Noise-Tolerant Deep Neighborhood Embedding for Remotely Sensed Images With Label Noise

Jian Kang, *Member, IEEE*, Ruben Fernandez-Beltran, *Senior Member, IEEE*, Xudong Kang, *Senior Member, IEEE*, Jingen Ni, *Senior Member, IEEE*, and Antonio Plaza, *Fellow, IEEE*

***Abstract*—Recently, many deep learning-based methods have been developed for solving remote sensing (RS) scene classification or retrieval tasks. Most of the adopted loss functions for training these models require accurate annotations. However, the presence of noise in such annotations (also known as label noise) cannot be avoided in large-scale RS benchmark archives, resulting from geo-location/registration errors, land-cover changes, and diverse knowledge background of annotators. To overcome the influence of noisy labels on the learning process of deep models, we propose a new loss function called noise-tolerant deep neighborhood embedding which can accurately encode the semantic relationships among RS scenes. Specifically, we target at maximizing the leave-one-out K-NN score for uncovering the inherent neighborhood structure among the images in feature space. Moreover, we down-weight the contribution of potential noisy images by learning their localized structure and pruning the images with low leave-one-out K-NN scores. Based on our newly proposed loss function, classwise features can be more robustly discriminated. Our experiments, conducted on two benchmark RS datasets, validate the effectiveness of the proposed approach on three different RS scene interpretation tasks, including classification, clustering, and retrieval. The codes of this article will be publicly available from https://github.com/jiankang1991.**

***Index Terms*—Deep metric learning, image characterization, image retrieval, label noise, remote sensing (RS).**

I. Introduction

WITH the rapid development of satellite sensors, remote sensing (RS) has entered the big data era. The availability of massive RS datasets now support a wide range of applications, such as object detection [1]–[5], land-cover characterization [6]–[16], or disaster monitoring [17], [18], among others. To successfully solve these tasks, accurate interpretation of the semantics within RS scenes is fundamental, with scene interpretation being a mainstream research topic [19].

Most existing scene interpretation methods can be categorized into two types: 1) handcrafted feature-based methods [20], [21]; and 2) data-driven feature-based methods [22]–[27]. Although handcrafted features can characterize most RS scenes, their performance is more limited as the complexity of the semantic contents increases. More recently, deep-learning methods (which directly summarize high-level semantics from large-scale RS data through end-to-end neural networks) have exhibited prominent performance for intelligently interpreting RS scenes [28], [29]. Specifically, deep metric learning methods have received particular attention in this context, owing to their good performance in the task of discriminating interclass features and discovering the inherent structure of intraclass features [30]–[32]. The main goal of these methods is to separate and group the features extracted from semantically similar and dissimilar images, respectively. To achieve this goal, most deep metric learning methods require supervised information (such as image annotations) for constructing image pairs or triplets with semantic relationships, where semantically similar images share the same label and dissimilar images have different labels. With the rapid growth of RS data archives, accurately annotating RS scenes has become a major challenge. Most RS scene benchmark datasets [33]–[38] are annotated by human experts with sufficient background knowledge. However, such labeling procedure is very expensive and time-consuming for large-scale RS datasets. Moreover, different annotators with different background knowledge may not reach an agreement when labeling semantically complex RS scenes, which may induce label noise in the dataset. An alternative way for scalable RS scene annotation is crowd-sourcing geospatial information, e.g., Google Maps, OpenStreetMap (OSM), CORINE Land Cover (CLC), etc. [39]–[41]. Although such approach can automatically annotate RS scenes in a scalable manner, some factors (such as geo-location/registration errors, land-cover changes, or even low-quality volunteered geographic information) could also introduce label noise. When noisy labels exist in the RS benchmark dataset, they will lead to performance degradation of trained deep models on different scene characterization tasks, such as classification or image retrieval, among others [42]. As a result, most existing methods for scene characterization based on deep metric learning are not robust to label noise.

Manuscript received January 6, 2021; revised January 29, 2021; accepted January 31, 2021. Date of publication February 2, 2021; date of current version February 24, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61871179, in part by the Ministry of Science, Innovation, and Universities of Spain under Grant RTI2018-098651-B-C54, in part by the Valencian Government of Spain under Grant GV/2020/167, in part by FEDER-Junta de Extremadura under Ref. GR18060, and in part by the European Union under the H2020 EOXPOSURE Project 734541. *(Corresponding author: Jingen Ni.)*

Jian Kang and Jingen Ni are with the School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China (e-mail: kangjian_1991@outlook.com; jni@suda.edu.cn).

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

Xudong Kang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: xudong_kang@163.com).

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

Digital Object Identifier 10.1109/JSTARS.2021.3056661

In order to solve this pressing issue, we propose a new deep metric learning loss function, termed as noise-tolerant deep neighborhood embedding (NTDNE), which can accurately capture the semantic relations among RS scenes in feature space. To model the inherent neighborhood structure of images in such space, we stochastically maximize the leave-one-out K-NN score, inspired by the concept of scalable neighborhood component analysis (SNCA) [43], [44]. Furthermore, considering the presence of label noise, we improve the robustness of the log-likelihood loss function of leave-one-out K-NN scores by replacing the logarithm function with the negative Box–Cox transformation [45], [46], which can down-weight the contribution of potentially noisy images by learning their localized structure in feature space. In addition, we apply a truncation mechanism on the loss function to further improve its robustness, especially when a large percentage of the datasets contain noisy labels. To this end, the contributions of this article can be summarized as follows.

1) We propose a novel deep metric learning loss function for modeling the inherent neighborhood structure of images in feature space in a robust manner.
2) The newly proposed loss function exhibits prominent performance in the task of feature generation (particularly in the presence of label noise) compared with other state-of-the-art deep metric learning losses.

The remainder of this article is organized as follows. Section II describes some related works. Section III details our newly proposed robust deep metric learning scheme to characterize RS images with label noise. Section IV presents the conducted experiments and discusses the obtained results. Section V concludes the article with some remarks and hints at plausible future research lines.

II. Related Work

A. RS Scene Characterization

Traditional RS scene characterization methods are developed in a handcrafted manner. Most of these methods exploit low-level visual descriptors to summarize color, texture, structure and shape information of land-use and land-cover objects within RS scenes. For example, [47] proposed a feature extraction method based on morphological texture for discriminating three mangrove species (and the surrounding environment) with multispectral IKONOS imagery. A gist feature-based method was proposed in [48] for automatically detecting and classifying targets in high-resolution broad-area satellite images. [49] exploited multiscale histogram of oriented gradients feature pyramids for extracting the features of objects and utilized support vector machine to achieve the classification. Other popular descriptors, such as local binary patterns and scale-invariant feature transform, have been also adopted to characterize the contents of RS scenes [50], [51]. Despite their advantages, they cannot sufficiently capture the features of some RS scenes with high semantic complexity. To solve this issue, data-driven RS scene characterization methods based on sparse coding, topic modeling, and auto-encoders [22], [52], [53] have been developed during the last decades. Most recently, deep learning techniques have attracted significant attention for characterizing the semantics of RS scenes, owing to the prominent capabilities of convolutional neural networks (CNN) in the task of extracting discriminative features from images [28]. For instance, Li *et al.* [54] integrated multilayer features of different pretrained CNN models for characterizing aerial scenes. Multiscale CNN features via spatial pyramid pooling were exploited in [55] for RS scene characterization. Zheng *et al.* [56] proposed a deep scene representation method utilizing pretrained CNN features, multiscale pooling, and Fisher vectors to achieve invariance of CNN features and enhance the discriminative capabilities. Huang *et al.* [26] first combined the multiscale deep feature learning strategy with manifold-learning-based dimension reduction for further feature embedding, which significantly improved the ability to embed local contextual information and learn discriminative features. Among various deep learning techniques, the so-called deep metric learning scheme has recently become a prominent trend to effectively encode the semantic contents of RS scenes with low-dimensional features, which can sufficiently model the semantic relationships among the images in the feature space. To improve the discriminative capability of CNN models, Cheng *et al.* [31] introduced a pairwise loss function as the regularizer of the cross-entropy loss, and proposed a discriminative CNN (D-CNN) accordingly. Yan *et al.* [32] exploited a deep metric learning scheme to reduce the data distribution bias in the embedding space, so that the scene classification accuracy on the target dataset (coming from a different domain) could be preserved. Cao *et al.* defined a content-based RS image retrieval framework based on the triplet loss, which exploits both positive and negative examples to learn the feature space more accurately. Yun *et al.* [57] proposed a new triangular loss function within a coarse-to-fine strategy for retrieving semantically similar images with certain variations of the image contents. Hong *et al.* [24] for the first time summarized several state-of-the-art fusion strategies into a general deep learning framework for multimodal RS image classification. This work has been widely recognized as a pioneering work in multimodal RS data analysis. Although the abovementioned deep metric learning methods can model the semantic relations among RS scenes, they are not robust to benchmark datasets containing noisy labels.

