been specially designed to characterize unlabeled RS scenes. Based on the first law of geography, the proposed approach defines spatial augmentation criteria to uncover semantic relationships among land cover tiles. Then, a queue of deep embeddings is constructed to enhance the semantic variety of RS tiles within the considered contrastive learning process, where an auxiliary CNN model serves as an updating mechanism. Our experimental comparison, including different state-of-the-art techniques and benchmark RS image archives, reveals that the proposed approach obtains remarkable performance gains when characterizing unlabeled scenes since it is able to substantially enhance the discrimination ability among complex land cover categories. The source codes of this article will be made available to the RS community for reproducible research.**

***Index Terms*— Deep learning (DL), metric learning, remote sensing (RS), scene characterization, self-supervised learning, unsupervised learning.**

Manuscript received April 21, 2020; revised June 18, 2020; accepted June 30, 2020. Date of publication July 14, 2020; date of current version February 25, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB 050500, in part by the Spanish Ministry of Economy under Grant RTI2018-098651-B-C54, in part by the FEDER-Junta de Extremadura under Grant GR18060, and in part by the European Union under the H2020 EOXPOSURE Project under Grant 734541. *(Corresponding author: Sicong Liu.)*

Jian Kang is with the Research Institute of Electronic Engineering Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: kangjian_1991@outlook.com).

Ruben Fernandez-Beltran is with the Institute of New Imaging Technologies, University Jaume I, 12071 Castellón de la Plana, Spain (e-mail: rufernan@uji.es).

Puhong Duan is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China.

Sicong Liu is with the College of Surveying and Geoinformatics, Tongji University, Shanghai 200092, China (e-mail: sicong.liu@tongji.edu.cn).

Antonio J. Plaza is with Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain (e-mail: aplaza@unex.es).

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

Digital Object Identifier 10.1109/TGRS.2020.3007029

I. Introduction

With the growing development of deep learning (DL) technologies, these kinds of methods have achieved tremendous success in many important remote sensing (RS) applications [1], [2], such as scene classification [3]–[7], object localization [8]–[12], and change detection [13]–[15], due to their prominent capabilities to uncover highly representative features from RS scenes [16]. In general, DL techniques aim at projecting the visual content of input images onto a particular label space, using a hierarchy of nonlinear layers to generate a high-level semantic abstraction that is very useful to characterize RS data. Most available DL-based image characterization methods in the RS field rely on a supervised learning scheme, in which a large amount of labeled scenes is required to properly train the models and prevent overfitting [17]. However, the task of obtaining relevant annotations for vast volumes of RS data can be very difficult and time consuming. This may severely constrain the applicability and potential of the supervised DL paradigm in operational RS environments, especially under the most challenging conditions [18].

In order to mitigate the need for labeled RS data, different strategies have been effectively explored in the literature [19]–[22]. One of the most popular schemes is based on the use of pretrained convolutional neural networks (CNNs) [23], where different predefined CNN architectures (e.g., AlexNet [24], VGGNet [25], GoogleNet [26], and ResNet [27]), trained on large-scale computer vision data sets (e.g., ImageNet [28]) are directly used as feature extraction methods for RS data. Despite its remarkable success [29]–[32], the existing limitations on the number of spectral bands and the data complexity make this transfer learning scheme unable to fully exploit the advantages of RS imagery [33]. An attractive option to relieve these limitations consists of using unsupervised methods to characterize unlabeled RS scenes. As a result, different methods have been successfully proposed within and outside the DL field [34]–[37]. However, the general unsupervised framework is often unable to introduce appropriate prior knowledge about the RS image domain, eventually constraining the resulting performance. Although a recently developed deep metric learning method—the Tile2Vec [38]—is certainly able to obtain promising results by using the geospatial information as prior knowledge, the unprecedented availability of massive RS archives, together with the constant

development of the acquisition technology, still make unsupervised DL-based image characterization a major challenge in RS. Note that the integration of the unsupervised mode into the deep metric learning approach [39] is highly limited by the contrasting land cover types that can be sampled in a single batch, which may eventually reduce the capacity of the model to distinguish a broader range of complex RS categories, and also motivates the development of novel techniques useful to deal with the large-scale variance complexity of the RS image domain [40].

With all these considerations in mind, this article proposes a new unsupervised deep metric learning approach, called spatially augmented momentum contrast (SauMoCo), which has been specially designed to characterize unlabeled RS scenes. Inspired by Tobler's first law of geography [41], the proposed approach provides a new perspective on unsupervised land cover characterization in which not only the semantic similarities among nearby scenes are exploited to learn the corresponding feature embeddings but also the inherent diversity within RS semantic concepts. To achieve this goal, we define spatial augmentation criteria to uncover enhanced semantic relationships among RS tiles for the embedding space. Then, we build a queue of deep embeddings, where the size of queue is forced to be larger than the batch size in order to further increase the semantic variety of contrasting land cover tiles during the training process. Moreover, we introduce an auxiliary CNN into our model to consistently update the deep embeddings of the RS tiles in the queue. With the objective of validating the proposed approach, we conduct a comprehensive experimental comparison, using two benchmark data sets and different state-of-the-art characterization techniques, which demonstrates the superior performance of the presented method in the task of categorizing RS scenes without using any land cover class information. In short, the main contributions of this article can be summarized as follows.

1) We propose a new unsupervised deep metric learning model (SauMoCo) to characterize unlabeled RS images. The presented approach pursues to exploit not only the semantic similarities among nearby geospatial locations but also the inherent diversity within land cover concepts, by using newly defined spatial augmentation criteria with a contrastive loss formulation and a momentum update-based optimization.
2) We investigate how the proposed SauMoCo model performs with large-scale training data, which gives us important insights about the working mechanism and practical advantages of the proposed method with respect to other unsupervised RS image characterization techniques available in the literature. The codes of this work will be released for reproducible research inside the research community.

The rest of this article is organized as follows. Section II reviews some related works on RS scene characterization while highlighting their main limitations. Section III details the proposed unsupervised deep metric learning model for RS images. Section IV presents the experimental part of the work. Finally, Section V concludes this article with some remarks and hints at plausible future research lines.

II. Related Work

Different strategies have been successfully adopted within the RS field to relieve the need for labeled data when characterizing aerial scenes. This section reviews some of the most relevant trends, including pretrained (Section II-A), unsupervised (Section II-B), and deep metric learning-based (Section II-C) methods. In addition, we also analyze their main limitations in the context of RS problems (Section II-D).

A. Pretrained Methods

One of the most popular schemes to characterize RS data is based on the use of pretrained CNNs. In more detail, these approaches make use of predefined CNN models, such as AlexNet [24], VGGNet [25], GoogleNet [26], or ResNet [27], which are pretrained on large-scale computer vision data sets, such as the ImageNet [28] collection. In this way, the amount of labeled RS scenes can be substantially reduced by transferring the knowledge from the standard image domain to the RS field. For instance, Hu *et al.* [42] defined two different schemes to take advantage of the VGGNet model pretrained on ImageNet. In the first scheme, the authors employ the last fully connected layers as image descriptors. In the second one, an additional encoding procedure is used to fuse the last convolutional feature maps. In both cases, a support vector machine (SVM) is adopted to finally classify the RS images. Precisely, Marmanis *et al.* [29] analyzed the effectiveness of classifying remotely sensed scenes using different CNN-based representations transferred from ImageNet. Similarly, Li *et al.* [31] presented a multilayer feature fusion framework that integrates several pretrained DL models for RS scene classification. Zheng *et al.* [43] built a holistic representation of RS images using a multiscale pooling over pretrained features. Kang *et al.* [44] also combined several pretrained CNN architectures to define a building-instance-level land-use classification framework. Othman *et al.* [30] proposed using a sparse autoencoder (AE) over pretrained features to generate the final representation of RS scenes.

