Robust Normalized Softmax Loss for Deep Metric Learning-based Characterization of Remote Sensing Images with Label Noise

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

Abstract

Most deep metric learning-based image characterization methods exploit supervised information to model the semantic relations among the remote sensing (RS) scenes. Nonetheless, the unprecedented availability of large-scale RS data makes the annotation of such images very challenging, requiring automated supportive processes. Whether the annotation is assisted by aggregation or crowd-sourcing, the RS large-variance problem, together with other important factors [e.g. geo-location/registration errors, land-cover changes, even low-quality Volunteered Geographic Information (VGI), etc.] often introduce so-called label noise, i.e. semantic annotation errors. In this paper, we first investigate the

This work was supported by the Major Program of the National Natural Science Foundation of China (No. 61890962), the National Natural Science Foundation of China (No. 61871179), the National Natural Science Fund of China for International Cooperation and Exchanges (No. 61520106001), the Natural Scince Foundation of Hunan Province (2019JJ50036), the Fund of Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province (No. 2018TP1013), the Ministry of Science, Innovation and Universities of Spain (RTI2018-098651-B-C54), the Valencian Government of Spain (GV/2020/167), FEDER-Junta de Extremadura (Ref. GR18060) and the European Union under the H2020 EOXPOSURE project (No. 734541). *(Corresponding author: Xudong Kang.)*

J. Kang is with the School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China (e-mail: kangjian_1991@outlook.com).

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

P. Duan and X. Kang are with the College of Electrical and Information Engineering, Hunan University, 410082 Changsha, China. (e-mail: puhong_duan@hnu.edu.cn; xudong_kang@163.com).

A. 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).

deep metric learning-based characterization of RS images with label noise and propose a novel loss formulation, named Robust Normalized Softmax Loss (RNSL), for robustly learning the metrics among RS scenes. Specifically, our RNSL improves the robustness of the Normalized Softmax Loss (NSL), commonly utilized for deep metric learning, by replacing its logarithm function with the negative Box-Cox transformation in order to down-weight the contributions from noisy images on the learning of the corresponding class prototypes. Moreover, by truncating the loss with a certain threshold, we also propose a truncated Robust Normalized Softmax Loss (t-RNSL) which can further enforce the learning of class prototypes based on the image features with high similarities between them, so that the intra-class features can be well grouped and inter-class features can be well separated. Our experiments, conducted on two benchmark RS datasets, validate the effectiveness of the proposed approach with respect to different state-of-the-art methods in three different downstream applications (classification, clustering and retrieval). The codes of this paper will be publicly available from https://github.com/jiankang1991.

Index Terms

Label noise, deep metric learning, image characterization, remote sensing, image retrieval.

I. Introduction

In recent years, the fast development of satellite sensor technology has created great opportunities to exploit remote sensing (RS) data in a wide range of important applications, such as target identification [1]–[6], land-cover analysis [7]–[14], ecosystem monitoring [15]–[18], agriculture [19]–[21] and demographics [22], [23]. In these (and many other relevant tasks [24]), the adequate characterization of RS scenes plays a key role for the semantic recognition of objects as well as the characterization of their spatial topology, due to the particular complexity of the RS image domain [25], [26].

In the literature, a large variety of methods have been developed to effectively characterize RS scenes and classify or retrieve their visual content in a satisfactory way [27], [28]. Among the most popular approaches, it is possible to find handcrafted feature-based methods [29]–[31], unsupervised characterization models [32]–[34] and deep learning-based techniques [35]–[38]. Regarding traditional handcrafted and unsupervised methods, these approaches typically exhibit limited performance within the RS field, owing to the inherent constraints of low-level descriptors and unlabeled data [39]. In contrast, the great potential of convolutional neural networks (CNN) to uncover highly discriminating features from RS scenes makes deep learning one of the most prominent and successful trends [40]. Specifically, the so-called deep metric learning approach has recently shown excellent results to characterize complex RS data [41]–[46]. In general, deep metric learning pursues to project semantically similar images to nearby positions in the corresponding CNN-based metric space while separating dissimilar images according to their semantic annotations. Consequently, these techniques generally require vast amounts of supervised data for properly learning the complex semantic relationships associated to aerial scenes [42].

With the proliferation of different Earth Observation missions, the big data era of RS is a present-day reality [47]–[50]. However, the annotation of massive data becomes a major challenge in RS, since manually generating relevant ground-truth information is very expensive and time-consuming, which often makes the process unaffordable from an operational perspective, and constrains the availability of labeled RS data for deep learning-based applications. In order to relieve this problem, two main annotation strategies have been adopted in the RS field: (i) aggregation and (ii) crowd-sourcing. On the one hand, aggregation techniques [51], [52] make use of some sort of unsupervised methodology to group the data into a limited number of clusters. Then, each group of samples is manually annotated with the same semantic labels to reduce the final effort in processing the whole archive. On the other hand, crowd-sourcing methods [53], [54] take advantage of the geospatial semantic information available in different crowd-sourcing platforms [e.g. Google Maps, OpenStreetMap (OSM), CORINE Land Cover (CLC), etc.] to automatically generate the corresponding semantic annotations for the aerial scenes according to their geographic coordinates [55].

Whether the RS data are labeled using aggregation or crowd-sourcing, both procedures inevitably introduce label errors due to the RS large-variance problem, as well as other important factors. That is, the high intra-class and low inter-class variability inherent to RS data [25] may produce that semantically similar scenes could be differently grouped and consequently mislabeled. Besides, additional factors, such as geo-location/registration errors, land-cover changes or even low-quality Volunteered Geographic Information (VGI), could also introduce label noise, i.e. scenes which are annotated with semantic labels different to the real ones. Precisely, all these deviations may potentially degenerate any supervised RS image characterization scheme, being deep metric learning no exception [42]. In fact, the frequent complexity of deep metric learning models (with many hyper-parameters) makes them particularly prone to suffer degradation by such noisy labels [56]. Although some efforts have been recently carried out in the literature to improve deep learning-based classifiers for some computer vision [57]–[59] and RS [60] applications, the lack of research on error-tolerant characterization methods within the RS field motivates the development of new deep metric learning models to effectively characterize complex RS archives with label noise.

In order to overcome these limitations, this paper proposes a new RS image characterization approach, based on the newly defined Robust Normalized Softmax Loss (RNSL), which has been specially developed to deal with RS scenes with noisy annotations. Specifically, we first investigate the general application of deep metric learning to characterize airborne and spaceborne images with label noise. After analyzing the state-of-the-art, we formulate the proposed RNSL by revisiting the Normalized Softmax Loss (NSL) based on the negative Box-Cox transformation [61], with the objective of naturally reducing the contribution of those images with potential label noise when learning the corresponding class prototypes. Additionally, we also define an extension of the proposed loss, named the truncated Robust Normalized Softmax Loss (t-RNSL), to enforce learning the class prototypes based on the image features with higher similarities. In this way, intra-class feature variations can be further reduced and inter-class features can be better separated in the resulting embedding space. To demonstrate the effectiveness of our contributions, we perform an extensive experimental comparison, involving two different RS benchmark archives and three downstream applications (K-NN classification, clustering and retrieval), that confirms the advantages of the proposed approach to characterize RS data with noisy annotations with respect to different state-of-the-art deep metric learning methods. Summarizing, the contributions of this paper can be condensed into the following points:

1) To our best knowledge, we investigate for the first time in the literature the problem of deep metric learning-based RS image characterization with label noise, exposing that noisy annotations may have a high impact on state-of-the-art losses when characterizing RS scenes.
2) We propose a new loss function (RNSL) and its truncated extension (t-RNSL) for increasing the noise-tolerant capability of the RS image characterization framework based on deep metric learning.
3) We widely explore how the presented approach performs in different downstream applications (K-NN classification, clustering and retrieval) over several benchmark RS archives with label noise. This provides important insights about the working mechanism and advantages of the proposed losses with respect to other state-of-the-art functions, especially under heavy uniform label noise.