B. Robust Loss Function for Deep Learning

Due to different factors, benchmark datasets may be strongly affected by noisy labels which can severely decrease the classification performance of pretrained CNN models [58]–[60]. As the scale of data rapidly grows, such issue has become more important in the machine learning field. Ghosh *et al.* [61] investigated the robustness of the commonly utilized categorical cross entropy (CCE) loss, together with the mean absolute value of error (MAE), and empirically demonstrated that MAE is more noise-tolerant than CCE. Zhang *et al.*[46] further illustrated that the robustness improvements of MAE over CCE are due to the weight scheme of the loss gradients, and proposed a novel loss function, named generalized cross entropy. Wang *et al.* [62] exploited the idea of symmetric Kullback–Leibler divergence and proposed a symmetric cross entropy loss to tackle noisy labels.

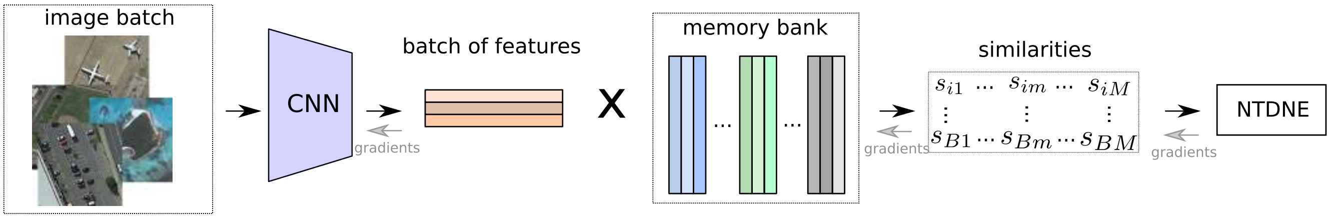

Fig. 1. Proposed framework based on NTDNE for RS scene characterization.

On the basis of $0-1$ loss (with some robust properties), Lyu *et al.* [63] proposed the curriculum loss as a tighter upper bound of the $0-1$ loss, which can be efficiently optimized. Moreover, robust deep learning methods have been recently reviewed in [64] and [65]. Although extensive research on robust deep learning has been done in machine learning and computer vision fields, most of the proposed losses are targeted to robustly predicting the categories of the input images. There is a lack of deep metric learning methods for robustly characterizing the semantics of RS scenes, which certainly motivates the development of new noise-tolerant deep metric learning models.

III. Noise-Tolerant Deep Neighborhood Embedding

With the guidance of supervised information, such as semantic labels, deep metric learning methods aim to produce CNN encoders which can preserve the semantic relations of images in the low-dimensional feature space. When label noise exists in the datasets, the trained CNN models cannot sufficiently capture the semantic contents of the images, which leads to a degraded discrimination capability of the generated features. To solve this issue, we introduce a new robust deep metric learning method in this section, which is mainly composed of two parts: 1) a backbone CNN architecture for modeling the semantics of RS images, based on low-dimensional features; and 2) a novel loss function, i.e., NTDNE, for learning the noise-tolerant CNN encoders, which can robustly capture the semantic similarities among the RS images. Fig. 1 provides a graphical illustration of our newly proposed method. In the following sections, we first introduce some notations, and then provide a review of SNCA and a description of the proposed NTDNE loss.

A. Notations

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a RS image dataset containing N images with label annotations, and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ be the associated set of labels, where each label is denoted by a one-hot vector, i.e., $\mathbf{y}_i \in \{0,1\}^C$ and C is the total number of categories. When the image $\mathbf{x}_i$ is annotated with the cth class, the cth element of $\mathbf{y}_i$ is 1, i.e., $y_i^c = 1$, and the others are 0. $\mathcal{F}(\cdot)$ is the nonlinear function modeled by the CNN encoders, which transform the input RS images into low-dimensional features $\mathbf{f}_i \in \mathbb{R}^D$ with dimensionality D. The normalized features, i.e., $\mathbf{f}_i = \mathcal{F}(\mathbf{x}_i)/\|\mathcal{F}(\mathbf{x}_i)\|_2$, are exploited in this article. When the dataset is corrupted by label noise, the noisy label set is denoted as $\hat{\mathcal{Y}} = \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N\}$, where $\hat{\mathbf{y}}_i$ represents the noisy label vector. In this article, we assume that the noise is conditionally independent of the input images given the true labels [46]

$$p(k|c, \mathbf{x}_i) = p(k|c) = \eta_{ck} \tag{1}$$

where η_{ck} is the noise rate, drawn as the (c, k)th component from a $C \times C$ probability transition matrix $\mathbf{Q}$ [66]. Two types of label noise are considered in this article, including *uniform* noise and *label-dependent* noise. In the case of uniform noise, a true label is randomly flipped into other labels with equal probability $\eta_{ck} = \frac{\eta}{C-1}$ or preserved as the true label with probability $\eta_{ck} = 1 - \eta$. In the case of label-dependent noise, a true label is more likely to be mistakenly labeled with a particular class with probability $\eta_{ck} = \eta$, or preserved as the true label with probability $\eta_{ck} = 1 - \eta$.

B. Review of SNCA and Its Limitations With Label Noise

Neighborhood component analysis [67] and its extension with a deep learning technique, i.e., SNCA [43] aim to maximize the averaged leave-one-out classification performance based on the input dataset. Specifically, given a pair of images $(\mathbf{x}_i, \mathbf{x}_j)$ and the corresponding features $(\mathbf{f}_i, \mathbf{f}_j)$, the semantic similarity of these two images can be measured by the cosine between the features

$$s_{ij} = \mathbf{f}_i^T \mathbf{f}_j. \tag{2}$$

For the image $\mathbf{x}_i$, we assume that an image $\mathbf{x}_j$ is located as its neighbor in feature space with probability p_{ij}, which can be defined as

$$p_{ij} = \frac{\exp(s_{ij}/\sigma)}{\sum_{k \neq i} \exp(s_{ik}/\sigma)}, \quad p_{ii} = 0 \tag{3}$$

where σ is a temperature parameter controlling the concentration level of the sample distribution [68], [69]. If s_{ij} is larger, $\mathbf{x}_j$ is more probably chosen as a neighbor of $\mathbf{x}_i$ in the feature space than another image $\mathbf{x}_k$. $p_{ii} = 0$ indicates that each image cannot be considered as its own neighbor. The probability p_i that $\mathbf{x}_i$ can be correctly classified is

$$p_i = \sum_{j \in \Omega_i} p_{ij} \tag{4}$$

where $\Omega_i = \{j | \mathbf{y}_i = \mathbf{y}_j\}$ is the index set of training images sharing the same label with $\mathbf{x}_i$. In order to maximize the leave-one-out classification score, the SNCA loss minimizes the expected negative log-likelihood over the training set, represented as

$$\mathcal{L}_{\text{SNCA}} = -\frac{1}{|\mathcal{T}|} \sum_i \log(p_i) \tag{5}$$

where $\mathcal{T}$ denotes the training set and $|\mathcal{T}|$ represents the number of training images. By calculating the gradients of $\mathcal{L}_{\text{SNCA}}$ with respect to $\mathbf{f}_i$ and $\mathbf{f}_j$ ($j \neq i$), we obtain

$$\frac{\partial \mathcal{L}_{\text{SNCA}}}{\partial \mathbf{f}_i} = \frac{1}{\sigma}\sum_{k} p_{ik}\mathbf{f}_k - \frac{1}{\sigma}\sum_{k\in\Omega_i} \tilde{p}_{ik}\mathbf{f}_k \tag{6}$$

$$\frac{\partial \mathcal{L}_{\text{SNCA}}}{\partial \mathbf{f}_j} = \begin{cases} \frac{1}{\sigma}(p_{ij} - \tilde{p}_{ij})\mathbf{f}_i, & j \in \Omega_i \\ \frac{1}{\sigma}p_{ij}\mathbf{f}_i, & j \notin \Omega_i \end{cases} \tag{7}$$

where $\tilde{p}_{ik} = p_{ik}/\sum_{j\in\Omega_i} p_{ij}$ is the normalized distribution of the ground-truth class. The robustness limitation of $\mathcal{L}_{\text{SNCA}}$ can be analyzed from the above two equations.