Despite the remarkable performance achieved by these and other pretrained models, there are still some important limitations that substantially reduce the applicability of such transfer learning strategies within the RS field. On the one hand, standard image collections, such as ImageNet, are often made up of RGB imagery, which makes the existing pretrained networks unable to take advantage of the additional spectral bands provided by airborne and space-borne optical sensors [45]. Note that RS instruments are often designed to provide valuable information outside the visible spectrum, and these data are essential in many important applications, such as biophysical parameter analysis [46] and land cover material study [47]. On the other hand, standard images often contain natural object-centric photographs that hardly represent the complexity of RS scenes, comprising fully focused multiband shots of the Earth surface with plenty of complex spatiospectral details within the same acquisition frame [48]. Precisely, these important differences often make it necessary to consider broader strategies than pretrained DL models.

B. Unsupervised Methods

A more general option to relieve the need for annotated RS data is based on using unsupervised image characterization methods. In more detail, these techniques work for characterizing aerial scenes without using any class label information, which becomes particularly attractive in RS problems [18]. Consequently, different unsupervised models (both within and outside the DL field) have been proposed to learn informative representations from unlabeled RS scenes. For instance, Cheriyadat [49] presented a feature learning approach for aerial scenes, which adopts a sparse coding framework to generate unsupervised data representations based on a set of basis functions derived from low-level measurements. Following this idea, other authors proposed using different unsupervised decomposition frameworks instead. This is the case of the works presented in [37], [50], and [51], which make use of probabilistic topic models to represent the RS data as probability distributions of feature patterns. Zhang *et al.* [52] exploited a sparse AE to effectively learn saliency-guided unsupervised features for RS scenes. Analogously, Hu *et al.* [35] utilized a spectral clustering procedure to uncover the intrinsic structures among image patches. Romero *et al.* [34] introduced a greedy layerwise unsupervised pretraining method for learning sparse features from aerial images. In the case of [36], the authors define a shallow weighted deconvolution network for extracting features from RS scenes by minimizing the Euclidean distance between the original and the reconstructed images. Alternatively, some works in the literature also show the utility of convolutional generative adversarial networks to characterize standard and remotely sensed imagery [53], [54].

C. Unsupervised Deep Metric Learning

Notwithstanding the positive results of these and other important unsupervised methods, all these works mainly rely on generic clustering, decomposition, or encoding procedures that are often unable to introduce relevant prior knowledge about the RS domain without using supervised information. Among all the conducted research, one of the most promising trends to adequately characterize RS images is based on the so-called deep metric learning approach [39]. In particular, deep metric learning aims at learning a low-dimensional metric space based on CNN models, where the feature embeddings of similar images should be close, and those of dissimilar images should be separated. Despite its great potential in RS problems [55]–[58], how to effectively define such semantic relationships for unlabeled aerial scenes is still an open-ended issue. However, this situation has undergone an important change with the latest research on unsupervised deep metric learning. Specifically, Jean *et al.* [38] developed the Tile2Vec, which is an algorithm to learn vector representations of RS images by using their geospatial information as prior knowledge. In more detail, it is based on the observation that those RS image tiles that are spatially closer on the Earth surface are more likely to comprise similar semantics, and consequently representations, than tiles which are far apart and hence expected to comprise dissimilar semantics. In this way, Jean *et al.* proposed learning a deep metric space where the feature embeddings of nearby RS image tiles should be close, and those of distant tiles should be separated, according to Tobler's first law of geography [41].

D. Current Limitations in RS

Certainly, the Tile2Vec algorithm sets a path for learning more informative CNN-based characterizations of RS data from a completely unsupervised perspective. However, the task of generating highly meaningful representations of aerial scenes without using any kind of class label information still remains a very important challenge in RS [59], [60]. The recent availability of massive data archives, together with the constant development of the airborne and space acquisition technology, is steadily increasing the complexity of RS data and, hence, their semantic understanding. Precisely, this growing complexity often produces a huge within-class diversity and between-class similarity that introduce important limitations within the aforementioned learning scheme [39]. When integrating the unsupervised mode into the deep metric learning approach [38], it is logically necessary to train the CNN model by sampling negative RS image tiles within each batch. However, this strategy significantly reduces the capacity of the model to distinguish between a broader range of contrasting land cover types since the learning process is constrained by the tiles that can be sampled in a single batch. Note that this point may become particularly critical in complex large-scale archives, which stimulates the development of more advanced unsupervised characterization techniques within the RS field [18] and ultimately motivates the research conducted in this work.

E. Novelty of the Proposed Method

To address all these challenges, this article proposes a new unsupervised deep metric learning model that jointly exploits two different aspects: a spatially augmented contrastive loss and a momentum update-based optimization. In contrast to Tile2Vec [38], the proposed approach integrates a new spatial augmentation criterion that allows considering not only semantic similarities among nearby RS scenes but also the inherent semantic diversity of land cover concepts when learning the corresponding metric space in an unsupervised fashion. Note that this within-class variability has not yet been exploited in the context of characterizing unlabelled RS scenes despite the fact that it may become very useful to relieve the large-scale variance problem of RS data [40]. Using this methodological improvement, the proposed approach is able to avoid the triplet loss limitations with scalable data while also taking advantage of additional contrastive RS image pairs during training. In order to further improve such contrasting land cover variety, the proposed approach also utilizes a momentum update-based optimization [61]. The general idea behind the momentum update is based on managing a dynamic dictionary of encoders to enhance the contrastive learning process. Following this idea, we build a queue of deep embeddings of RS scenes in order to force the length of such queue to be larger than the minibatch size. Unlike the standard momentum scheme that shows limited results with large-scale

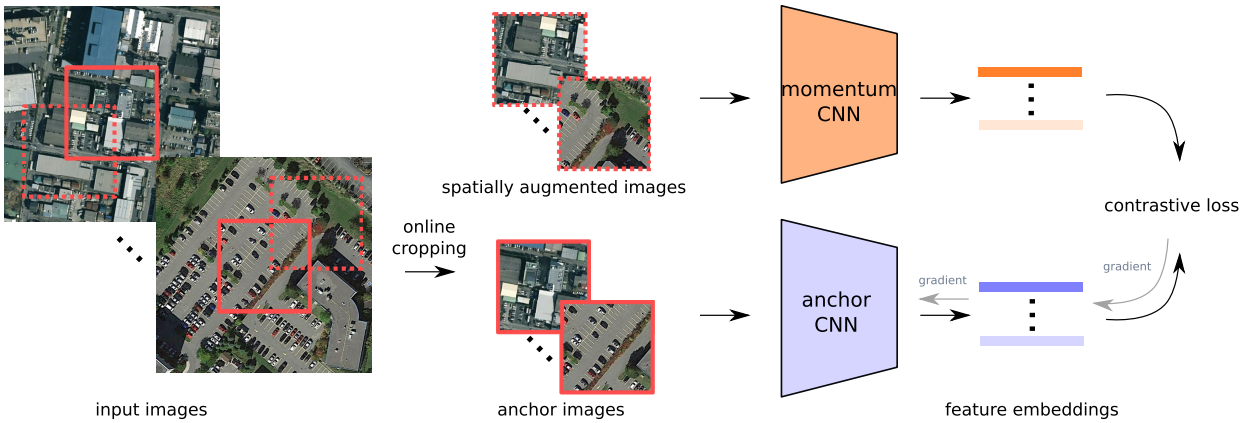

Fig. 1. Graphical illustration of the proposed unsupervised deep metric learning framework (SauMoCo), which has been specially designed to characterize unlabeled RS scenes. With the proposed approach, we aim to encode RS images into the learned metric space through the anchor CNN model, where nearby cropped tiles are grouped together and distant tiles are separated. The momentum CNN model is used to update the queue of deep embeddings.

data [61], the proposed end-to-end approach is designed to exploit vast unlabeled RS archives by using a CNN-based backbone architecture to jointly characterize land cover scenes and update the queue. Compared with different state-of-the-art methods to characterize unlabeled RS scenes, the proposed approach is able to achieve a better performance than the methods in [29], [38], [54], and [62], which also reveals the novelty and advantages provided by this work for the RS community.