The reminder of this paper is structured as follows. Section II describes some related works while highlighting the novelty of this work. Section III presents our deep metric learning-based RS image characterization scheme, including the two newly proposed loss functions. Section IV reports the conducted experimental comparison and discusses the results. Finally, Section V concludes the work and provides some hints at plausible future research lines.

II. Related Work

A. RS Image Characterization

Generally, existing RS scene characterization methods can be categorized into three different types depending on the nature of the considered features [28]: (1) handcrafted; (2) unsupervised; and (3) deep learning-based. In handcrafted feature-based methods, low-level visual descriptors are extracted to represent RS images according to different elementary features, such as color, texture, shape, etc. For instance, some of the most popular descriptors used in RS are color histograms, local binary patterns and scale-invariant feature transform (SIFT) [29]–[31]. Despite their advantages, these straightforward methods are often unable to provide satisfactory results to characterize airborne and space-borne optical data owing to the high semantic complexity of the RS image domain [25]. To improve the generalization capability, unsupervised methods make use of different unsupervised learning paradigms to encode the extracted features into a higher-level feature space. Among the most representative techniques used in RS, we can find sparse coding, topic modeling and auto-encoders [32]–[34]. Despite the positive results achieved by these and other unsupervised alternatives, the lack of supervised information generally reduces their intra-class discrimination ability which may eventually becomes an important limitation to deal with the large-variance problem in RS [39].

The recent development of deep learning technologies has attracted the attention of the RS research community owing to the excellent capabilities of CNNs to extract highly discriminating features from visual data [40]. In particular, the objective of these models is based on projecting the input data onto its corresponding label space using multiple nonlinear mappings and layers that are able to produce high-level characterizations very useful in RS [62]. For instance, it is the case of Li *et al.* who present in [63] a RS image classification approach which integrates multi-layer features of different pre-trained CNN models for characterizing the aerial scenes. Liu *et al.* also develop in [64] an image characterization method for RS that exploits multi-scale CNN features based on an spatial pyramid pooling. Similarly, Zheng *et al.* propose in [65] a deep scene representation technique that makes use of pre-trained CNN features, multi-scale pooling and Fisher vectors to characterize RS images.

Notwithstanding the good performances of these and other related approaches [66], the so-called deep metric learning scheme has recently become a prominent trend to effectively represent RS scenes. Deep metric learning pursues to project semantically similar images to nearby locations in the resulting CNN-based characterization space. As a result, this framework becomes highly suitable for modeling the complex semantic relationships inherent to large-scale variance RS data [28]. In general, it is possible to categorize most of the existing deep metric learning methods based on two formulations: (i) the contrastive loss; and (ii) the triplet loss. On the one hand, the contrastive embedding [67] is trained with pared data to minimize the distance between the two samples if they share the same class, and to increase such distance (by a certain margin) if they belong to different classes. On the other hand, the triplet loss [68] considers triplets of samples (i.e. anchor, positive and negative) with the objective of minimizing the distance between the anchor and its positive exemplar and also pushing the negative sample away from the anchor by a certain margin. Different works in the RS literature exemplify these two alternatives. For instance, Cheng *et al.* develop in [42] the discriminative CNN (D-CNN), which imposes a contrastive-based metric learning regularization over an off-the-shelf CNN architecture. Following a similar inspiration, Yan *et al.* present in [43] a cross-domain adaptation based on hybrid color features to reduce the bias of the data distribution in the resulting embedding space. In [69], Cao *et al.* make use of the triplet loss formulation to define a content-based RS image retrieval framework, that considers both positive and negative aerial scenes when generating the embedding space.

B. Deep Metric Learning with Label Noise

In spite of the advantages of the deep metric learning scheme to characterize RS scenes, the task of sampling informative pairs or triplets from large-scale RS archives becomes very challenging because the probability of considering samples with inconsistent semantic annotations and relationships is logically affected by the data volume and noise [42]. Note that both factors (the data size and the presence of label noise) are important problems in RS due to the unprecedented availability of massive RS data, together with the operational limitations of annotating such large-scale image archives [47], [48]. Precisely, different strategies have been developed in the literature to alleviate these problems. For example, it is the case of the scalable neighborhood component analysis (SNCA) [70]. Specifically, this approach is built upon the neighborhood component analysis (NCA) [71] by including an augmented non-parametric memory for training the CNN model with a larger data scope to produce more general and robust features. In [72], Zhai *et al.* propose the normalized softmax loss (NSL) which considers a classification-based metric learning approach to characterize and retrieve images by content. In more details, NSL pursues to maximize the agreement between the corresponding class prototypes and the associated features of the same class in order to make the most challenging samples more relevant during training, which may certainly enhance the generalization capability of the model. Additionally, Deng *et al.* present in [73] the additive angular margin (ArcFace) to maximize the class separability while producing highly discriminating image features. In particular, ArcFace uses the arc-cosine function with an additive angular margin to optimize the distance between features and target weights with the objective of estabilising the training process under the most complicated situations, e.g intra-class large-variance and label noise. In [74], Yuan *et al.* also define an alternative distance metric based on the signal-to-noise ratio (SNR) to improve the discrimination ability and feature robustness. Despite all the conducted research, there is a lack of deep metric learning methods specifically designed for characterizing large-scale RS archives with label noise, which motivates the development of new error-tolerant models to account for the particular complexity of the RS image domain (where instrument types, sensing positions or atmospheric effects may also generate important semantic deviations [28]).

C. Novelty of our Work

In order to face these challenges, this paper presents a novel deep metric learning scene characterization method which is particularly designed to deal with large-scale RS archives with noisy annotations. Whereas state-of-the-art deep metric learning models try to improve their robustness and generalization capability by exploiting data diversity and separability (e.g. [42], [68], [70], [72], [73]), even a few mislabeled RS scenes on class boundaries may have a strong impact on the final embedding space, since some decision boundaries could be easily modified by possible noisy label fluctuations. Note that the large-scale nature and inherent semantic complexity of RS data make the problem of label noise particularly challenging over these transitional regions. In this context, the proposed approach aims at reducing the negative effect of label noise over such critical regions by means of two novel loss functions: RNSL and its truncated extension t-RNSL. More specifically, we take advantage of the negative Box-Cox transformation [61] to enforce normal and symmetry conditions that allow the proposed RNSL loss to reduce the contribution of samples belonging to semantically uncertain regions that are particularly affected by noisy labels. Additionally, t-RNSL further improves the model robustness to label noise by thresholding the contribution of potentially mislabeled RS scenes.