1) According to (6), the gradient of $\mathcal{L}_{\text{SNCA}}$ with respect to $\mathbf{f}_i$ depends on the two terms:[1] $\sum_k p_{ik}\mathbf{f}_k$ and $\sum_{k\in\Omega_i} \tilde{p}_{ik}\mathbf{f}_k$. Regarding the first term, its value does not depend on the annotated labels of the images, since it just calculates a weighted summation of all the other image features except $\mathbf{f}_i$, with the associated weight p_{ij}. Regarding the second term, its value is basically dependent on the weighted summation of all the other image features sharing the same label with respect to image $\mathbf{x}_i$, where the weights are mainly determined by the similarities s_{ik}. When label noise exists in the set Ω_i, the similarity s_{ik} between the image $\mathbf{x}_i$ and the noisy ones $\mathbf{x}_k$ will be small, and this will lead to a large gradient value $\frac{\partial \mathcal{L}_{\text{SNCA}}}{\partial \mathbf{f}_i}$. In other words, the images with lower similarities (which are most probably the ones with label noise) with respect to the image $\mathbf{x}_i$ have a stronger impact on the learning of feature $\mathbf{f}_i$.

2) For (7), when $j \in \Omega_i$, $(p_{ij} - \tilde{p}_{ij})$ can be formulated as

$$p_{ij}\left(1 - \frac{1}{\sum_{k\in\Omega_i} p_{ik}}\right). \tag{8}$$

Since $1 - \frac{1}{\sum_{k\in\Omega_i} p_{ik}} \leq 0$, a low value of p_{ij} will lead to a large gradient $\frac{\partial \mathcal{L}_{\text{SNCA}}}{\partial \mathbf{f}_j}$. Thus, when label noise exists in Ω_i, the associated images with lower similarity with respect to the image $\mathbf{x}_i$ will contribute more to the learning of the feature $\mathbf{f}_j$ than the other images. In the case of $j \notin \Omega_i$, when some images are annotated with wrong labels, $j \in \Omega_i$ may happen for some values of j. Thus, the associated p_{ij} will be large, which will also lead to a large gradient $\frac{\partial \mathcal{L}_{\text{SNCA}}}{\partial \mathbf{f}_j}$ when learning $\mathbf{f}_j$.

Based on the above analysis, we can note that the images with noisy labels will contribute more to the optimization of features $\mathbf{f}_i$ and $\mathbf{f}_j$ than the other images, which will cause the overfitting of the learned CNN models due to label noise.

C. Proposed NTDNE

1) Loss Function: To improve the noise-tolerant capability of SNCA, we observe that the SNCA loss is the negative summation of all the leave-one-out classification scores wrapped by the logarithm function. The gradient values with respect to the learned features are mainly dependent on the adopted wrap function. Inspired by [46], we exploit the negative Box–Cox transformation [45] to replace the logarithm function in SNCA, owing to its down-weighting effect on small values of p_i, which results in the proposed NTDNE loss

$$\mathcal{L}_{\text{NTDNE}} = \frac{1}{|\mathcal{T}|}\sum_i \frac{1-(p_i)^q}{q}, \quad q \in (0,1). \tag{9}$$

Lemma 1: $\lim_{q\to 0} \mathcal{L}_{\text{NTDNE}} = \mathcal{L}_{\text{SNCA}}$.

Proof: Based on L'Hôpital's rule, we have

$$\begin{aligned}\lim_{q\to 0}\frac{1}{|\mathcal{T}|}\sum_i \frac{1-(p_i)^q}{q} &= \frac{1}{|\mathcal{T}|}\sum_i \lim_{q\to 0}\frac{\frac{d}{dq}\left(1-(p_i)^q\right)}{\frac{d}{dq}q} \\ &= \frac{1}{|\mathcal{T}|}\sum_i \lim_{q\to 0} -(p_i)^q\log(p_i) \\ &= -\frac{1}{|\mathcal{T}|}\sum_i \log(p_i).\end{aligned} \tag{10}$$

As q gets smaller, the proposed NTDNE loss will approximate the SNCA loss. By calculating the gradients of $\mathcal{L}_{\text{NTDNE}}$ with respect to $\mathbf{f}_i$ and $\mathbf{f}_j$, we can obtain

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{NTDNE}}}{\partial \mathbf{f}_i} &= (p_i)^q\left(-\frac{1}{p_i}\frac{\partial p_i}{\partial \mathbf{f}_i}\right) = (p_i)^q\left(-\frac{\partial\left(\log(p_i)\right)}{\partial \mathbf{f}_i}\right) \\ &= (p_i)^q\left(\frac{1}{\sigma}\sum_k p_{ik}\mathbf{f}_k - \frac{1}{\sigma}\sum_{k\in\Omega_i}\tilde{p}_{ik}\mathbf{f}_k\right)\end{aligned} \tag{11}$$

$$\frac{\partial \mathcal{L}_{\text{NTDNE}}}{\partial \mathbf{f}_j} = \begin{cases} (p_i)^q\left(\frac{1}{\sigma}(p_{ij} - \tilde{p}_{ij})\mathbf{f}_i\right), & j \in \Omega_i \\ (p_i)^q(\frac{1}{\sigma}p_{ij}\mathbf{f}_i), & j \notin \Omega_i. \end{cases} \tag{12}$$

Differently to the gradients of the SNCA loss with respect to the image features in (6) and (7), the gradients in (11) and (12) involve one scaling factor $(p_i)^q$. Since $q < 1$ and $p_i \leq 1$, $(p_i)^q$ will have a down-weighting effect on the gradients $\frac{\partial \mathcal{L}_{\text{SNCA}}}{\partial \mathbf{f}_i}$ and $\frac{\partial \mathcal{L}_{\text{SNCA}}}{\partial \mathbf{f}_j}$. When label noise exists in the dataset, large gradients induced by the noisy samples for learning the features can be down-weighted by $(p_i)^q$. Thus, the proposed NTDNE loss function can improve the robustness to existing label noise in the dataset as compared with the SNCA loss. Although this down-weighting scheme can suppress the contributions of noisy images when learning the image features, they still can affect the optimization progress of the features, especially when the noise level is large. To further increase robustness, we adopt a "pruning" strategy during the training phase with the following modification of the NTDNE loss

$$\mathcal{L}_{\text{NTDNE}} = \frac{1}{|\mathcal{T}|}\sum_i \begin{cases} \frac{1-k^q}{q}, & \text{if } p_i \leq k \\ \frac{1-(p_i)^q}{q} & \text{if } p_i > k \end{cases} \tag{13}$$

where $k \in (0,1)$ denotes a threshold value. When the leave-one-out-classification score is less than k, the loss induced by the associated image will be cut to a constant value. Only if it is greater than k, the score can contribute to the overall NTDNE loss. By doing so, the samples with lower value of p_i (most probably these are the samples with label noise) can be "filtered out" during the learning progress of the features which, in turn, can further improve the noise-tolerance of NTDNE. To this end, a graphical illustration demonstrating the difference between SNCA and the proposed loss is given in Fig. 2.

[1] For simplicity, $\frac{1}{\sigma}$ is omitted here.

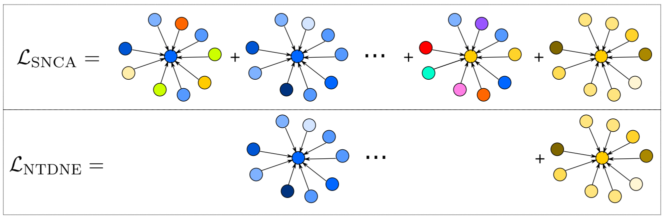

Fig. 2. Simplified graphical illustration of the differences between SNCA and NTDNE, where different colors refer to the features with different classes. Since the SNCA loss is constructed by the summation of all the leave-one-out K-NN classification scores, it cannot avoid that existing noisy labels in the dataset negatively impact the learning of features. By pruning some images with potentially noisy labels, our NTDNE loss will be less affected by the associated classification scores, which can robustly model the semantic relations among the images based on the learned features.