III. SauMoCo

Our newly proposed end-to-end unsupervised deep metric learning model for characterizing unlabeled RS scenes (SauMoCo) can be summarized in the following three parts.

1) A backbone architecture (called *anchor CNN*) which is used to generate the corresponding feature embedding of the input RS scenes. Note that this CNN architecture can be defined according to a specific off-the-shelf topology, such as AlexNet [24], VGGNet [25], GoogleNet [26], and ResNet [27].
2) A spatially augmented loss, based on the contrastive loss formulation and a newly defined spatial augmentation criterion, which exploits not only the semantic similarities among nearby RS scenes but also the inherent diversity within land cover semantic concepts.
3) The corresponding optimization algorithm, which learns the proposed model parameters using a momentum contrast update. To achieve this goal, a queue of deep embeddings is constructed, and an additional CNN model (called *momentum CNN*) is introduced to update such queue. It is important to highlight that this network should be defined with the same architecture with regard to one of the anchor CNNs for a scalable training.

Fig. 1 shows in a graphical manner the proposed unsupervised deep metric learning framework. In the following sections, we detail the newly defined loss function and the considered optimization algorithm.

A. Spatially Augmented Loss

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ be a collected RS data set that consists of M images. From each image $\mathbf{x}_i$, an anchor patch $\mathbf{x}_i^a$ (located in its center) can be cropped with a certain size $W \times W$. With a certain distance d of the anchor patch $\mathbf{x}_i^a$, a neighborhood patch cropped from $\mathbf{x}_i$ is defined as its spatial augmentation, which is $\mathbf{x}_i^n$. If the distance is 100 pixels (in both vertical and horizontal directions), the center of $\mathbf{x}_i^n$ should be within 100 pixels with respect to the center of $\mathbf{x}_i^a$. Let $\mathbf{f}_i^a \in \mathbb{R}^D$ denote the deep embedding of $\mathbf{x}_i^a$ obtained by a CNN model $\mathcal{F}(\cdot;\theta)$ on the unit sphere (i.e., $\mathbf{f}_i^a = \mathcal{F}(\mathbf{x}_i^a;\theta)/\|\mathcal{F}(\mathbf{x}_i^a;\theta)\|_2$), where D is its dimension and θ represents the parameters of the CNN model. We identify this model as *anchor CNN*.

As noted by the first law of geography [41], everything is related to everything else, but nearby things are more related than distant things. Following this rule, the proposed method relies on the assumption that images that are geographic neighbors should be semantically more similar than distant images [38]. Therefore, the embeddings of nearby images should be closer than those of distant images in the metric space. However, it is important to highlight that the proposed spatial augmentation criterion is different from the one considered in other works, such as in [38]. Specifically, we do not fix the position of the spatially augmented patches to a specific neighbor position but to a neighborhood region of the anchor patch. As a result, our augmentation criterion allows certain spatial variations on the cropped areas in order to increase the variety of spatially augmented patches that are extracted online. To achieve this in a scalable way, we adopt a contrastive learning mechanism [63], [64], where the contrastive loss of $\mathbf{x}_i^a$ can be defined as

$$\mathcal{L}_i = -\log \frac{\exp\left(\left\langle \mathbf{f}_i^a, \mathbf{f}_i^n \right\rangle / \tau\right)}{\sum_{j=1}^{M} \exp\left(\left\langle \mathbf{f}_i^a, \mathbf{f}_j^a \right\rangle / \tau\right)}. \tag{1}$$

In this equation, the inner product $\langle \mathbf{f}_i^a, \mathbf{f}_i^n \rangle$ measures the cosine similarity between the embedding $\mathbf{f}_i^a$ of the anchor patch $\mathbf{x}_i^a$ and the one $\mathbf{f}_i^n$ of its spatial augmented patch $\mathbf{x}_i^n$. Besides, τ represents a temperature parameter controlling the concentration level of the sample distribution [65]. Intuitively, (1) describes the log-likelihood of the spatial augmented patch, which can be classified as its anchor patch among all the anchor patches in $\mathcal{X}$. Then, the corresponding contrastive loss

over the whole data set can be formally expressed as

$$\mathcal{L} = \sum_{i=1}^{M} \mathcal{L}_i. \tag{2}$$

By optimizing (2), we can obtain the deep embeddings of $\mathcal{X}$ and the trained CNN model, which is useful for characterizing unlabeled RS scenes and conducting the corresponding downstreaming land-cover categorization tasks.

B. Optimization via Momentum Update

In order to sufficiently train the CNN model based on (2) in an unsupervised fashion, a scalable data set is logically required to be fed into the deep model. For scalable data sets, how to sufficiently sample the negative patches (i.e., $\mathbf{x}_j^a$) with respect to $\mathbf{x}_i^a$ should be carefully defined since the number of spatially augmented patches and their consistency are both critical aspects within the proposed contrastive unsupervised learning scheme.

One common strategy adopted in the literature is based on sampling the negative patches within each minibatch [39]. However, this optimization mechanism has important limitations for training our deep model with scalable spatially augmented data. In more detail, this minibatch sampling process assumes that each patch can be seen once during one epoch of training, and hence, $\mathbf{x}_i^a$ only exists in one minibatch for the current iteration. Consequently, the CNN model is only able to see its corresponding negative patches $\mathbf{x}_j^a$ belonging to this minibatch, while other important samples outside the minibatch cannot be considered. Precisely, this fact can substantially reduce the semantic variety of contrasting land cover types during training, which is a key factor to allow learning more informative RS image representations from an unsupervised perspective.

To solve this problem, we adopt the momentum update rule [61], [66] for training our newly proposed unsupervised RS image characterization model. Specifically, a *queue* of the deep embeddings of image patches $\mathbf{x}_j^a$ is constructed, where the size of the queue is forced to be larger than that of the minibatch. In this way, the unsupervised learning process can be substantially enhanced by considering contrasting patches beyond a single batch. During the training phase, the embeddings of the current minibatch are compared with the ones in the queue, as they are progressively replaced. The embeddings of the current minibatch are enqueued and the oldest ones are dequeued. Moreover, in order to consistently update the deep embeddings in the queue, an auxiliary CNN model with parameter set θ_{aux} is introduced. We identify the θ_{aux} model as *momentum CNN* and it is updated as follows:

$$\theta_{\text{aux}}^{(t+1)} \leftarrow m\theta_{\text{aux}}^{(t)} + (1-m)\theta^{(t)} \tag{3}$$

where $m \in [0, 1)$ is a momentum coefficient. It is worth noting that only the CNN with θ is updated by means of backpropagation. The momentum CNN with parameters θ_{aux} can be evolved more smoothly than the CNN with θ. Then, the embeddings in the queue (encoded by the momentum CNN) are updated by

$$\hat{\mathbf{f}}_i^{(t+1)} \leftarrow \hat{\mathbf{f}}_i^{(t)}. \tag{4}$$

In this expression, $\hat{\mathbf{f}}_i$ denotes the features generated by the momentum CNN. In other words, the embeddings in the queue are replaced by the ones encoded by the momentum CNN after each training epoch. To this end, the proposed optimization mechanism is detailed in Algorithm 1.

Algorithm 1 Optimization Mechanism of SauMoCo

Require: $\mathbf{x}_i$

Initialize θ, θ_{aux}, τ, W, d, and D. Randomly initialize the queue.
for $t = 0$ to maxEpoch **do**
  Sample a mini-batch of $\mathbf{x}_i$.
  Within $\mathbf{x}_i$, randomly generate the center pixels of $\mathbf{x}_i^n$.
  Crop the patches $\mathbf{x}_i^n$ and $\mathbf{x}_i^a$ online.
  Obtain $\mathbf{f}_i^{(t)}$ based on CNN with $\theta^{(t)}$.
  Obtain $\hat{\mathbf{f}}_i^{(t)}$ based on the momentum CNN with $\theta_{\text{aux}}^{(t)}$.
  Calculate the loss in Equation (1) over the mini-batch and back-propagate the gradients.
  Update the parameters θ_{aux} of the momentum CNN via (3).
  Update the embeddings in the queue via Equation (4).
end for

IV. Experiments

A. Data Set Description

In this work, we use two benchmark RS image data sets to validate the effectiveness of the proposed method. A detailed description of the data sets is provided in the following.