Unlike other works available in the literature, the proposed approach gathers two important facets for the RS domain in an innovative manner: data scalability and label noise. On the one hand, we present a RS image characterization method which is built upon the rationale of the NSL function, owing to its prominent generalization capabilities with large-scale data [72], which is certainly a key factor in RS [47], [48]. On the other hand, we formulate two novel loss functions (RNSL and t-RNSL) in order to account for label noise when learning the corresponding RS image embeddings, since the presence of erroneous annotations is an important problem in the most challenging RS archives. In contrast to NSL [72], the proposed approach has been developed assuming that existing labels can be corrupted by noise. Hence, we integrate an effective mechanism to control the contribution of such noise to the gradient update with the aim of not misleading the process of learning the corresponding class prototypes. In other words, we reformulate the standard normalized version of the soft-max loss in order to deal with the particular requirements of RS image collections with label noise. When compared to different state-of-the-art techniques, the presented RS image characterization model is able to provide remarkable performance improvements with respect to the methods in [42], [68], [70], [72], [73], [75], which also indicates the novelty and advantages of the proposed approach when dealing with label noise.

III. Robust Normalized Softmax Loss

The proposed deep metric learning method for characterizing RS scenes with noisy labels mainly contains two parts: 1) a backbone CNN architecture for encoding the RS images into the associated features of a low-dimensional metric space; 2) a new loss function (RNSL) and its truncated extension (t-RNSL) for robustly learning the distance metrics of the RS images with label noise. Figure 1 provides a graphical illustration of the proposed framework, where deep features, class prototypes and noisy labels are involved in the defined loss formulation. As follows, we describe all the framework details, including the considered notations (Section III-A), a technical analysis of NSL (Section III-B) and the proposed losses (Section III-C).

A. Notations

Let $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be an RS image dataset consisting N images with category labels, and $\mathcal{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$ 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 c-th class, the c-th element of $\mathbf{y}_i$ is 1, i.e., $y_i^c = 1$, and the other elements are 0. In the context of deep metric learning, we denote $\mathcal{F}(\cdot)$ as the CNN model which encodes the input image $\mathbf{x}_i$ into a low-dimensional feature $\mathbf{f}_i \in \mathbb{R}^D$ with the dimension size of D. In this paper, the features are normalized, i.e., $\mathbf{f}_i = \mathcal{F}(\mathbf{x}_i)/\|\mathcal{F}(\mathbf{x}_i)\|_2$. Considering the label noise, we denote the noisy label set by $\hat{\mathcal{Y}} = \{\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_N\}$, where $\hat{\mathbf{y}}_i$ represents the noisy label vector. Here, we also assume that the noise is conditionally independent of input images given the true labels [76]:

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

where η_{ck} describes the noise rate, drawn as the (c, k)-th component from a $C \times C$ probability transition matrix $\mathbf{Q}$ [77]. Two different kinds of noise are considered in this paper, including *uniform* noise, where a true label is randomly flipped into other labels with equal probability $\eta_{ck} = \frac{\eta}{C-1}$ or preserves as the true label with the probability $\eta_{ck} = 1 - \eta$, and *label-dependent* noise, where a true label is more likely to be mistakenly labeled with a particular class with the probability $\eta_{ck} = \eta$ or preserves as the true label with the probability $\eta_{ck} = 1 - \eta$. For example, Figure 2 illustrates two kinds of noise by a RS scene of *farmland*. For uniform noise, *farmland* is randomly mislabeled by other classes with equal probability, such as *bareland*, *forest*, etc. For label-dependent noise, it is mislabeled by *forest* with a certain probability, since their semantic contents are similar. Deep metric learning methods are developed for learning a CNN model $\mathcal{F}(\cdot)$ to effectively encode the semantic contents of images by the features $\mathbf{f}$ in the embedding space, where semantically similar images are close and dissimilar images are separated. With the label noise existing in RS image archives, such semantic relationships among the scenes should be robustly uncovered via $\mathcal{F}(\cdot)$.

B. Details on NSL

One of the state-of-the-art methods for deep metric learning is based on NSL [72], which is formally described as:

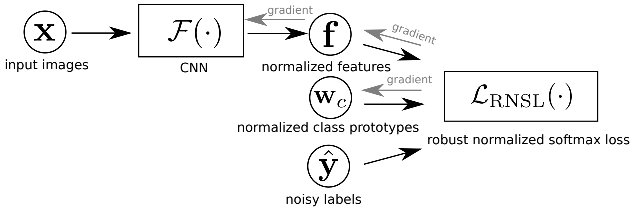


Fig. 1. A graphical illustration of the proposed framework. Given the training data with label noise, we exploit the proposed robust losses to learn class prototypes w_c and the noise-tolerant CNN models for extracting the image features.

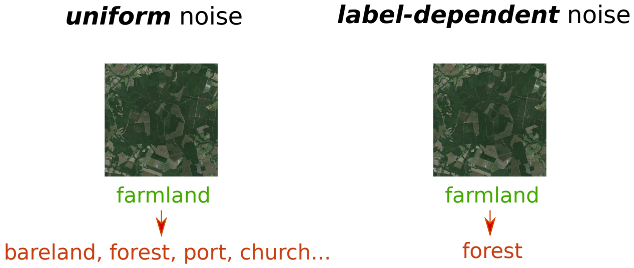


Fig. 2. Two types of label noise are considered in this paper. (Left) Uniform noise: a true label is flipped into other labels with equal probability. (Right) Label-dependent noise: a true label is more likely to be mistakenly labeled with a particular class.

$$\begin{aligned}\mathcal{L}_{\mathrm{NSL}} &= -\frac{1}{N}\sum_i\sum_c y_i^c \log\Big(\frac{\exp(\mathbf{w}_c^T\mathcal{F}(\mathbf{x}_i)/\sigma)}{\sum_k \exp(\mathbf{w}_k^T\mathcal{F}(\mathbf{x}_i)/\sigma)}\Big)\\ &= -\frac{1}{N}\sum_i\sum_c y_i^c(\mathbf{w}_c^T\mathcal{F}(\mathbf{x}_i)/\sigma)\\ &+\frac{1}{N}\sum_i\sum_c y_i^c \log\Big(\sum_k \exp(\mathbf{w}_k^T\mathcal{F}(\mathbf{x}_i)/\sigma)\Big),\end{aligned} \tag{2}$$

where $\mathbf{w}_c \in \mathbb{R}^D$ denotes the normalized weight vector of the class c, (i.e., $\|\mathbf{w}_c\|_2 = 1$), and σ is the temperature parameter that controls the concentration of the sample distribution. According to $-\frac{1}{N}\sum_i\sum_c y_i^c(\mathbf{w}_c^T\mathcal{F}(\mathbf{x}_i)/\sigma)$, minimizing $\mathcal{L}_{\mathrm{NSL}}$ given the true labels tends to maximize the agreement between $\mathbf{w}_c$ and the associated features of the same class. Therefore, $\mathbf{w}_c$ is often termed as *class prototype*. Let p_i^c represent the probability that the feature $\mathcal{F}(\mathbf{x}_i)$ is aligned with the c-th class prototype among all the others, (i.e., $p_i^c = \frac{\exp(\mathbf{w}_c^T\mathcal{F}(\mathbf{x}_i)/\sigma)}{\sum_k \exp(\mathbf{w}_k^T\mathcal{F}(\mathbf{x}_i)/\sigma)}$). By calculating the gradient of $\mathcal{L}_{\mathrm{NSL}}$ with respect to $\mathbf{w}_c$, we can obtain:

$$\frac{\partial\mathcal{L}_{\mathrm{NSL}}}{\partial\mathbf{w}_c} = -\frac{1}{N}\sum_i\sum_c \frac{y_i^c}{p_i^c}\frac{\partial p_i^c}{\partial\mathbf{w}_c}. \tag{3}$$