2) Optimization Strategy: To optimize $\mathcal{L}_{\text{NTDNE}}$ in (13), it can be reformulated as follows:

$$\mathcal{L}_{\text{NTDNE}} = \frac{1}{|\mathcal{T}|}\sum_i \left([p_i > k]\frac{1-(p_i)^q}{q} + (1-[p_i > k])\frac{1-k^q}{q}\right) \tag{14}$$

where $[\cdot]$ denotes the Iverson bracket. To practically optimize it, an indicator vector $\mathbf{w}$ is created, where each element indicates whether the condition $p_i > k$ is triggered or not. Within the first several training epochs, the CNN models cannot discriminate enough image features for accurate similarity measurement. Therefore, we first exploit the NTDNE loss in (9) to optimize the CNN models for T epochs. Then, the loss function is switched to the truncated version in (14). Following [43], a *memory bank* $\mathcal{B}$ is adopted to store the normalized features online, i.e., $\mathcal{B} = \{\mathbf{f}_i, \ldots, \mathbf{f}_{|\mathcal{T}|}\}$. In each iteration, the gradients are obtained through (11) and (12), and the parameters are updated by using the backpropagation technique. The update of $\mathcal{B}$ is done with the following empirical weighted average

$$\mathbf{f}_i^{(t+1)} \leftarrow \lambda\mathbf{f}_i^{(t)} + (1-\lambda)\mathbf{f}_i \tag{15}$$

where $\mathbf{f}_i^{(t+1)}$ denotes the updated feature in $\mathcal{B}$, $\mathbf{f}_i^{(t)}$ denotes the obtained feature based on the CNN model at the current iteration, and $\mathbf{f}_i$ is the previous feature stored in $\mathcal{B}$.

3) Complexity Analysis: The storage complexity of the online memory bank $\mathcal{B}$ is $\mathcal{O}(DN)$. The indicator matrix $\mathbf{w}$ requires an $\mathcal{O}(N)$ increase in memory. Let us assume that the batch size is b. The image similarities and leave-one-out classification scores both require $\mathcal{O}(bN)$ complexity for storing the values and their gradients. Compared with SNCA, the proposed method only introduces the indicator matrix $\mathbf{w}$, which will lead to a storage memory increase with $\mathcal{O}(N)$ complexity.

IV. Experiments

A. Experimental Setup

Several RS benchmark datasets have been considered in our experiments, including: 1) aerial image dataset (AID) [36]; and 2) NWPU-RESISC45 [19]. Both the AID and NWPU-RESISC45 datasets are designed for land-cover or land-use classification. The datasets are randomly split into training, validation, and test sets with percentages of 70% (training), 10% (validation), and 20% (testing). Two kinds of noise, including uniform and label-dependent, are added to the training sets with different noise rates, i.e., $\eta = 0.1, 0.3, 0.5, 0.7$. The design of label-dependent noise is consistent with our previous work [42]. To evaluate the performance of our newly proposed deep metric learning method, three downstream tasks are considered, including: 1) K-NN classification; 2) clustering; and 3) image retrieval.

1) K-NN Classification: The labels of the test images can be decided by majority voting based on their K nearest neighbors retrieved from the training sets. The Euclidean distance is exploited for the metric measurement in feature space. Note that we utilize the training sets with true labels in the evaluation phase. The overall accuracy is calculated for evaluation purposes.

2) Clustering: Based on the extracted features of the test images in the feature space, we first apply K-means clustering, then the clustered results are evaluated by normalized mutual information (NMI) [70] and unsupervised clustering accuracy (ACC), formulated as follows:

$$\text{NMI} = \frac{2 \times I(\mathbf{Y};\mathbf{C})}{H(\mathbf{Y}) + H(\mathbf{C})} \tag{16}$$

where $\mathbf{Y}$ represents the ground-truth class labels, and $\mathbf{C}$ denotes the cluster labels based on the clustering method. $I(\cdot;\cdot)$ and $H(\cdot)$ represent the mutual information and entropy function, respectively.

$$\text{ACC} = \max_{\mathcal{M}} \frac{\sum_{i=1}^{N} \delta(l_i = \mathcal{M}(c_i))}{N} \tag{17}$$

where l_i denotes the ground-truth class, c_i is the assigned cluster of image $\mathbf{x}_i$, and $\delta(\cdot)$ represents the Dirac delta function. $\mathcal{M}$ is a function than finds the best mapping between the estimated and ground-truth labels. These two metrics are utilized for validating the discrimination of the extracted features based on deep metric learning methods.

3) Image Retrieval: Given the query images, image retrieval aims to accurately and effectively find the most semantically similar images in a database by measuring the similarities of the features through the Euclidean distance in feature space. To evaluate the image retrieval performance, we demonstrate the precision-recall (PR) curve and calculate the mean average precision (MAP) with the form

$$\text{AP} = \frac{1}{Q}\sum_{r=1}^{R} P(r)\delta(r) \tag{18}$$

where Q is the number of ground-truth RS images in the dataset that are relevant with respect to the query image, $P(r)$ denotes the precision for the top r retrieved images, and $\delta(r)$ is an indicator function to specify whether the rth relevant image is truly relevant to the query. During the evaluation phase, the test sets are exploited for querying and the training set is the database to be retrieved.

TABLE I
K-NN ($K = 10$) Classification Accuracies (%) of the Considered Methods on Two Benchmark Datasets With Two Types of Noise At Different Levels (Uni:uniform, LD:Label-Dependent)

			D-CNN	Triplet	NSL	SNCA	ArcFace	MAE	SCE	t-RNSL	NTDNE
AID	Uni	0.1	92.40	91.80	89.35	90.65	90.30	82.15	92.95	**94.05**	93.40
		0.3	86.80	85.35	84.35	82.70	79.95	82.70	85.85	91.80	**92.25**
		0.5	75.25	77.15	75.60	64.55	87.30	79.90	77.70	89.50	**90.25**
		0.7	60.10	55.35	63.90	42.90	82.30	80.65	70.20	78.25	**86.10**
	LD	0.1	92.45	93.30	90.20	90.40	90.80	83.45	91.70	**93.80**	92.9
		0.3	88.95	90.50	87.25	81.95	81.35	81.05	83.50	90.50	**90.75**
		0.5	84.40	85.80	85.15	75.40	86.55	82.25	68.60	86.05	**91.55**
		0.7	84.25	85.60	84.15	74.45	83.60	81.15	47.30	81.55	**90.15**
NWPU-RESISC45	Uni	0.1	90.05	86.92	87.46	87.90	87.75	78.69	88.70	**92.30**	91.49
		0.3	81.92	75.19	78.73	76.08	88.68	78.53	77.97	**90.76**	89.37
		0.5	**88.62**	61.57	65.57	59.75	85.52	75.43	61.16	88.56	88.33
		0.7	40.59	50.68	45.38	30.03	80.71	74.03	35.54	79.84	**85.03**
	LD	0.1	90.06	90.06	88.27	87.17	87.16	77.94	88.84	**92.03**	91.14
		0.3	85.70	87.76	84.14	77.97	71.25	77.65	81.79	89.30	**89.97**
		0.5	80.00	83.41	80.19	68.02	80.35	76.97	72.35	84.84	**86.30**
		0.7	77.19	79.25	78.00	63.44	78.25	77.17	67.51	76.84	**87.41**

We utilize ResNet18 [71] as the CNN backbone to extract image features. Other CNN architectures, such as ResNet50, can be also adopted. The input images are all resized to 256×256 pixels. Three data augmentation methods including: 1) RandomGrayscale; 2) ColorJitter; and 3) RandomHorizontalFlip are exploited for increasing the variation of the training sets. The parameters D, σ, k, q, and T are empirically set to $128, 0.1, 0.3, 0.7$, and 20, respectively. The stochastic gradient descent (SGD) optimizer (with initial learning rate set to 0.01) is utilized for the optimization, and the learning rate is decayed by a factor of 0.5 every 30 epochs. The batch size is set to 256 and the CNN models are trained for 100 epochs. We compare the proposed method with several state-of-the-art deep metric learning methods including the following:

1) D-CNN [31] where a joint loss composed of cross entropy term and metric learning term is proposed to achieve better discrimination of CNNs;
2) Triplet [72] where the negative features should be pushed away with a certain distance from the anchors with respect to the positive features;
3) SNCA [43];
4) NSL [73] which is a normalized version of the cross entropy loss;
5) ArcFace [74] which is a marginalized version of the cross entropy loss;
6) MAE [61] which minimizes the mean absolute errors between the softmax scores and the one-hot label vectors;
7) SCE [62] which is a symmetric cross entropy loss for robust classification; and
8) t-RNSL [42] which is the robust version of the normalized cross entropy loss in the framework of deep metric learning.