1) *National Agriculture Imagery Program (NAIP) [38]:* This data set was generated for validating the Tile2Vec framework [38]. Specifically, it was collected from high-resolution RS images provided by the NAIP from the United States Department of Agriculture (USDA). All the images are located within the latitude from 36.45 to 37.05 and longitudes from -120.25 to -119.65. In this data set, there are totally 1000 images with a size of 50×50 pixels, a spatial resolution of 0.6 m, and four spectral bands (red, green, blue, and infrared). Each image is labeled using 28 classes obtained from cropland data layer (CDL), which are Corn, Cotton, Barley, Shrubland, Winter Wheat, Oats, Alfalfa, Grassland, Onions, Tomatoes, Fallow, Grapes, Other Tree Crops, Citrus, Almonds, Walnuts, Triticale, Pistachios, Garlic, Oranges, Pomegranates, Dbl Crop WinWht/Corn, Dbl Crop WinWht/Sorghum, Open Water, Developed/Open Space, Developed/Low Intensity, Developed/Med Intensity, and Developed/High Intensity. The NAIP data set is publicly available.[1]
2) *EuroSAT [67]:* This data set was created for land-use and land-cover classification based on multispectral RS images. In particular, it consists of 27 000 labeled and georeferenced Sentinel-2 images with a size of 64×64 pixels, a spatial resolution of 10 m, and 13 spectral bands covering the wavelength region from

[1] https://github.com/ermongroup/tile2vec

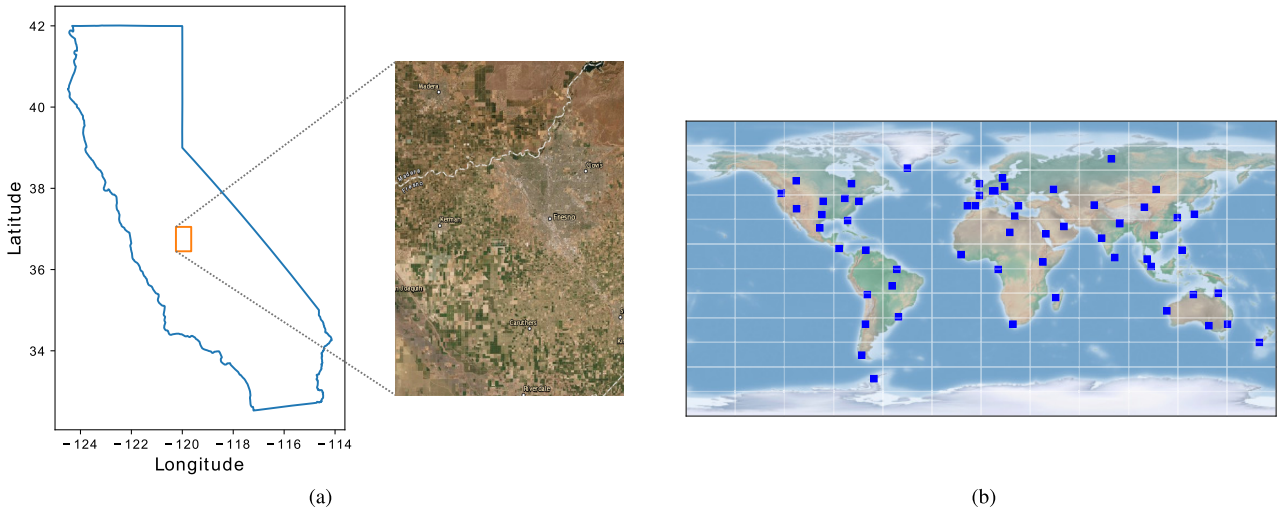

Fig. 2. Geolocations of the collected training data for the proposed method, evaluated on the NAIP and EuroSAT data sets. (a) We download 100 NAIP tiles near Fresno, CA, USA. (b) 100 Sentinel-2 tiles are downloaded all over the world.

443 to 2190 nm of the electromagnetic spectrum. Each image belongs to one class from a total of ten semantic land-cover categories: Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River, and Sea Lake. The EuroSAT archive is also publicly available[2].

These two RS archives have been selected to evaluate the performance of the unsupervised RS image characterization process from a single-source land cover acquisition perspective because they are two popular benchmark collections that also have available supplementary open access data for training the models with unlabeled scenes, that is, the NAIP and EuroSAT data sets are only used for assessment purposes, once the corresponding unsupervised characterization models have been trained with unlabeled NAIP and Sentinel-2 images, respectively. Specifically, we build two large-scale unlabeled training sets (one for NAIP and another for EuroSAT) using the following procedures.

1) In the case of NAIP, we download 100 NAIP full-scenes located in Central Valley areas near Fresno, CA, USA, through the USGS EarthExplorer[3] tool. The geolocations of the downloaded scenes are inside the orange rectangle in Fig. 2(a). Then, we randomly select a total of 100 000 images (with a size of 250×250 pixels) from the downloaded tiles.
2) In the case of EuroSAT, we downloaded 100 Sentinel-2 Level-1C image products that have been globally sampled from the entire globe. Fig. 2(b) shows the geolocations of the downloaded Sentinel-2 products. Then, we select 100 000 random images (with a size of 264×264 pixels) from the downloaded products.

Fig. 3 shows an example of the creation of the training data set. From one Sentinel-2 tile, we randomly crop one image (with the size of 264×264 pixels), considering that the sizes of the anchor and spatially augmented patches are of 64×64 pixels, and the distance between their centers is 100 pixels, according to the defined spatial augmentation criteria.

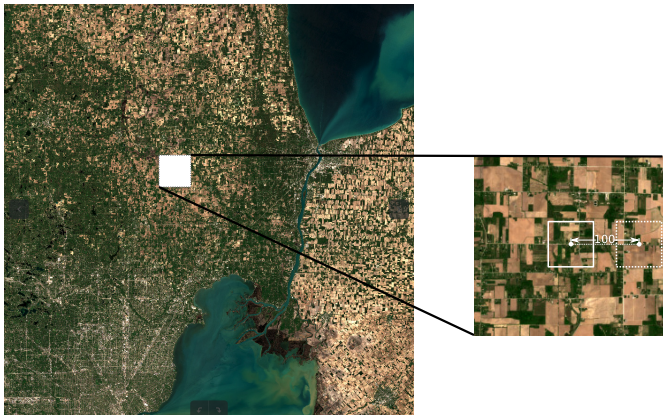

Fig. 3. Creation of a training data set from the downloaded tile. As an example, we randomly crop one image from a Sentinel-2 tile with the size of 264×264 pixels since the distance between the centers of the anchor patch and the spatially augmented patch is 100 pixels.

B. Experimental Setup

The proposed method is implemented in PyTorch [68]. The ResNet18 [27] network has been selected as elemental backbone architecture for extracting the corresponding deep embeddings of the RS images, that is, we use the ResNet18 model on both the anchor and momentum CNNs of the proposed approach. It is important to note that other architectures, such as ResNet50 or ResNet101, can be used within the proposed framework. Nonetheless, the ResNet18 model has been selected in this work because it usually provides a positive balance between complexity and performance in many different RS applications. During the training phase, the anchor and spatially augmented patches cropped from NAIP and Sentinel-2 images are with the size of 50×50 and 64×64 pixels, respectively, in order to be consistent with the

[2]http://madm.dfki.de/files/sentinel/EuroSATallBands.zip
[3]https://earthexplorer.usgs.gov/

benchmark data sets. *RandomFlip* and *RandomRotation* are adopted for the data augmentation. Regarding the considered parameters, τ and D are set to 0.25 and 128, respectively. In addition, the distance parameter d is set to 100 following the settings used in [38] and after contrasting this configuration on the NAIP data set. The stochastic gradient descent (SGD) optimizer is adopted for training. The initial learning rate is set to 0.01 and it is decayed by 0.5 every 30 epochs. The batch size is 256, and we totally train the CNN model for 100 epochs. In order to validate the effectiveness of the proposed approach with respect to different state-of-the-art methods, we include three different RS image characterization techniques in the experimental comparison: 1) the deep convolutional generative adversarial network (DCGAN) [54]; 2) MARTAaa GAN [62]; 3) the ResNet18 model pretrained on ImageNet while considering the most discriminating principal components (pretrained CNN+PCA) [29]; and 4) the Tile2Vec [38]. In the case of the pretrained CNN+PCA, it is important to highlight that we make use of the PCA method after extracting the pretrained features to generate the corresponding deep embeddings with the same dimensionality as the other methods. All the experiments are conducted on an NVIDIA Tesla P100 graphics processing unit (GPU).