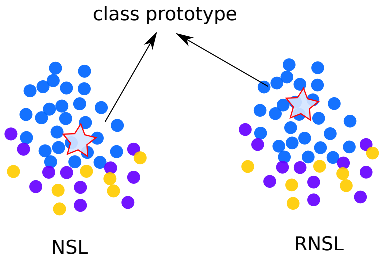
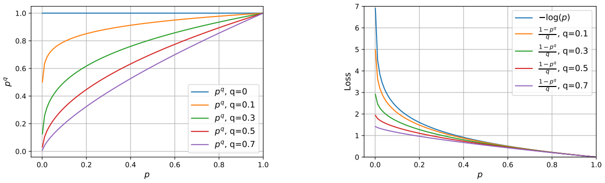


Fig. 3. An illustration of the different learning effects on the class prototypes based on NSL and RNSL.



(a) (b)

Fig. 4. Function plots of p^q in (a), $-\log(p)$ and $\frac{1-p^q}{q}$ in (b).

For the images belonging to the c-th class, when their features are not well aligned with respect to $\mathbf{w}_c$, larger weights $\frac{y_i^c}{p_i^c}$ are enforced on the term of $\frac{\partial p_i^c}{\partial \mathbf{w}_c}$. In other words, during the learning phase of $\mathbf{w}_c$, hard images receive more attention than the images that can be easily discriminated, and they contribute more on the gradient update of $\mathbf{w}_c$. However, when the true labels are corrupted by noise, the wrongly categorized images can dominate the contribution of the gradient update, which will mislead the learning of the associated class prototype. As illustrated in Figure 3, the produced class prototype may be closer to the features of the wrongly categorized images in the feature space. This will also limit the learning of the CNN models which are utilized for generating the features for the input RS images, since they can be positioned by the models towards the features from different classes in the feature space.

C. RNSL

To solve these problems, we propose a novel loss formulation for deep metric learning which is more robust for images with noisy labels. Inspired by the results achieved in other domains [76], we utilize the negative Box-Cox transformation [61] due to its normal and symmetrical properties as the loss function with the expression:

$$\mathcal{L}_{\mathrm{RNSL}} = \frac{1}{N}\sum_i \sum_c y_i^c \frac{\big(1-(p_i^c)^q\big)}{q}, \quad q \in (0,1). \tag{4}$$

Differently with respect to the generalized cross entropy (GCE) proposed in [76], RNSL has the capability of robustly learning the distance metrics of the images by enforcing the alignment between the class-wise prototypes and the associated image features in the feature space. By calculating the gradient of $\mathcal{L}_{\mathrm{RNSL}}$ with respect to $\mathbf{w}_c$, we can obtain:

$$\frac{\partial \mathcal{L}_{\mathrm{RNSL}}}{\partial \mathbf{w}_c} = \frac{1}{N}\sum_i \sum_c y_i^c (p_i^c)^q \big(-\frac{1}{p_i^c}\frac{\partial p_i^c}{\partial \mathbf{w}_c}\big). \tag{5}$$

Due to $p_i^c \in [0,1]$ and $q \in (0,1)$, $(p_i^c)^q$ has a down-weighting effect on $-\frac{1}{p_i^c}\frac{\partial p_i^c}{\partial \mathbf{w}_c}$ for each image. We plot the values of p^q with respect to the variations of p, when $q = 0.1, 0.3, 0.5, 0.7$, in Figure 4(a). When p decreases from 1 to 0, the decreasing speed of p^q becomes faster. In other words, for the images with noisy labels, smaller weights $(p_i^c)^q$ are imposed on $-\frac{1}{p_i^c}\frac{\partial p_i^c}{\partial \mathbf{w}_c}$ compared with the other images. Thus, the optimization of $\mathbf{w}_c$ is more dependent on the gradients calculated on the images with the truth labels than those with noisy labels. From the loss function perspective, we display the values of $\frac{1-p^q}{q}$ ($q = 0.1, 0.3, 0.5, 0.7$) and $-\log(p)$ with respect to the different values of p, in Figure 4(b). Compared to the loss function $-\log(p)$ utilized in NSL, $\frac{1-p^q}{q}$ of $\mathcal{L}_{\mathrm{RNSL}}$ puts less emphasis on the smaller values of p. Therefore, minimizing the loss values contributed from the images with noisy labels will not give more performance gain, which will improve the robustness for learning $\mathbf{w}_c$ and $\mathcal{F}(\cdot)$ against label noise.

Lemma 1. $\lim_{q\to 0} \mathcal{L}_{\mathrm{RNSL}} = \mathcal{L}_{\mathrm{NSL}}$.

Proof. Based on L'Hôpital's rule, we have [76]:

$$\begin{aligned}\lim_{q\to 0}\frac{1}{N}\sum_i\sum_c y_i^c \frac{1-(p_i^c)^q}{q} &= \frac{1}{N}\sum_i\sum_c y_i^c \lim_{q\to 0}\frac{\frac{d}{dq}\big(1-(p_i^c)^q\big)}{\frac{d}{dq}q}\\ &= \frac{1}{N}\sum_i\sum_c y_i^c \lim_{q\to 0} -(p_i^c)^q \log(p_i^c)\\ &= -\frac{1}{N}\sum_i\sum_c y_i^c \log(p_i^c)\end{aligned} \tag{6}$$

As it can be observed, the smaller the value of q will introduce higher approximation of $\mathcal{L}_{\mathrm{RNSL}}$ with respect to $\mathcal{L}_{\mathrm{NSL}}$. Therefore, we can consider the proposed loss function $\mathcal{L}_{\mathrm{RNSL}}$ as the robust version of NSL against label noise for deep metric learning.

Although the loss values of $\mathcal{L}_{\mathrm{RNSL}}$ with respect to small values of p_i^c are suppressed, the images with noisy labels still contribute to the learning of class prototypes $\mathbf{w}_c$. In order to further improve the robustness capability of $\mathcal{L}_{\mathrm{RNSL}}$, it is better to avoid the updating of $\mathbf{w}_c$ influenced by the gradient directions produced by the images with noisy labels. To achieve this goal, a truncated version of $\mathcal{L}_{\mathrm{RNSL}}$ is also proposed:

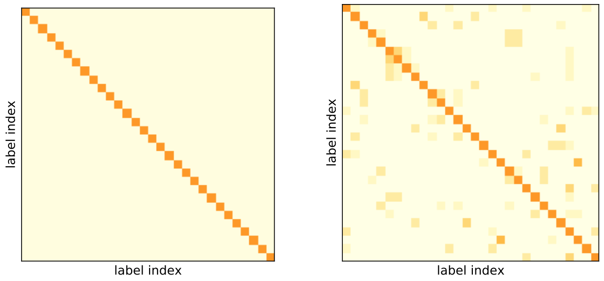
$$\mathcal{L}_{\mathrm{t-RNSL}} = \frac{1}{N}\sum_i\sum_c y_i^c \begin{cases} \frac{1-k^q}{q}, & \text{if } p_i^c \leq k \\ \frac{1-(p_i^c)^q}{q}, & \text{if } p_i^c > k \end{cases} \tag{7}$$

where $k \in (0,1)$ represents a threshold value. When p_i^c is less than a certain threshold k, the loss induced by p_i^c is cut to a certain value. This will lead to the zero gradients of $\mathcal{L}_{\mathrm{t-RNSL}}$ with respect to $\mathbf{w}_c$. Thus, the associated p_i^c cannot make any contributions to the learning of $\mathbf{w}_c$. In other words, the images with noisy labels have a great potential to be ”thrown out” during the learning phase based on $\mathcal{L}_{\mathrm{t-RNSL}}$. To practically optimize the $\mathcal{L}_{\mathrm{t-RNSL}}$, it can be further formulated as follows:

$$\begin{split}\mathcal{L}_{\mathrm{t-RNSL}} = \frac{1}{N}\sum_i\sum_c y_i^c\big([p_i^c > k]\frac{1-(p_i^c)^q}{q} \\ + (1-[p_i^c > k])\frac{1-k^q}{q}\big),\end{split} \tag{8}$$

where $[\cdot]$ denotes the Iverson bracket function, which takes 1 for those input values that make the argument statement true and 0 otherwise. In practice, one can store an indicator array representing whether $p_i^c > k$ is triggered for each image $\mathbf{x}_i$. However, at the initial state of CNN models (less than a certain number of training epochs), the generated features from the images cannot be discriminative enough to be exploited for ”filtering out” some plausible images with noisy labels. In this regard, the training of CNN models based on $\mathcal{L}_{\mathrm{t-RNSL}}$ is conducted with the following two steps:

- Within the first T epochs, the CNN models are trained with $\mathcal{L}_{\mathrm{RNSL}}$.
- After T epochs, the loss function is switched to $\mathcal{L}_{\mathrm{t-RNSL}}$.



(a)

(b)

Fig. 5. The probability transition matrix $\mathbf{Q}$ for the uniform (a) and label-dependent (b) noise based on the AID labels.

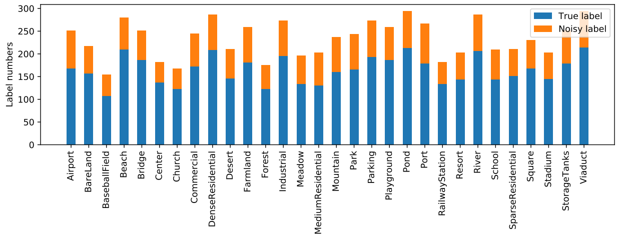


Fig. 6. Class-wise numbers of true and noisy labels on the AID dataset when $\eta = 0.3$.

The first step is to utilize $\mathcal{L}_{\mathrm{RNSL}}$ to train the CNN models until a state that the generated features can be discriminative, and then $\mathcal{L}_{\mathrm{t-RNSL}}$ takes action to prune some hard images for fine-tuning the state of CNN models which may not be affected by some noisy images.

IV. Experiments

A. Experimental Setup

We conduct extensive experiments based on two RS benchmark datasets including: 1) Aerial Image Dataset (AID)[1] [78] and 2) NWPU-RESISC45[2] [28]. We specifically select these two collections because they are both complex RS image archives in terms of data volume and semantic complexity, that became particularly challenging while reliable under label noise. For details about the datasets, we refer the readers to the associated papers. We randomly split the datasets into training, validation and test sets with percentages of 70%, 10%, and 20%,

[1] https://captain-whu.github.io/AID/

[2] http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html

respectively. For the training sets, the associated labels are corrupted by the uniform and label-dependent noises with the noise rate η equals to $0.1, 0.3, 0.5, 0.7$. For example, when $\eta = 0.5$, we plot the probability transition matrix $\mathbf{Q}$ for the uniform and label-dependent noise based on the AID labels in Figure 5(a) and Figure 5(b), respectively. For uniform noise, the probability that the original labels are preserved is $1 - \eta$, and they are randomly changed to other labels with the equal probability $\frac{\eta}{C-1}$. For label-dependent noise, we preserve the original labels with a probability of $1 - \eta$. Then, each noisy label is flipped into another class according to the probability transition matrices detailed in the appendix section. Note that, in the case of label-dependent noise, we design $\mathbf{Q}$ in such a way that noisy annotations are based on the semantic similarities among the land-use or land-cover classes, in order to make the mislabeling process as realistic as possible. As an example, Figure 5(b) shows the probability transition matrix for AID when $\eta = 0.5$. Besides, Figure 6 demonstrates the number of class-wise true and noisy labels of the AID dataset when the uniform noise exists with $\eta = 0.3$. The validation and test sets are exploited for the evaluation in the training and test phases, respectively. To evaluate the performance of the proposed method on the generation of distance metrics among RS images, we carry out several downstream tasks including: 1) K-NN classification; 2) clustering; and 3) image retrieval.

1) KNN classification: For the test images, their labels can be determined by a majority voting based on their K nearest neighbors retrieved from the training set via the measurement of the associated Euclidean distances in the feature space. Note that, in this evaluation phase, the truth labels of the training set are exploited for K-NN classification. We evaluate the classification performance based on the overall accuracy.

2) Clustering: We apply K-means clustering on the extracted features from the test images. Then, the clustering results are evaluated by normalized mutual information (NMI) [79] and unsupervised clustering accuracy (ACC) described as follows:

$$\mathrm{NMI} = \frac{2 \times I(\mathbf{Y};\mathbf{C})}{H(\mathbf{Y}) + H(\mathbf{C})}, \tag{9}$$

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.

$$\mathrm{ACC} = \max_{\mathcal{M}} \frac{\sum_{i=1}^{N} \delta(l_i = \mathcal{M}(c_i))}{N}, \tag{10}$$

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 cluster assigned labels and the ground-truth labels. These two metrics are utilized for measuring the discrimination of the generated features in the feature space.

3) Image retrieval: Image retrieval aims to accurately and effectively find the most semantically-similar images in a database given the query images based on their similarities of features. Such similarities are often measured by the Euclidean distance in the 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:

$$\mathrm{AP} = \frac{1}{Q} \sum_{r=1}^{R} P(r)\delta(r), \tag{11}$$

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.

For image retrieval, the test sets are exploited for query, and the training sets are the databases to be retrieved. The proposed method is implemented in PyTorch [80]. In this paper, we make use of ResNet18 [81] as CNN backbone architecture for feature extraction. Although other models could be also adopted, we utilize ResNet18 for the sake of simplicity, since it offers a reasonable trade-off between complexity and performance, considering the multi-task nature of the work. Additionally, the images are resized to 256×256 pixels, and data augmentation strategies including 1) RandomGrayscale; 2) ColorJitter; and 3) RandomHorizontalFlip are utilized. The parameters D, σ, k, q and T are set to 128, 0.05, 0.5, 0.7 and 40, respectively. Stochastic gradient descent (SGD) optimizer with the initial learning rate as 0.01 is adopted for optimizing the loss. The learning rate is decayed by 0.5 every 30 epochs. The batch size is 256, and we totally train the CNN models for 100 epochs. For validating the effectiveness of the proposed method, we compare it with several state-of-the-art deep metric learning methods including: 1) **D-CNN** [42]; 2) **Triplet** [68]; 3) **SNCA** [70]; 4) **NSL** [72], 5) **ArcFace** [73] and 6) **MAE** [75]. For training all these methods, we fine-tune the associated learning rates (on the corresponding validation sets) while considering the aforementioned parameters to conduct a fair experimental comparison.