The parameters of the baseline methods are tuned to obtain optimal performance. The proposed method is implemented in PyTorch [75]. All the experiments are performed on an NVIDIA Tesla P100 graphics processing unit.

B. *Experimental Results*

1) KNN Classificaiton: Table I provides the K-NN classification ($K = 10$) results on the test sets with the aim of evaluating the trained CNN models on the noisy data via all the considered losses. It can be observed that the proposed method can achieve the best performance on both datasets with different levels of label noise. Since D-CNN, Triplet, NSL, SNCA, and ArcFace are not robust to noisy labels, their classification performances decrease as the noise rate increases. In comparison, NTDNE exhibits performance stability against label noise, especially when the noise rate is large ($\eta = 0.7$). Compared with MAE and SCE, NTDNE also achieves better classification results. Since both MAE and SCE are designed for learning deep classifiers, they are not focused on modeling the semantic relations among the deep features extracted from the images. Although they are robust deep classifiers, their K-NN classification performances (determined by the distance measurements among the deep features) cannot achieve comparable accuracies with respect to deep metric learning losses, such as t-RNSL and NTDNE. The t-RNSL is utilized for learning optimal class prototypes by pulling within-class deep features toward the associated prototype during the training phase. However, the t-RNSL cannot discover the inherent neighborhood structure of the images in feature space. In comparison, for each image in the feature space, NTDNE aims at selecting the most semantically similar images as its neighbors, so that the localized feature structure can be well preserved.

2) Clustering: In order to evaluate the clustering performance of CNN models trained with different losses, we first extract the deep features of the test sets and calculate the NMI and ACC scores displayed in Tables II and III, separately. As in the previous section, NTDNE outperforms the other methods in terms of both metrics by a large margin. For example, when $\eta = 0.7$, NTDNE can achieve a significant performance improvement (more than 5%) compared with other methods for

TABLE II
NMI SCORES (%) OF THE CONSIDERED METHODS ON TWO BENCHMARK DATASETS WITH TWO TYPES OF NOISE AT DIFFERENT LEVELS (UNI:UNIFORM, LD:LABEL-DEPENDENT)

			D-CNN	**Triplet**	**NSL**	**SNCA**	**ArcFace**	**MAE**	**SCE**	**t-RNSL**	**NTDNE**
AID	Uni	0.1	87.27	87.47	83.94	86.56	87.40	67.65	88.83	**91.45**	91.19
		0.3	78.52	78.10	72.87	76.19	72.26	68.10	79.51	88.11	**89.56**
		0.5	58.90	65.38	47.02	51.96	80.32	61.24	70.80	85.57	**86.82**
		0.7	30.41	37.67	22.29	25.40	71.67	61.40	63.50	66.23	**80.55**
	LD	0.1	88.24	89.78	86.95	87.53	87.39	71.43	86.81	**91.62**	90.76
		0.3	83.44	85.34	79.60	75.27	75.54	63.59	75.27	86.13	**86.79**
		0.5	75.06	77.20	74.63	70.09	75.18	63.95	51.47	79.87	**86.94**
		0.7	72.88	71.94	72.57	67.55	71.42	63.57	22.85	68.08	**85.46**
NWPU-RESISC45	Uni	0.1	85.26	82.32	81.31	84.47	83.98	61.72	83.53	**89.28**	**89.28**
		0.3	74.38	64.78	66.84	69.37	83.03	61.67	68.82	**88.01**	86.24
		0.5	84.54	48.59	41.86	50.02	77.90	57.61	46.25	**84.81**	84.51
		0.7	17.60	35.60	11.55	19.19	70.45	53.76	14.10	74.04	**80.72**
	LD	0.1	84.05	86.23	83.74	84.02	83.53	60.59	85.00	**89.17**	88.45
		0.3	79.21	82.54	77.48	73.73	67.80	59.13	77.11	85.04	**86.93**
		0.5	72.93	76.18	72.00	64.69	69.95	59.70	67.91	79.47	**82.89**
		0.7	69.56	69.28	69.65	59.29	67.08	58.13	63.30	68.32	**83.54**

TABLE III
ACC SCORES (%) OF THE CONSIDERED METHODS ON TWO BENCHMARK DATASETS WITH TWO TYPES OF NOISE AT DIFFERENT LEVELS (UNI:UNIFORM, LD:LABEL-DEPENDENT)

			D-CNN	**Triplet**	**NSL**	**SNCA**	**ArcFace**	**MAE**	**SCE**	**t-RNSL**	**NTDNE**
AID	Uni	0.1	91.65	89.20	85.10	90.00	90.55	63.85	92.50	**94.00**	93.25
		0.3	80.90	79.40	71.85	81.75	79.25	66.45	81.55	91.40	**92.05**
		0.5	65.50	65.25	45.30	60.20	78.50	53.30	68.30	86.60	**89.45**
		0.7	28.85	34.20	21.75	26.30	68.95	57.75	55.30	69.55	**82.10**
	LD	0.1	87.70	89.70	87.25	90.15	87.90	67.15	90.50	**93.50**	93.00
		0.3	79.50	85.20	77.95	79.05	77.00	57.20	78.25	86.65	**90.50**
		0.5	73.35	70.95	70.70	63.80	65.30	56.45	57.75	76.55	**87.80**
		0.7	64.50	65.30	68.55	59.55	62.00	56.65	23.70	63.20	**84.95**
NWPU-RESISC45	Uni	0.1	86.86	79.21	80.97	88.00	87.70	52.13	85.57	**91.48**	89.29
		0.3	77.86	60.40	69.41	76.06	78.48	52.97	75.90	**90.27**	88.90
		0.5	85.00	42.54	43.81	58.59	75.41	50.70	53.90	84.98	**86.14**
		0.7	17.41	25.30	10.75	24.62	68.38	46.46	15.51	78.16	**83.13**
	LD	0.1	85.71	84.41	83.19	86.94	86.87	53.33	84.95	89.22	**90.94**
		0.3	78.65	80.94	74.95	76.21	68.49	50.13	73.83	86.63	**89.06**
		0.5	66.40	72.60	69.35	60.03	52.73	51.84	66.13	75.10	**83.22**
		0.7	59.78	57.54	65.65	47.02	48.62	49.37	58.02	61.22	**81.43**

both NMI and ACC metrics. For illustrative purposes, we project the features of the AID test set, which are obtained via the CNN models trained on all the considered losses, into a 2-D space by the t-distributed stochastic neighbour embedding (t-SNE) and visualize the obtained results in Fig. 3. As it can be seen, when the noise rate increases, most losses cannot be utilized for learning the discriminative classwise features. In comparison, the interclass separability and the intraclass compactness of the features produced via NTDNE can be very well preserved when η changes. Therefore, when NTDNE is exploited, the pseudo-labels generated by K-means clustering of the features can match better the ground-truth labels.