To measure the effectiveness of the proposed approach compared with the other methods, we extract the deep embeddings for the NAIP and EuroSAT collections after training. Then, we use the available annotations to compute the corresponding classification results for each data set. In more detail, we provide five different experiments for validating and analyzing the results from several perspectives:

1) Evaluation of Deep Embeddings Based on Random Forest (RF) Classification: We first utilize the RF classifier to measure the classification performance based on the extracted feature embeddings of the two data sets obtained by the considered methods. For each data set, we randomly select 80% images for training the classifier and evaluate its performance on the rest 20% images. In order to obtain a mean score of the overall accuracy, a total of 100 trials are conducted. Then, we calculate the mean and standard deviation values of the obtained accuracy scores.

2) Visualization of Image Retrieval: In this experiment, we conduct a retrieval test to explore, from a qualitative perspective, the performance of the considered characterization methods. In particular, we extract one query image patch from a complete NAIP scene. Then, we use the pretrained CNN+PCA, Tile2Vec, and SauMoCo models to obtain the deep embeddings of the selected patch as well as the rest of the patches in the scene. Finally, we calculate their corresponding similarity maps with respect to the query and retrieve the ten nearest neighbor patches within the whole scene.

3) Evaluation of CNN Model Initialization: In this experiment, we utilize the ResNet18 network as a classifier by training this model with two different initializations, the pretrained ImageNet parameters and the parameters pretrained by our SauMoCo method. The objective is to evaluate the effectiveness of the proposed method as CNN model initialization. Specifically, we train the ResNet18 model on the EuroSAT data set using 80% of the images for training and 20% of the images for testing. To quantify the corresponding performance, we calculate the overall accuracy on the test set after each training epoch and observe the corresponding learning curve.

TABLE I

RF CLASSIFICATION PERFORMANCES BASED ON THE DEEP EMBEDDINGS EXTRACTED FROM THE CONSIDERED METHODS: DCGAN, MARTA GAN, PRETRAINED CNN+PCA, TILE2VEC, AND SAUMOCO

	NAIP	EuroSAT
DCGAN	62.2±2.8	63.9±0.6
MARTA GAN	60.2±2.9	72.6±0.6
Pre-trained+PCA	62.6±3.5	73.7±0.5
Tile2Vec	66.1±3.3	74.5±0.6
SauMoCo	**73.5±2.9**	**76.5±0.5**

4) Hyperparameter Analysis of SauMoCo: We investigate the sensitivity of the proposed model to the τ parameter. For each data set, we test eight different values in a range from 0.05 to 0.5. Then, we calculate the corresponding RF-based classification accuracy considering 80% of the images for training and 20% of the images for testing.

5) Comparison of Different CNN Backbone Architectures for SauMoCo: In the abovementioned experiments, we utilize ResNet18 as the CNN backbone architecture for extracting the feature embeddings. For evaluating the scene characterization performance by using different CNN backbone architectures of SauMoCo, we also utilize ResNet50 on the RF classification carried on the NAIP data set. The experiment setup is consistent with Section IV-B1.

C. Experimental Results

1) Evaluation of Deep Embeddings Based on RF Classification: In order to monitor the learning effectiveness, Fig. 4 shows the RF performances based on the deep embeddings extracted via the proposed SauMoCo, Tile2Vec, DCGAN, and MARTA GAN models during the training phase, that is, after each training epoch, we use the generated deep embeddings to calculate the corresponding RF-based classification results. As it is possible to observe, the deep embeddings based on SauMoCo outperform those extracted from the other compared methods in the training phases when the RF classification is applied. Regarding the DCGAN and MARTA GAN, this model provides the most unstable results while also leading to a clearly lower performance than both SauMoCo and Tile2Vec, which consistently achieve the best and second best performances. In the case of Tile2Vec, the triplet loss makes this model highly demanding because it requires a set of triplets with about $\mathcal{O}(|\mathcal{X}|^3)$ samples, which becomes unaffordable for scalable data sets. Precisely, this limitation may lead to a constrained training of the model so that the learned deep embeddings cannot properly represent a broader variety of land cover semantic concepts. In the proposed approach, the semantic similarities are calculated based on the images within each minibatch and also all the other images in the data set, due to the use of a queue of deep embeddings. Then, the SauMoCo model can be trained by capturing all

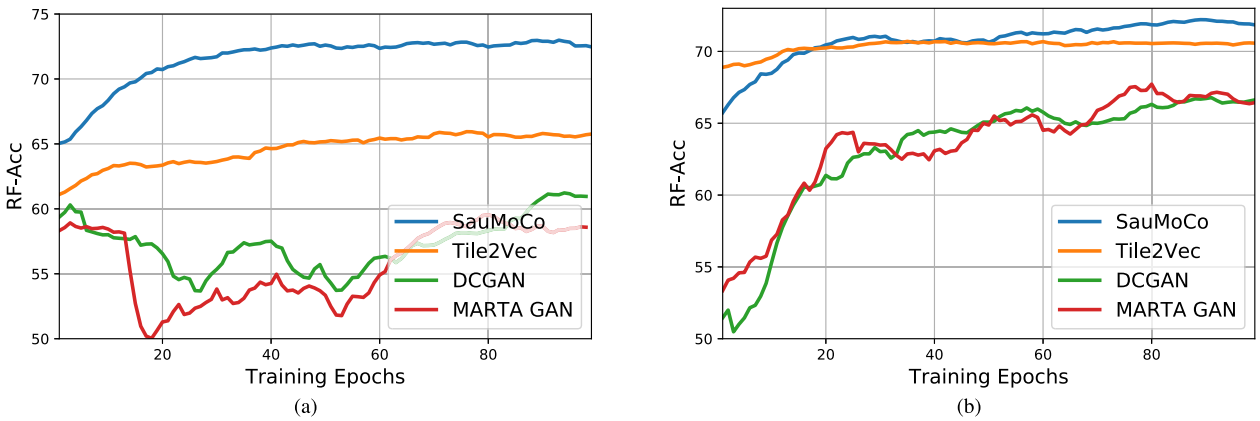

Fig. 4. RF classification performances based on the deep embeddings extracted from (a) NAIP and (b) EuroSAT data sets via the proposed method (SauMoCo), Tile2Vec, and DCGAN (during the training phase).

TABLE II
CLASSWISE F1 SCORES OBTAINED BY THE RF CLASSIFIER (WITH THE CONSIDERED UNSUPERVISED LEARNING METHODS) ON THE EUROSAT DATA SET

	DCGAN	MARTA GAN	Pre-trained CNN+PCA	Tile2Vec	SauMoCo
Annual Crop	67.15	67.01	**72.89**	69.64	72.60
Forest	83.20	86.54	89.78	**92.17**	91.35
Herb. Vegetation	69.69	68.60	**78.37**	75.54	77.45
Highway	29.50	34.47	40.15	**41.68**	36.43
Industrial	78.52	79.63	82.53	75.73	**84.04**
Pasture	65.63	66.31	73.73	**75.96**	74.09
Permanent Crop	63.84	62.32	67.43	64.24	**70.83**
Residential	68.18	63.41	**71.55**	66.82	69.79
River	77.73	**90.11**	81.93	84.83	80.22
Sea Lake	97.42	**99.67**	98.84	99.42	99.10

the possible distance metrics among the RS images in $\mathcal{X}$. Moreover, the spatially augmented images are cropped online, which also provides additional advantages as data augmentation strategy. By doing so, a higher semantic variety of similar images can be generated during the training phase (with respect to the anchor images). In comparison, the triplet set utilized in Tile2Vec is constructed beforehand and it does not exhibit any data augmentation capability during the training. Therefore, the proposed approach can also take advantage of the proposed spatial augmentation criteria.