TABLE I

K-NN ($K = 10$) CLASSIFICATION ACCURACIES (%) OF THE CONSIDERED METHODS ON THE TWO BENCHMARK DATASETS WITH TWO TYPES OF NOISE AT DIFFERENT LEVELS.

	AID								NWPU-RESISC45							
	Uniform				Label-dependent				Uniform				Label-dependent			
	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
D-CNN	92.40	86.80	75.25	60.10	92.45	88.95	84.40	84.25	90.05	81.92	**88.62**	40.59	90.06	85.70	80.00	77.19
Triplet	91.80	85.35	77.15	55.35	93.30	**90.50**	85.80	**85.60**	86.92	75.19	61.57	50.68	90.06	87.76	83.41	**79.25**
NSL	89.35	84.35	75.60	63.90	90.20	87.25	85.15	84.15	87.46	78.73	65.57	45.38	88.27	84.14	80.19	78.00
SNCA	90.65	82.70	64.55	42.90	90.40	81.95	75.40	74.45	87.90	76.08	59.75	30.03	87.17	77.97	68.02	63.44
ArcFace	90.30	79.95	87.30	**82.30**	90.80	81.35	**86.55**	83.60	87.75	88.68	85.52	**80.71**	87.16	71.25	80.35	78.25
MAE	82.15	82.70	79.90	80.65	83.45	81.05	82.25	81.15	78.69	78.53	75.43	74.03	77.94	77.65	76.97	77.17
RNSL	93.25	91.15	81.10	54.25	93.15	85.65	79.65	76.10	92.25	88.92	80.54	49.63	91.30	84.32	74.35	68.21
t-RNSL	**94.05**	**91.80**	**89.50**	78.25	**93.80**	**90.50**	86.05	81.55	**92.30**	**90.76**	88.56	79.84	**92.03**	**89.30**	**84.84**	76.84

TABLE II

NMI SCORES (%) OF THE CONSIDERED METHODS ON THE TWO BENCHMARK DATASETS WITH TWO TYPES OF NOISE AT DIFFERENT LEVELS.

	AID								NWPU-RESISC45							
	Uniform				Label-dependent				Uniform				Label-dependent			
	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
D-CNN	87.27	78.52	58.90	30.41	88.24	83.44	75.06	**72.88**	85.26	74.38	84.54	17.60	84.05	79.21	72.93	69.56
Triplet	87.47	78.10	65.38	37.67	89.78	85.34	77.20	71.94	82.32	64.78	48.59	35.60	86.23	82.54	76.18	69.28
NSL	83.94	72.87	47.02	22.29	86.95	79.60	74.63	72.57	81.31	66.84	41.86	11.55	83.74	77.48	72.00	**69.65**
SNCA	86.56	76.19	51.96	25.40	87.53	75.27	70.09	67.55	84.47	69.37	50.02	19.19	84.02	73.73	64.69	59.29
ArcFace	87.40	72.26	80.32	**71.67**	87.39	75.54	75.18	71.42	83.98	83.03	77.90	70.45	83.53	67.80	69.95	67.08
MAE	67.65	68.10	61.24	61.40	71.43	63.59	63.95	63.57	61.72	61.67	57.61	53.76	60.59	59.13	59.70	58.13
RNSL	91.09	86.70	67.93	24.25	90.60	78.64	68.75	62.00	**89.84**	85.40	72.13	23.43	88.36	78.97	67.83	61.31
t-RNSL	**91.45**	**88.11**	**85.57**	66.23	**91.62**	**86.13**	**79.87**	68.08	89.28	**88.01**	**84.81**	**74.04**	**89.17**	**85.04**	**79.47**	68.32

It is important to note that we use ResNet18 as feature extractor model for all the considered methods. Besides, we only focus on techniques that can be framed within the scheme of deep metric learning with label noise due to the multi-task nature of our work. All the experiments are performed on one NVIDIA Tesla P100 graphics processing unit (GPU).

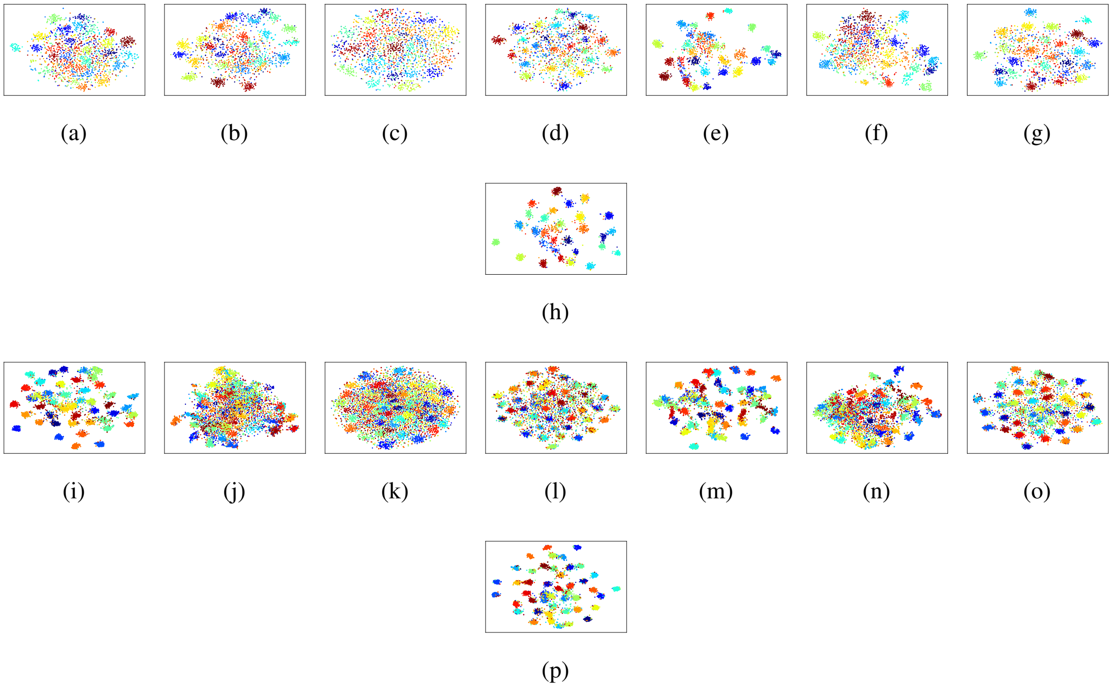
B. Experimental Results

1) KNN classification: Based on the trained CNN models of the considered methods with the noisy data of different noise rates, we conduct K-NN classification ($K = 10$) experiments on the

TABLE III

ACC SCORES (%) OF THE CONSIDERED METHODS ON THE TWO BENCHMARK DATASETS WITH TWO TYPES OF NOISE AT DIFFERENT LEVELS.