3) Image Retrieval: Fig. 4 shows the PR curves of all the considered methods (based on the trained CNN models) when the datasets are corrupted by uniform label noise with $\eta = 0.5$. Compared with other methods, the proposed approach exhibits the best image retrieval performance, since it is robust to the label noise and can discover the neighborhood structure of the image features. Moreover, the MAP results are demonstrated in Table IV. It can be seen that our NTDNE can achieve the best retrieval accuracy with all the experimental settings, which indicates that the proposed method can be exploited for large-scale RS image retrieval in a robust manner. Fig. 5 provides some image retrieval examples based on SNCA, t-RNSL, and NTDNE. For the considered query images (with complex semantic content) both SNCA and t-RNSL cannot accurately retrieve the most semantically similar images from the database. For example, for the t-RNSL, *intersection* is confused with *basketball court* and

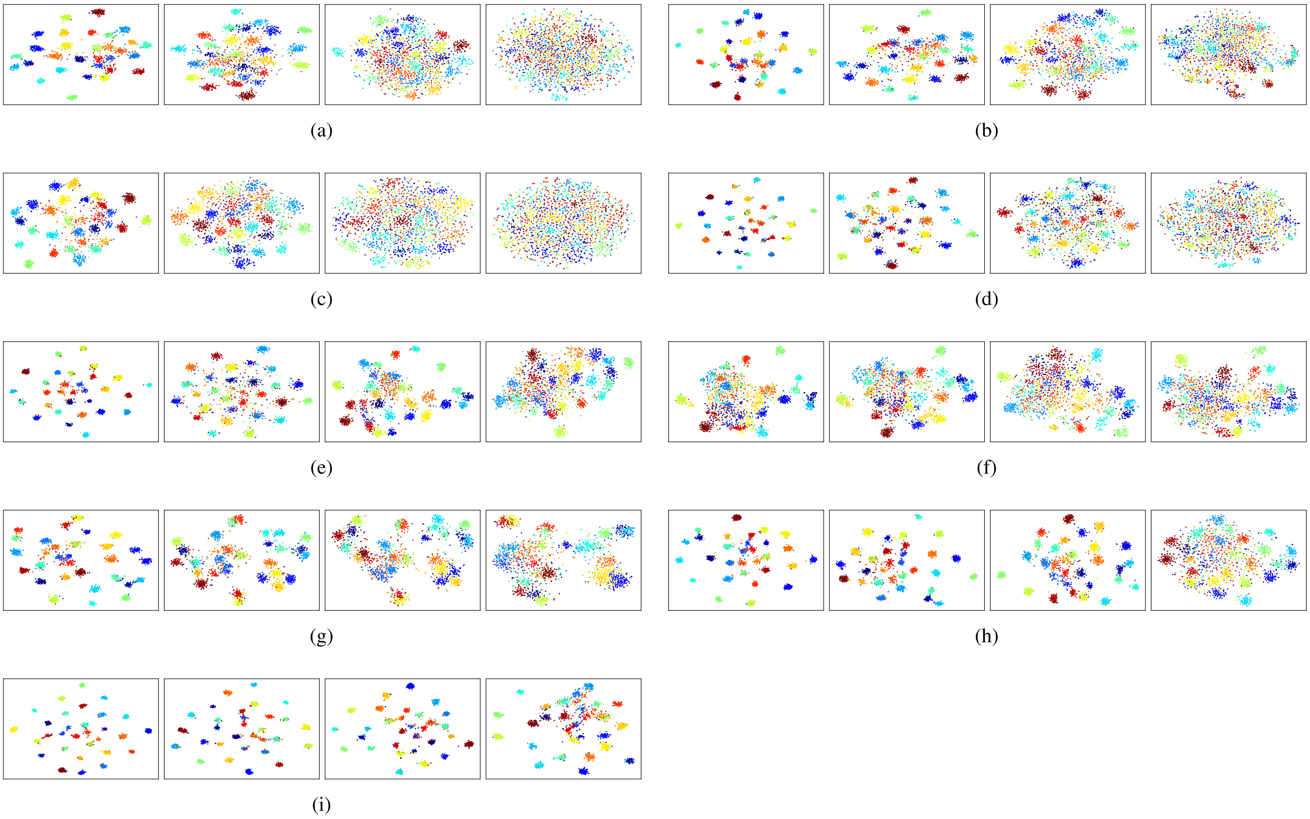

Fig. 3. 2-D projection of the extracted features from the AID test set (via the trained CNN models) on the training set with different levels of uniform label noise, i.e., $\eta = 0.1, 0.3, 0.5, 0.7$ –from left to right–. (a) D-CNN. (b) Triplet. (c) NSL. (d) SNCA. (e) ArcFace. (f) MAE. (g) SCE. (h) t-RNSL. (i) NTDNE.

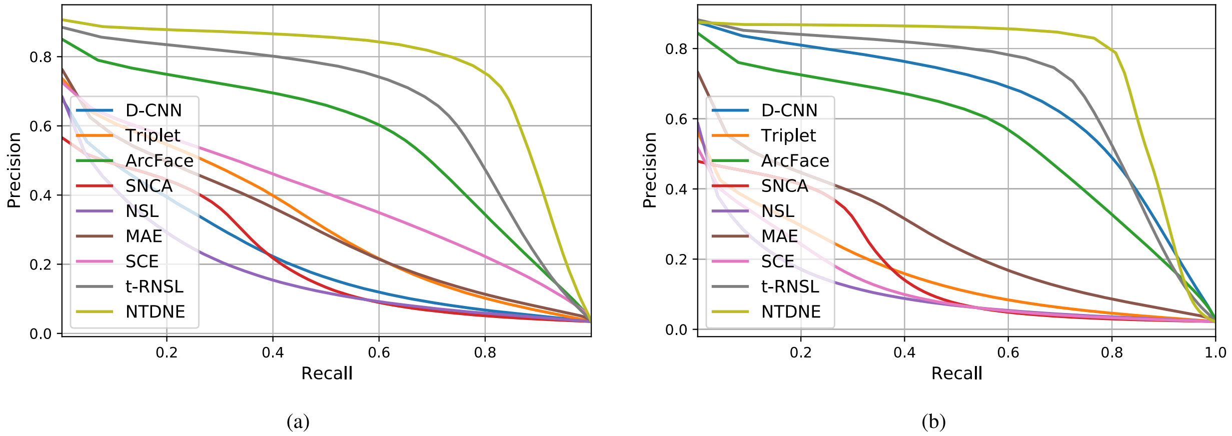

Fig. 4. Image retrieval results described by the PR curves of all the methods when the training sets contain uniform label noise with $\eta = 0.5$. (a) AID. (b) NWPU-RESISC45.

dense residential, while NTDNE can correctly retrieve the most semantically similar images for the same classes.

4) Hyperparameter Analysis: One of the main hyperparameters in the proposed method is k, which is the threshold value controlling when the loss starts to be effective. We analyze its sensitivity based on K-NN classification on the features of the test sets, considering the case when the training sets are with uniform label noise ($\eta = 0.5$). Fig. 6 shows the classification results obtained when k varies from 0.1 to 0.5. It can be observed that the performance of NTDNE is stable when k is in the range from 0.1 to 0.5. As a result, it is suggested to set q to a constant value (e.g. 0.7) [42], [46], and σ to a relatively small number (e.g. 0.05 or 0.1). Another hyperparameter T is the number of epochs from which the truncation starts. Empirically, it can be set as a number during the early stage of the training phase, e.g., 20 out of 100 total epochs.

TABLE IV

MAP SCORES (%) OF THE CONSIDERED METHODS ON TWO BENCHMARK DATASETS WITH TWO TYPES OF NOISE AT DIFFERENT LEVELS AND $R = 20$ (UNI:UNIFORM, LD:LABEL-DEPENDENT)

			D-CNN	Triplet	NSL	SNCA	ArcFace	MAE	SCE	t-RNSL	NTDNE
AID	Uni	0.1	93.25	93.13	90.61	96.81	96.04	79.18	94.21	96.04	**97.44**
		0.3	84.63	85.87	81.01	89.12	85.55	80.09	85.38	93.71	**95.89**
		0.5	69.17	74.96	67.96	68.92	85.41	74.04	75.58	90.55	**93.15**
		0.7	50.98	56.34	55.26	48.05	77.27	74.77	69.20	77.41	**89.18**
	LD	0.1	93.44	93.79	90.40	95.41	94.72	81.88	94.62	95.38	**96.96**
		0.3	87.28	90.64	85.50	85.72	83.88	76.29	85.66	91.37	**94.39**
		0.5	81.46	86.44	81.84	78.64	84.47	77.04	66.56	86.53	**93.54**
		0.7	78.61	83.56	78.96	76.99	81.73	76.27	45.82	77.79	**91.94**
NWPU-RESISC45	Uni	0.1	91.60	87.99	88.71	96.09	96.17	76.03	92.69	95.10	**97.42**
		0.3	80.94	75.10	76.53	85.49	88.59	75.57	80.62	94.11	**94.62**
		0.5	90.28	61.50	62.30	67.07	85.20	73.13	61.89	91.33	**93.02**
		0.7	41.53	51.79	44.44	42.06	78.30	67.74	40.88	83.29	**89.64**
	LD	0.1	91.02	90.95	89.39	92.71	93.84	81.88	91.99	94.83	**96.61**
		0.3	85.24	89.57	83.42	83.24	77.83	76.29	83.86	91.32	**93.59**
		0.5	77.80	84.42	78.14	72.64	79.69	77.04	73.33	85.35	**90.21**
		0.7	73.44	78.85	75.52	69.38	77.37	76.27	68.09	76.23	**90.40**