Table I tabulates the RF classification performances using the considered RS image characterization methods, where the mean accuracy and the standard deviation scores are carried out based on 100 trials. From the reported results, it is possible to make some important observations. Specifically, it can be seen that SauMoCo achieves a remarkable improvement with respect to all the compared methods, being the accuracy gains between 7% and 10% for NAIP and between 2% and 12% for EuroSAT. In more detail, Tile2Vec consistently obtains the second best performance, followed by the pretrained CNN+PCA, DCGAN, and MARTA GAN. On the NAIP data set, DCGAN achieves a similar performance with regard to the one achieved by the pretrained CNN+PCA, while Tile2Vec and, especially, SauMoCo are able to provide superior results. On EuroSAT, the pretrained CNN+PCA and MARTA GAN perform significantly better than DCGAN. Compared with DCGAN, the introduced multiple-layer feature matching in MARTA GAN can improve the encoding performance of images via the discriminative model. However, Tile2Vec improves all these classification results and the proposed approach remarkably achieves the best performance. The results obtained in both collections reveal a similar trend concerning the good performance of Tile2Vec and the superior effectiveness of SauMoCo when characterizing unlabeled RS scenes.

To analyze the differences between Tile2Vec and SauMoCo in more detail, Table II provides the corresponding classwise F1 scores obtained by the RF classifier on the EuroSAT data set, where the two best results are highlighted in bold and gray-shaded font. As it is possible to observe, the proposed approach obtains the best and second best performances in two and five land-cover classes, respectively. Although Tile2Vec also exhibits positive results in five categories, the performance for the rest of the classes is rather limited, being even worse than that of the pretrained CNN+PCA and MARTA GAN in some cases. Precisely, these important differences make SauMoCo more stable and accurate from a global perspective, indicating that the proposed approach is able to extract

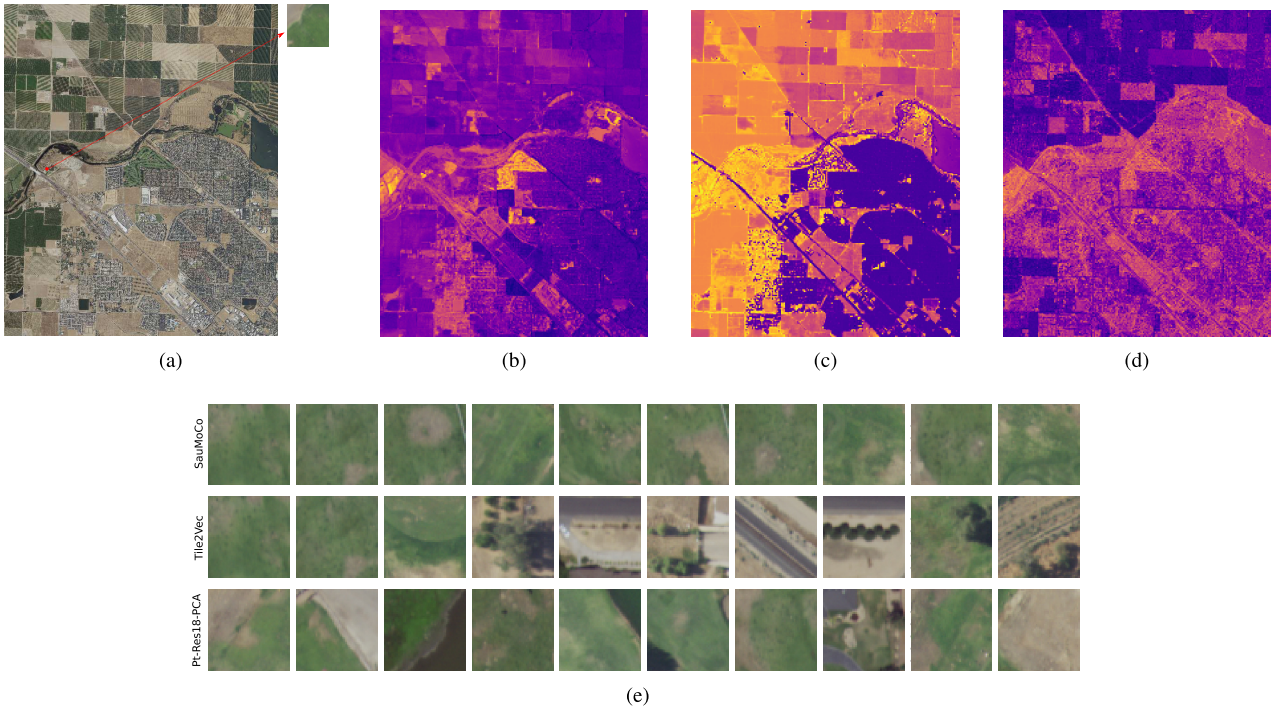

Fig. 5. Given the deep embedding of one image patch in an NAIP tile, similarity heatmaps can be obtained by calculating the similarities between the query patch and the rest of patches in the scene. (a) NAIP tile and one query image patch. (b)–(d) Similarity heatmaps of SauMoCo, Tile2Vec, and the pretrained CNN+PCA. (e) Top ten nearest neighbors with respect to the query image.

more relevant information about a wider range of semantic concepts.

2) Visualization of Image Retrieval: As shown in Fig. 5, we extract one query image patch from an NAIP tile. Then, we obtain its deep embedding and the ones of the rest of the patches in the tile based on the considered methods. Subsequently, we calculate their similarities and obtain the corresponding heatmaps for SauMoCo [Fig. 5(b)], Tile2Vec [Fig. 5(c)], and pretrained CNN+PCA [Fig. 5(d)], where brighter colors denote a higher similarity in the embedding space. In addition, the ten nearest neighbor patches are shown in Fig. 5(e). As it is possible to observe in the heatmaps, the locations of the most similar patches with respect to the query can be more clearly identified in Fig. 5(b). Precisely, these results indicate that the semantic information of the RS scene is not properly encoded based on Tile2Vec and the pretrained CNN+PCA model since there are larger parts in the image that are considered to be similar to the query in the embedding space. Regarding the nearest neighbor results, it is possible to see that the image patches retrieved from the embedding space generated by SauMoCo are the most visually similar with regard to the query, that is, the proposed approach is able to model the semantic content of the query more accurately than the other methods since all the retrieved images display similar land-cover patterns.

3) Evaluation of CNN Model Initialization: Fig. 6 shows the learning curves of the ResNet18 model over the EuroSAT collection when using two different initialization strategies: with the parameters obtained by the proposed approach and

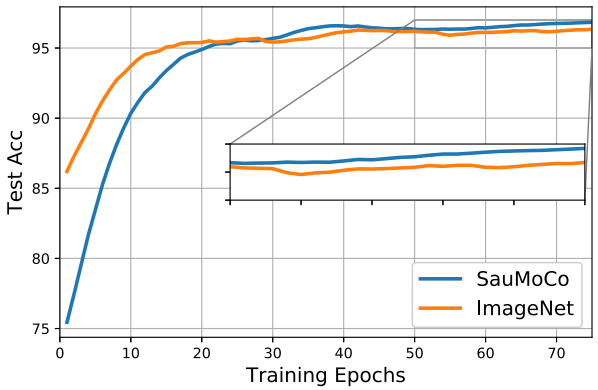

Fig. 6. Learning curves of the ResNet18 model on the EuroSAT data set when the model is initialized with different parameters, i.e., the pretrained parameters (via SauMoCo) and the ones from the pretrained ResNet18 model on ImageNet.

with the pretrained ImageNet parameters. According to the displayed results, it can be seen that the classification accuracy can be slightly improved when the parameters of the CNN model are initialized via the proposed approach. Although the pretrained ImageNet initialization exhibits higher classification accuracies at the beginning of the training process, SauMoCo is able to consistently achieve better results after 30 epochs. This fact reveals that the proposed approach is able to capture richer semantic information in the corresponding embedding space since a better minimum location of the loss function can be discovered by the parameters pretrained via SauMoCo.