	AID								**NWPU-RESISC45**							
	Uniform				Label-dependent				Uniform				Label-dependent			
	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
D-CNN	91.65	80.90	65.50	28.85	87.70	79.50	73.35	64.50	86.86	77.86	**85.00**	17.41	85.71	78.65	66.40	59.78
Triplet	89.20	79.40	65.25	34.20	89.70	85.20	70.95	65.30	79.21	60.40	42.54	25.30	84.41	80.94	72.60	57.54
NSL	85.10	71.85	45.30	21.75	87.25	77.95	70.70	**68.55**	80.97	69.41	43.81	10.75	83.19	74.95	69.35	**65.65**
SNCA	90.00	81.75	60.20	26.30	90.15	79.05	63.80	59.55	88.00	76.06	58.59	24.62	86.94	76.21	60.03	47.02
ArcFace	90.55	79.25	78.50	68.95	87.90	77.00	65.30	62.00	87.70	78.48	75.41	68.38	86.87	68.49	52.73	48.62
MAE	63.85	66.45	53.30	57.75	67.15	57.20	56.45	56.65	52.13	52.97	50.70	46.46	53.33	50.13	51.84	49.37
RNSL	90.90	90.45	72.60	23.75	90.70	76.75	62.90	54.85	89.41	88.11	75.59	26.86	88.33	79.68	64.14	53.41
t-RNSL	**94.00**	**91.40**	**86.60**	**69.55**	**93.50**	**86.65**	**76.55**	63.20	**91.48**	**90.27**	84.98	**78.16**	**89.22**	**86.63**	**75.10**	61.22

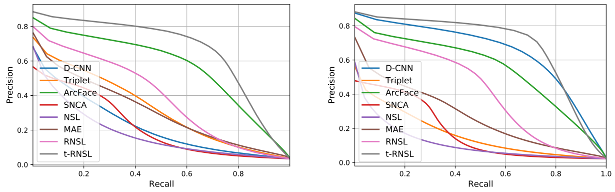

(a) (b) (c) (d) (e) (f) (g)

(h)

(i) (j) (k) (l) (m) (n) (o)

(p)

Fig. 7. 2-D projection of the extracted features of the two test sets based on the CNN model trained with uniform label noise when $\eta = 0.5$. AID: (a) D-CNN (b) Triplet (c) NSL (d) SNCA (e) ArcFace (f) MAE (g) RNSL (h) t-RNSL. NWPU-RESISC45: (i) D-CNN (j) Triplet (k) NSL (l) SNCA (m) ArcFace (n) MAE (o) RNSL (p) t-RNSL.



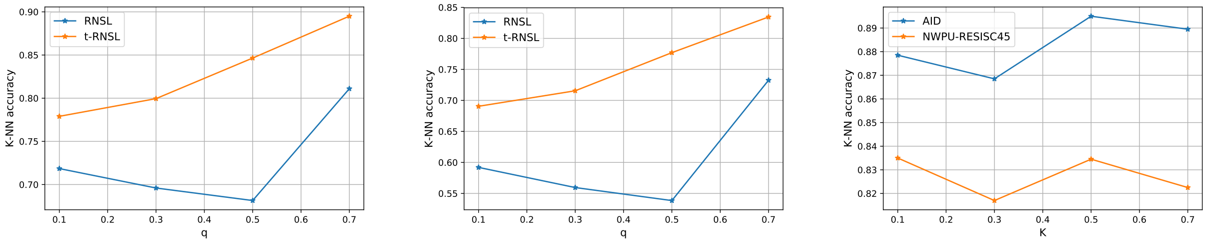
(a) (b)

Fig. 8. PR curves describing the image retrieval performances of all the methods based on the trained CNN models using training sets with uniform noise: AID (a) and NWPU-RESISC45 (b), when $\eta = 0.5$.

TABLE IV

MAP scores (%) of the considered methods on the two benchmark datasets with two types of noise at different levels when $R = 20$.

	AID								**NWPU-RESISC45**							
	Uniform				Label-dependent				Uniform				Label-dependent			
	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
D-CNN	93.25	84.63	69.17	50.98	93.44	87.28	81.46	78.61	91.60	80.94	90.28	41.53	91.02	85.24	77.80	73.44
Triplet	93.13	85.87	74.96	56.34	93.79	90.64	86.44	**83.56**	87.99	75.10	61.50	51.79	90.95	89.57	84.42	**78.85**
NSL	90.61	81.01	67.96	55.26	90.40	85.50	81.84	78.96	88.71	76.53	62.30	44.44	89.39	83.42	78.14	75.52
SNCA	**96.81**	89.12	68.92	48.05	**95.41**	85.72	78.64	76.99	96.09	85.49	67.07	42.06	92.71	83.24	72.64	69.38
ArcFace	96.04	85.55	85.41	77.27	94.72	83.88	84.47	81.73	**96.17**	88.59	85.20	78.30	93.84	77.83	79.69	77.37
MAE	79.18	80.09	74.04	74.77	81.88	76.29	77.04	76.27	76.03	75.57	73.13	69.74	81.88	76.29	77.04	76.27
RNSL	95.53	93.28	82.23	51.48	94.78	85.72	76.92	71.12	95.26	92.43	84.49	53.21	94.11	86.24	73.25	66.62
t-RNSL	96.04	**93.71**	**90.55**	**77.41**	95.38	**91.37**	**86.53**	77.79	95.10	**94.11**	**91.33**	**83.29**	**94.83**	**91.32**	**85.35**	76.23



(a) (b) (c)

Fig. 9. Sensitivity analysis of hyperparameters q and k. (a) K-NN classification performances on AID with different q. (b) K-NN classification performances on NWPU-RESISC45 with different q. (c) K-NN classification performances on both datasets with different k in t-RNSL.

test sets and show the results in Table I. As for the uniform noise, t-RNSL can achieve the best performances under different noise rates on the two datasets. For RNSL, when η is less than 0.5, it achieves the second-best performances on the two RS datasets. It indicates that the utilized negative Box-Cox transformation indeed mitigates the effects on the learning of CNN models induced by the images with noisy labels. However, when the noise rate is large, e.g., $\eta = 0.7$, large amounts of noisy data can still confuse the learning of class prototypes and the CNN models. Under such case, its truncated version (t-RNSL) can further improve its robustness through the pruning strategy compared with RNSL. Therefore, when $\eta = 0.7$, the classification results from t-RNSL outperform RNSL by large margins. Although the losses induced by the images with lower p_i^c are down-weighted in RNSL, they still make more contributions to the learning of class prototypes and the trained CNN models cannot accurately capture the geometry structures of the features in the feature space. Although MAE is also a robust loss for classification with noisy labels, its gradients are treated equally with respect to both easily-classified or hard-classified images. However, this effect will not be beneficial for capturing discriminative features of RS scenes with complex semantics, i.e., hard-positives. Therefore, MAE can be robust to the different noise levels, while the overall performance cannot be comparable with respect to t-RNSL. Considering the other losses except ArcFace, the classification performances significantly degrade as the noise level increases. For ArcFace, the angular margin parameter is utilized for enforcing the compactness of the intra-class features. It also demonstrates the robustness against noisy labels compared with the other baseline losses. As for the label-dependent noise, t-RNSL can achieve the comparable classification accuracies with respect to the state-of-the-art method, e.g., D-CNN or Triplet, when the labels are corrupted with different noise rates. To this end, for both two types of label noise with different noise rates, the proposed method can achieve the superior K-NN classification performances compared with the other state-of-the-art methods.