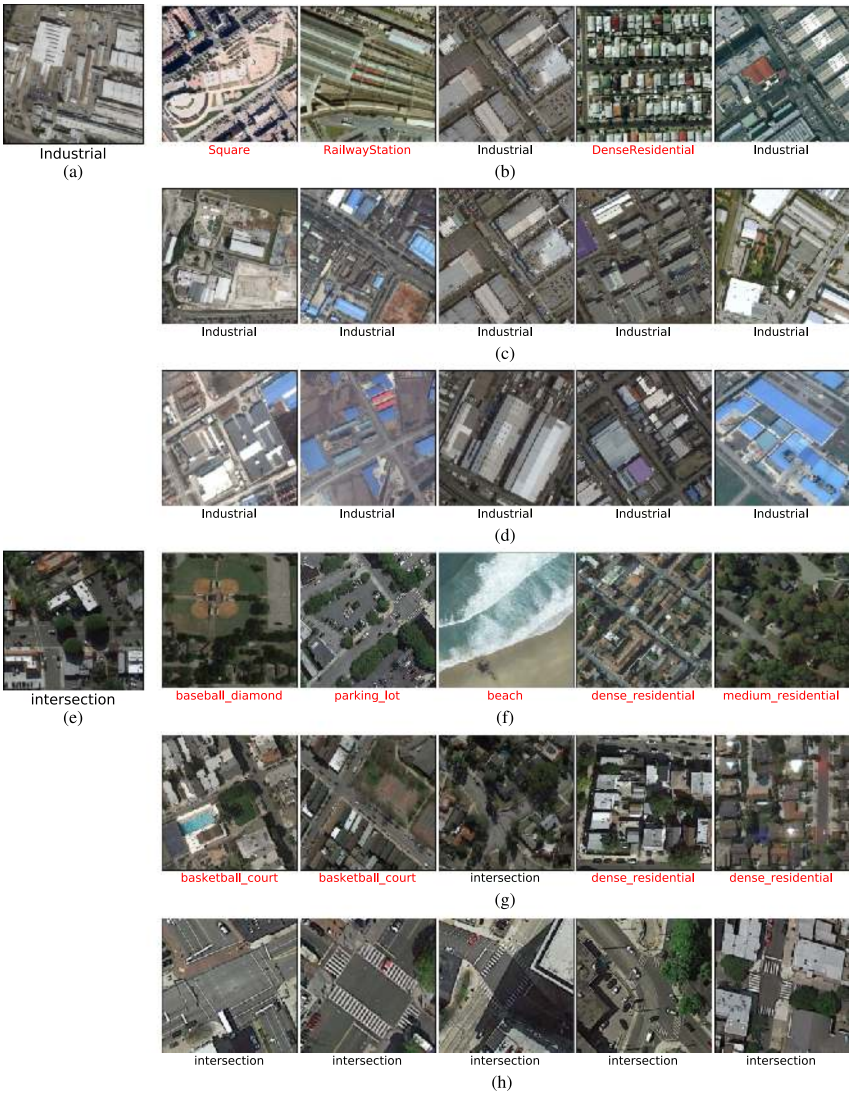

Fig. 5. Top five nearest neighbors retrieved from the training sets with respect to the query images based on SNCA, t-RNSL, and NTDNE. (a) and (e) are the query images from AID and NWPU-RESISC45, respectively. (b) and (f) are the retrieved results based on SNCA. (c) and (g) are the retrieved results based on t-RNSL. (d) and (h) are the retrieved results based on NTDNE.

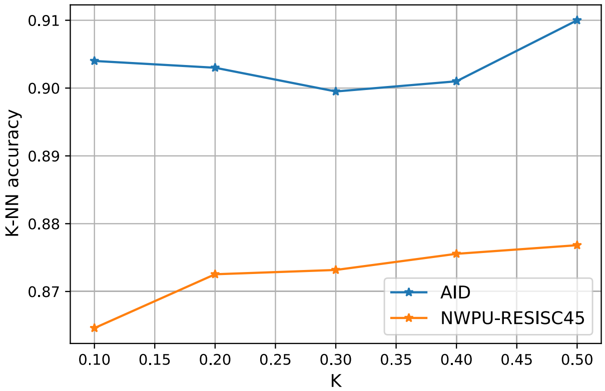

Fig. 6. Sensitivity analysis of hyperparameter k on two benchmark datasets when k varies from 0.1 to 0.5.

V. Conclusion

We propose a novel deep metric learning loss function for characterizing the semantic content of RS scenes in a robust manner. We stochastically maximize the leave-one-out K-NN score, which is inspired by SNCA, to model the inherent neighborhood structure of images in feature space. Considering the existence of label noise, we improve the robustness of the log-likelihood loss function of the leave-one-out K-NN score by replacing the logarithm function with the negative Box–Cox transformation, which can down-weight the contribution of potentially noisy images by learning their localized structure in feature space. In addition, we apply a truncation mechanism on the loss function to further improve the robustness capability, especially when a large percentage of the database contains noisy labels. Compared with several state-of-the-art metric learning losses, the proposed method exhibits superior performance on three downstream tasks, including K-NN classification, clustering, and image retrieval based on the features extracted the trained CNN models. In practice, the proposed method can be utilized for large-scale RS scene classification and retrieval tasks, without the need for accurate land-use or land-cover labels. As a future work, we will extend the proposed method to the multilabel case.

References

[1] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[2] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, “Orsim detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.

[3] 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.

[4] G.-S. Xia *et al.*, “DOTA: A. large-scale dataset for object detection in aerial images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[5] X. Sun, Y. Liu, Z. Yan, P. Wang, W. Diao, and K. Fu, “SRAF-Net: Shape robust anchor-free network for garbage dumps in remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3023928.

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

[7] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, “Learning-shared cross-modality representation using multispectral-Lidar and hyperspectral data,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1470–1474, Aug. 2020.

[8] R. Fernandez-Beltran, A. Plaza, J. Plaza, and F. Pla, “Hyperspectral unmixing based on dual-depth sparse probabilistic latent semantic analysis,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6344–6360, Nov. 2018.

[9] P. Duan, J. Lai, J. Kang, X. Kang, P. Ghamisi, and S. Li, “Texture-aware total variation-based removal of sun glint in hyperspectral images,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 359–372, 2020.

[10] J. Kang, D. Hong, J. Liu, G. Baier, N. Yokoya, and B. Demir, “Learning convolutional sparse coding on complex domain for interferometric phase restoration,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 826–840, Feb. 2021.

[11] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “An augmented linear mixing model to address spectral variability for hyperspectral unmixing,” *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[12] R. Huang, Y. Xu, and U. Stilla, “GraNet: Global relation-aware attentional network for ALS point cloud classification,” 2020, *arXiv:2012.13466*.

[13] Y. Huang, L. Zhang, J. Li, Z. Chen, and X. Yang, “Reweighted tensor factorization method for SAR narrowband and wideband interference mitigation using smoothing multiview tensor model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3298–3313, May 2020.

[14] R. Huang, Y. Xu, W. Yao, L. Hoegner, and U. Stilla, “Robust global registration of point clouds by closed-form solution in the frequency domain,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 171, pp. 310–329, 2020.

[15] Y. Xu, Z. Ye, R. Huang, L. Hoegner, and U. Stilla, “Robust segmentation and localization of structural planes from photogrammetric point clouds in construction sites,” *Autom. Construction*, vol. 117, 2020, Art. no. 103206.

[16] Y. Xu, R. Boerner, W. Yao, L. Hoegner, and U. Stilla, “Pairwise coarse registration of point clouds in urban scenes using voxel-based 4-planes congruent sets,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 106–123, 2019.