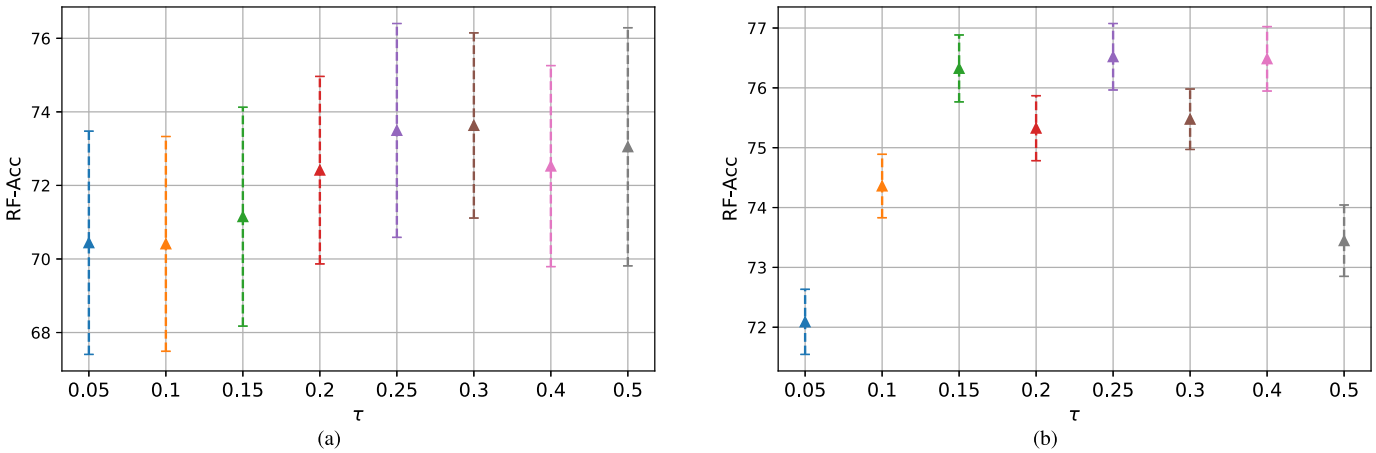

Fig. 7. Sensitivity analysis of τ on the two benchmark data sets. (a) NAIP and (b) EuroSAT, where we calculated the mean and standard deviation values of the RF classification results.

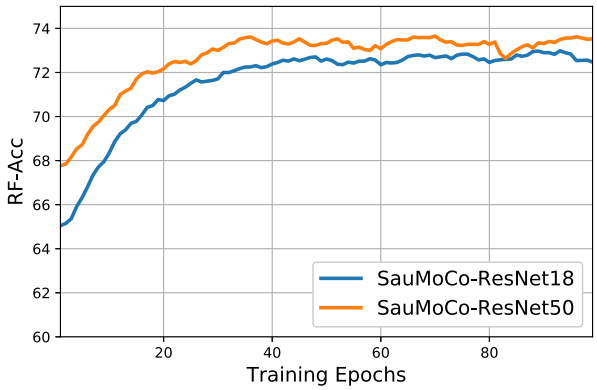

Fig. 8. RF classification performances based on the deep embeddings encoded by different CNN architectures (ResNet18 and ResNet50) on the NAIP data set under the training mechanism of SauMoCo.

4) Hyperparameter Analysis of SauMoCo: An important hyperparameter of the proposed approach is τ, which controls the concentration level of the sample distribution. To investigate the sensitivity of the proposed model to τ, we conduct several additional classification experiments on the embedding spaces generated by different hyperparameter values. In particular, Fig. 7 shows the effectiveness of the RF classification based on SauMoCo with respect to eight different values of τ on the two benchmark data sets: 1) NAIP and 2) EuroSAT. As it is possible to observe, the best classification performance in both data sets can be achieved when τ is 0.25. Nonetheless, the corresponding classification results are very consistent in the range from 0.15 to 0.4, which also indicates an adequate stability of the proposed approach with respect to the τ hyperparameter.

5) Comparison of Different CNN Backbone Architectures for SauMoCo: Fig. 8 shows the RF classification performances based on the deep embeddings encoded by different CNN architectures (ResNet18 and ResNet50) on the NAIP data set under the training mechanism of SauMoCo. It can be observed that the quality of the feature embeddings extracted from ResNet50 is slightly improved in comparison with ResNet18. As shown in Table III, compared with ResNet18, the classification accuracy based on the deep embeddings from ResNet50 can be improved with a score of 0.5% on the NAIP data set.

TABLE III
RF CLASSIFICATION PERFORMANCES BASED ON THE DEEP EMBEDDINGS ENCODED BY RESNET18 AND RESNET50 ON THE NAIP DATA SET

	RF
SauMoCo-ResNet18	73.5±2.9
SauMoCo-ResNet50	74.0±2.8

V. CONCLUSION AND FUTURE LINES

This article presents a new unsupervised deep metric learning framework (SauMoCo) to characterize unlabeled RS scenes. Specifically, the proposed approach initially defines a spatial augmentation criterion to uncover semantically similar RS images based on the first law of geography. Then, a queue of deep embeddings is built such that the size of queue is forced to be substantially larger than the batch size, to improve the semantic variety of contrasting land cover types during the training. To achieve this goal, an auxiliary CNN model is also used to consistently update the deep embeddings in the queue. The experimental part of the work, conducted over two benchmark data sets and based on the use of different characterization methods, reveals that the proposed unsupervised deep metric learning model is able to provide competitive advantages with respect to other state-of-the-art techniques in the task of representing unlabeled RS images.

One of the main conclusions that arises from this work is the relevance of considering a broader variety of land cover types when learning unsupervised RS image characterizations. In this regard, the proposed approach takes advantage of the defined spatial augmentation criteria and the considered queue of deep embeddings to enrich the semantic information of different RS categories during the contrastive learning process. Precisely, this feature allows our SauMoCo to enhance the global discrimination ability among unsupervised land-cover

classes and also to provide a more robust behavior with different data sets and settings. Due to the remarkable performance achieved by the presented method, our future work will be directed toward adapting it to intersensor data and other important RS tasks, such as dimensionality reduction of hyperspectral imagery or fine-grained land-use categorization.

REFERENCES

[1] 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.

[2] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep," 2020, *arXiv:2003.02822*. [Online]. Available: http://arxiv.org/abs/2003.02822

[3] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[4] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7109–7121, Dec. 2018.

[5] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.

[6] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[7] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, Jan. 2019.

[8] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[9] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[10] Z. Xiao, Y. Gong, Y. Long, D. Li, X. Wang, and H. Liu, "Airport detection based on a multiscale fusion feature for optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1469–1473, Sep. 2017.

[11] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[12] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.

[13] Y. Li, C. Peng, Y. Chen, L. Jiao, L. Zhou, and R. Shang, "A deep learning method for change detection in synthetic aperture radar images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5751–5763, Aug. 2019.

[14] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[15] 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.

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

[17] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[18] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.

[19] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[20] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "P-CNN: Part-based convolutional neural networks for fine-grained visual categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 6, 2019, doi: 10.1109/TPAMI.2019.2933510.

[21] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Apr. 27, 2020, doi: 10.1109/TGRS.2020.2985989.

[22] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, early access, May 18, 2020, doi: 10.1109/TGRS.2020.2991407.

[23] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–51.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[25] 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

[26] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] 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.

[29] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[30] E. Othman, Y. Bazi, N. Alajlan, H. Alhichri, and F. Melgani, "Using convolutional features and a sparse autoencoder for land-use scene classification," *Int. J. Remote Sens.*, vol. 37, no. 10, pp. 2149–2167, May 2016.

[31] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[32] P. Du, E. Li, J. Xia, A. Samat, and X. Bai, "Feature and model level fusion of pretrained CNN for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2600–2611, Aug. 2019.

[33] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.

[34] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

[35] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.

[36] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[37] Q. Zhu, Y. Zhong, S. Wu, L. Zhang, and D. Li, "Scene classification based on the sparse homogeneous-heterogeneous topic feature model," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2689–2703, May 2018.

[38] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2vec: Unsupervised representation learning for spatially distributed data," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3967–3974.

[39] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[40] B. Zhang *et al.*, "Remotely sensed big data: evolution in model development for information extraction [point of view]," *Proc. IEEE*, vol. 107, no. 12, pp. 2294–2301, Dec. 2019.

[41] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Econ. Geogr.*, vol. 46, pp. 234–240, Jun. 1970.

[42] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.

[43] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.

[44] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 44–59, Nov. 2018.

[45] A. S. Belward and J. O. Skøien, "Who launched what, when and why; trends in global land-cover observation capacity from civilian Earth observation satellites," *ISPRS J. Photogramm. Remote Sens.*, vol. 103, pp. 115–128, May 2015.

[46] R. Fernandez-Beltran, F. Pla, and A. Plaza, "Sentinel-2 and Sentinel-3 intersensor vegetation estimation via constrained topic modeling," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1531–1535, Oct. 2019.

[47] R. Fernandez-Beltran, F. Pla, and A. Plaza, "Endmember extraction from hyperspectral imagery based on probabilistic tensor moments," *IEEE Geosci. Remote Sens. Lett.*, early access, Jan. 13, 2020, doi: 10.1109/LGRS.2019.2963114.

[48] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Single-frame super-resolution in remote sensing: A practical overview," *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 314–354, Jan. 2017.

[49] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.

[50] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

[51] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "Remote sensing image fusion using hierarchical multimodal probabilistic latent semantic analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4982–4993, Dec. 2018.

[52] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[53] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[54] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.

[55] Y. Wang *et al.*, "Learning a discriminative distance metric with label consistency for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4427–4440, Aug. 2017.

[56] Z. Gong, P. Zhong, Y. Yu, and W. Hu, "Diversity-promoting deep structural metric learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 371–390, Jan. 2018.

[57] Y. Xing, M. Wang, S. Yang, and L. Jiao, "Pan-sharpening via deep metric learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 165–183, Nov. 2018.

[58] R. Cao *et al.*, "Enhancing remote sensing image retrieval using a triplet deep metric learning network," *Int. J. Remote Sens.*, vol. 41, no. 2, pp. 740–751, Jan. 2020.

[59] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

[60] P. Zhu *et al.*, "Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4047–4060, Jun. 2020.

[61] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2019, *arXiv:1911.05722*. [Online]. Available: http://arxiv.org/abs/1911.05722

[62] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: Unsupervised representation learning for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2092–2096, Nov. 2017.

[63] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.

[64] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*. [Online]. Available: http://arxiv.org/abs/1807.03748

[65] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

[66] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Trans. Geosci. Remote Sens.*, early access, May 12, 2020, doi: 10.1109/TGRS.2020.2991657.

[67] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[68] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

Jian Kang (Member, IEEE) received the B.S. and M.E. degrees in electronic engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2013 and 2015, respectively, and the Dr.-Ing. degree from Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, in 2019.

In August of 2018, he was a Guest Researcher with the Institute of Computer Graphics and Vision (ICG), TU Graz, Graz, Austria. He is with the Research Institute of Electronic Engineering Technology, Harbin Institute of Technology, Harbin, China, and with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin (TU Berlin), Berlin, Germany. His research focuses on signal processing and machine learning, and their applications in remote sensing. In particular, he is interested in multidimensional data analysis, geophysical parameter estimation based on InSAR data, SAR denoising, and deep learning-based techniques for remote sensing image analysis.

Dr. Kang obtained First Place of the Best Student Paper Award in EUSAR 2018, Aachen, Germany.

Ruben Fernandez-Beltran (Senior Member, IEEE) received the B.Sc. degree in computer science, the M.Sc. degree in intelligent systems, and the Ph.D. degree in computer science from Universitat Jaume I, Castellón de la Plana, Spain, in 2007, 2011, and 2016, respectively.

He is a Post-Doctoral Researcher with the Computer Vision Group, University Jaume I, as a member of the Institute of New Imaging Technologies. He has been a Visiting Researcher with the University of Bristol, Bristol, U.K., the University of Cáceres, Cáceres, Spain, and the Technische Universität Berlin, Berlin, Germany. He is a member of the Spanish Association for Pattern Recognition and Image Analysis (AERFAI), which is part of the International Association for Pattern Recognition (IAPR). His research interests lie in multimedia retrieval, spatiospectral image analysis, pattern recognition techniques applied to image processing, and remote sensing.

Dr. Fernandez-Beltran was awarded with the Outstanding Ph.D. Dissertation Award at Universitat Jaume I in 2017.

Puhong Duan (Member, IEEE) received the B.Sc. degree from Suzhou University, Suzhou, China, in 2014, and the M.S. degree from the Hefei University of Technology, Hefei, China, in 2017. He is pursuing the Ph.D. degree with the Laboratory of Vision and Image Processing, Hunan University, Changsha, China.

His research interests include image classification, visualization, object detection, and image fusion.

Sicong Liu (Member, IEEE) received the B.Sc. degree in geographical information system and the M.E. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2009 and 2011, respectively, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, in 2015.

He is an Assistant Professor with the College of Surveying and Geo-Informatics, Tongji University, Shanghai, China. His research interests include multitemporal remote sensing data analysis, change detection, multispectral/hyperspectral remote sensing, signal processing, and pattern recognition.

Dr. Liu was the winner (ranked Third Place) of the Paper Contest of the 2014 IEEE GRSS Data Fusion Contest. He is the Technical Co-Chair of the Tenth International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp 2019). He served as the Session Chair for many international conferences, such as the International Geoscience and Remote Sensing Symposium. He is also a referee for more than 20 international journals.

Antonio J. Plaza (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in computer engineering from Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura, Cáceres, Spain, in 1999 and 2002, respectively.

He is the Head of the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura. He has authored more than 600 publications, including over 200 JCR journal articles (over 160 in IEEE journals), 23 book chapters, and around 300 peer-reviewed conference proceeding papers. His research interests include hyperspectral data processing and parallel computing of remote sensing data.

Dr. Plaza is a fellow of the IEEE for contributions to hyperspectral data processing and parallel computing of Earth observation data. He was a member of the Editorial Board of the IEEE GEOSCIENCE AND REMOTE SENSING NEWSLETTER from 2011 to 2012 and the *IEEE Geoscience and Remote Sensing Magazine* in 2013. He was also a member of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS). He was a recipient of the Most Highly Cited Paper (2005–2010) in the *Journal of Parallel and Distributed Computing*, the 2013 Best Paper Award of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS), and the Best Column Award of the *IEEE Signal Processing Magazine* in 2015. He received the Best Paper Awards at the IEEE International Conference on Space Technology and the IEEE Symposium on Signal Processing and Information Technology. He received the recognition as a Best Reviewer of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2009 and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2010, for which he has served as an Associate Editor from 2007 to 2012. He has served as the Director of Education Activities for the IEEE Geoscience and Remote Sensing Society (GRSS) from 2011 to 2012 and as the President of the Spanish Chapter of the IEEE GRSS from 2012 to 2016. He has reviewed more than 500 manuscripts for over 50 different journals. He has served as the Editor-in-Chief for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING from 2013 to 2017. He has guestedited ten special issues on hyperspectral remote sensing for different journals. He is also an Associate Editor of the IEEE ACCESS (received the recognition as an Outstanding Associate Editor of the journal in 2017). Additional information: http://www.umbc.edu/rssipl/people/aplaza