[17] N. Yokoya *et al.*, “Breaking limits of remote sensing by deep learning from simulated data for flood and debris-flow mapping,” *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3035469.

[18] B. Adriano *et al.*, “Learning from multimodal and multitemporal earth observation data for building damage mapping,” 2020, *arXiv:2009.06200*.

[19] 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.

[20] H. Li, H. Gu, Y. Han, and J. Yang, “Object-oriented classification of high-resolution remote sensing imagery based on an improved colour structure code and a support vector machine,” *Int. J. Remote Sens.*, vol. 31, no. 6, pp. 1453–1470, 2010.

[21] Y. Yang and S. Newsam, “Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery,” in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 1852–1855.

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

[23] M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, “Land-use classification with compressive sensing multifeature fusion,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2155–2159, Oct. 2015.

[24] D. Hong *et al.*, “More diverse means better: Multimodal deep learning meets remote sensing imagery classification,” *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3016820.

[25] D. Hong, L. Gao, J. Yao, B. Zhang, P. Antonio, and J. Chanussot, “Graph convolutional networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3015157.

[26] R. Huang, Y. Xu, D. Hong, W. Yao, P. Ghamisi, and U. Stilla, “Deep point embedding for urban classification using ALS point clouds: A new perspective from local to global,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 163, pp. 62–81, 2020.

[27] X. Sun, A. Shi, H. Huang, and H. Mayer, “Bas-Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images,” *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5398–5413, 2020, doi: 10.1109/JSTARS.2020.3021098.

[28] 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.

[29] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," in *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020, doi: 10.1109/JSTARS.2020.3005403.

[30] 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.

[31] 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.

[32] L. Yan, R. Zhu, N. Mo, and Y. Liu, "Cross-domain distance metric learning framework with limited target samples for scene classification of aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3840–3857, Jun. 2019.

[33] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.

[34] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *Proc. ISPRS TC VII Symp.*, 2010, vol. 38, pp. 298–303.

[35] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[36] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[37] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[38] Z. Dong *et al.*, "Registration of large-scale terrestrial laser scanner point clouds: A. review and benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 163, pp. 327–342, 2020.

[39] J. E. V. Munoz, S. Srivastava, D. Tuia, and A. X. Falcao, "Openstreetmap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geosci. Remote Sens. Mag.*, to be published.

[40] H. Li *et al.*, "RSI-CB: A large scale remote sensing image classification benchmark via crowdsource data," 2017, *arXiv:1705.10450*.

[41] Y. Long *et al.*, "DiRS: On creating benchmark datasets for remote sensing image interpretation," 2020, *arXiv:2006.12485*.

[42] J. Kang, R. Fernandez-Beltran, P. Duan, X. Kang, and A. Plaza, "Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3042607.

[43] Z. Wu, A. A. Efros, and S. X. Yu, "Improving generalization via scalable neighborhood component analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 685–701.

[44] 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.*, vol. 58, no. 12, pp. 8905–8918, Dec. 2020.

[45] G. E. Box and D. R. Cox, "An analysis of transformations," *J. Roy. Statist. Soc., Ser. B. (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[46] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8778–8788.

[47] X. Huang, L. Zhang, and L. Wang, "Evaluation of morphological texture features for mangrove forest mapping and species discrimination using multispectral IKONOS imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 3, pp. 393–397, Jul. 2009.

[48] Z. Li and L. Itti, "Saliency and gist features for target detection in satellite images," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2017–2029, Jul. 2011.

[49] G. Cheng *et al.*, "Object detection in remote sensing imagery using a discriminatively trained mixture model," *ISPRS J. Photogrammetry Remote Sens.*, vol. 85, pp. 32–43, 2013.

[50] C. Chen, B. Zhang, H. Su, W. Li, and L. Wang, "Land-Use scene classification using multi-scale completed local binary patterns," *Signal, Image Video Process.*, vol. 10, no. 4, pp. 745–752, 2016.

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

[52] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "Multimodal probabilistic latent semantic analysis for Sentinel-1 and sentinel-2 image fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1347–1351, Sep. 2018.

[53] E. Li, P. Du, A. Samat, Y. Meng, and M. Che, "Mid-level feature representation via sparse autoencoder for remotely sensed scene classification," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 1068–1081, Mar. 2017.

[54] 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.

[55] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 117–126, Jan. 2018.

[56] 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.

[57] M.-S. Yun, W.-J. Nam, and S.-W. Lee, "Coarse-to-fine deep metric learning for remote sensing image retrieval," *Remote Sens.*, vol. 12, no. 2, 2020, Art. no. 219.

[58] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 567–574.

[59] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," ICLR (Workshop). 2015.

[60] S. Sukhbaatar and R. Fergus, "Learning from noisy labels with deep neural networks," 2014, *arXiv:1406.2080*.

[61] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," 2017, *arXiv:1712.09482*.

[62] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 322–330.

[63] Y. Lyu and I. W. Tsang, "Curriculum loss: Robust learning and generalization against label corruption," 2019, *arXiv:1905.10045*.

[64] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," 2019, *arXiv:1912.05170*.

[65] H. Song, M. Kim, D. Park, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," 2020, *arXiv:2007.08199*.

[66] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," *3rd Int. Conf. Learn. Representations*, ICLR 2015. 2015.

[67] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 513–520.

[68] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[69] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3007029.

[70] C. D. Manning, H. Schütze, and P. Raghavan, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

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

[72] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[73] A. Zhai and H.-Y. Wu, "Classification is a strong baseline for deep metric learning," 2018, *arXiv:1811.12649*.

[74] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.

[75] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

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. From 2019 to 2020, he was with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin (TU Berlin), Berlin, Germany. He is currently with the School of Electronic and Information Engineering, Soochow University, Suzhou, China. His research interests include signal processing and machine learning techniques, 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. His collaborated work was selected as one of the 10 Student Paper Competition Finalists in IGARSS 2020.

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, Plana, Spain, (Castellon de la Plana, Spain) in 2007, 2011, and 2016, respectively.

He is currently a Postdoctoral 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., University of Cáceres, Cáceres, Spain, and Technische Universität Berlin, Berlin, Germany. His research interests include multimedia retrieval, spatio-spectral 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. He is member of the Spanish Association for Pattern Recognition and Image Analysis (AERFAI), which is part of the International Association for Pattern Recognition (IAPR).

Xudong Kang (Senior Member, IEEE) received the B.Sc. degree from Northeast University, Shenyang, China, in 2007, and the Ph.D. degree from Hunan University, Changsha, China, in 2015.

In 2015, he joined the College of Electrical Engineering, Hunan University, Changsha, China. His research interests include hyperspectral feature extraction, image classification, image fusion, and anomaly detection.

Dr. Kang was awarded the Second Prize in the Student Paper Competition in IGARSS 2014. In IGARSS 2017, he was selected as the Best Reviewer for *IEEE Geoscience and Remote Sensing Letters*, in 2016.

Jingen Ni (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2004 and the Ph.D. degree in circuits and systems from Fudan University, Shanghai, China, in 2011.

Since February 2011, he has been with the School of Electronic and Information Engineering, Soochow University, Suzhou, China, where he is currently a Professor. From November 2014 to November 2015, he was also a Visiting Assistant Professor with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pok Fu Lam, Hong Kong. His research interests include adaptive signal processing, distributed optimization and learning, remote sensing image processing, and artificial neural networks.

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

He is currently 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 around 300 JCR journal articles (over 170 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 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 Top. in Applied Earth Observations and Remote Sensing (JSTARS)*. 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 was also 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 Top. in Applied Earth Observations and Remote Sensing (JSTARS)*, and the Best Column Award of the *IEEE Signal Processing Magazine* in 2015. He received Best Paper Awards at the IEEE International Conference on Space Technology and the IEEE Symposium on Signal Processing and Information Technology. 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 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 of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING from 2013 to 2017. He has guest edited ten special issues on hyperspectral remote sensing for different journals. He is also an Associate Editor of *IEEE ACCESS* (received the recognition as an Outstanding Associate Editor of the journal in 2017). He is currently serving as Editor-in-Chief of the *IEEE Journal on Miniaturization for Air and Space Systems*. He has been included in the Highly Cited Researchers list from Clarivate Analytics in 2018, 2019, and 2020. For more information please visit: http://www.umbc.edu/rssipl/people/aplaza