2) Clustering: Table II and Table III report the NMI and ACC scores of the considered methods carried out on the two benchmark datasets with two types of noise at different noise rates. For uniform noise, consistently with the above analysis, the K-means clustering results based on t-RNSL can reach the best or second-best performances on the degree of the agreement that the produced pseudo-labels can match the associated ground-truth labels. When the labels are corrupted by heavy noise, e.g., $\eta = 0.7$, t-RNSL can present much higher NMI and ACC scores than most of the other considered methods on the two RS datasets. Moreover, in Figure 7, we first extract the features of the two test sets via the CNN model trained with labels corrupted by uniform noise ($\eta = 0.5$), and project them into the 2-D space by the t-distributed stochastic neighbour embedding (t-SNE). It can be obviously observed that the intra-class features are more compact and the inter-class features are better separated via t-RNSL than the other considered methods. Therefore, the produced features based on t-RNSL can be better clustered through K-means clustering in the feature space, which will lead to the higher NMI and ACC scores than the other methods. For label-dependent noise, t-RNSL can also achieve the best performance compared with the other losses when $\eta = 0.1, 0.3$ and 0.5.

3) Image retrieval: Figure 8 shows the PR curves describing the image retrieval performances of all the methods based on the trained CNN models on the training sets with uniform noise: AID (a) and NWPU-RESISC45 (b), when $\eta = 0.5$. It can be observed that t-RNSL exhibits higher precision and recall scores compared to the other considered methods, when the CNN models are trained on the datasets with uniform label noise. Therefore, t-RNSL can be exploited for large-scale RS image retrieval task when the label annotations are not accurate. Table IV displays the MAP scores of the considered methods on the two benchmark datasets with two types of noise at different levels when $R = 20$. Consistently with the above experiments, for the uniform noise, the image retrieval results obtained by t-RNSL can reach the superior performances with respect to the other losses. When $\eta = 0.1$, SNCA can achieve the best retrieval performance, while the proposed losses are slightly lower than it. By stochastically maximizing the leave-one-out K-NN score, SNCA can better discover the inherent neighborhood structure among the images in the feature space than the other class-prototype based deep metric learning losses, such as NSL. However, t-RNSL can outperform SNCA by large margins when the training data contain uniform label noise with high proportions. As for the label-dependent noise, t-RNSL can preserve the best retrieval performance of all the considered methods when η is less than 0.7.

4) Hyperparameter analysis: Two main parameters in the proposed losses are q and k, where q controls the power of p_i^c and k is the threshold from which loss starts to affect the learning of class prototypes $\mathbf{w}_c$. We analyze their performance sensitivities based on K-NN classification conducted on the features from the two test sets, extracted via the trained CNN model on the training sets under uniform noise when $\eta = 0.5$. For the temperature parameter σ, following the theoretical analysis in [82], $1/\sigma$ denotes the radius of the hypersphere which the features are projected on. Larger values of $1/\sigma$ can ensure sufficient hyperspace for feature learning with an expected large margin so that the class-wise features can be discriminately separated. Therefore, it is recommended to set this as a relative small number [82], [83], e.g., $\sigma = 0.05$, and we keep it constant for all the experiments in this paper. Figure 9 shows the parameter sensitivity analysis results. It suggests that the proposed losses are favored with a relatively big value of q. For example, for both datasets, the highest performances are observed when $q = 0.7$. This is consistent with Figure 4. Larger q will lead to stronger down-weighting effects on the learning of $\mathbf{w}_c$ based on the hard images. As for k, the effectiveness of the proposed methods are not influenced much with respect to the variations of its values.

C. Discussion

According to our experimental results, we observe that t-RNSL can achieve the best performance on all the three considered tasks (K-NN classification, clustering and image retrieval) under the deep metric learning scheme when the training labels are corrupted by the uniform loss. The associated performances do not degrade much as the noise rate increases. Compared with RNSL, t-RNSL performs more robustly against the label noise, since a pruning strategy is taken through a truncated loss to avoid the high contributions on learning the class prototypes from the images with wrong labels, especially when the noise rates are high. For practical applications, both RNSL and t-RNSL can be adopted for robustly learning the image features when small amounts of images are wrongly labeled (i.e., $\eta < 0.5$). However, when the noise rates are high (i.e., $\eta \geq 0.5$), t-RNSL is recommended to be exploited. As indicated by the t-SNE results, most of the compared methods cannot robustly discover the locality structure for the features in the feature space, as the intra-class features cannot be separated class-wisely. As for class-dependent noise, we observe that all the considered methods can perform in a stable manner when the noise level is not high, and the proposed method can achieve the best accuracies. When $\eta = 0.7$, D-CNN or Triplet losses can achieve the best performance. The plausible reason is that the D-CNN and Triplet losses are optimized in a relationship-based training mechanism, i.e., pairwise or triplet. When the pairwise or triplet relationships among images are correct, the associated losses can be well-performed on the learning, no matter if the label of the individual image is correct or not. Compared with the relationship-based losses, the performance of t-RNSL is slightly lower. As for the hyperparameter setting on t-RNSL, we can set q and k as constant values (i.e., $q = 0.7$ and $k = 0.5$) for all the experiments with different noise rates, so that the parameter tuning will not cost much efforts and the effectiveness of the proposed loss can be preserved. Another parameter T, which triggers the pruning procedure of noisy images, can be set to a constant value, e.g., $T = 30$, when the noise level is low. As for heavy noise, e.g., $\eta = 0.7$, its value can be adapted so that most noisy images do not mislead the learning of class prototypes in the early stages.

V. Conclusion

This paper presents a novel loss formulation for robust deep metric learning of RS images annotated with label noise. To improve the robustness of the NSL commonly utilized for deep metric learning, we introduce a new RNSL which replaces the logarithm function by the negative Box-Cox transformation in order to down-weight the contributions from the noisy images on the learning of class prototypes. Moreover, by truncating the loss with a certain threshold, the proposed t-RNSL can enforce the learning of class prototypes based on the features with high similarities between them, so that intra-class features can be well grouped and inter-class features can be well separated. Compared with several state-of-the-art metric learning losses, the proposed losses can demonstrate better performances on the K-NN classification, clustering and image retrieval conducted on the extracted features through the trained CNN models. In practice, the proposed losses can be utilized for feature learning and image retrieval with large-scale RS data without precise annotations. As a future work, we would like to extend the proposed losses to the multi-label case by exploring the use of semantic prototype mixtures. In addition, noise rate estimation is worth to be further studied in future developments.

Appendix

In order to simulate the realistic cases of label-dependent noise for RS scenes, we manually create its probability transition matrix $\mathbf{Q}$ based on the visual semantic-similarities among different land-use and land-cover classes. For the two benchmark datasets, we flip each label to other labels with similar semantic-contents based on the probabilities shown in Table V and table VI when $\eta = 0.5$. Note that, when considering other noise rates, original labels are preserved with a probability of $1 - \eta$ and the remaining probability values are proportionally modified.

References

[1] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, "Weakly supervised learning for target detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 4, pp. 701–705, 2014.

TABLE V

Label transition probabilities for creating label-dependent noise of AID dataset when $\eta = 0.5$.

Airport →	[Airport:0.5, BareLand:0.1, Industrial:0.1, Parking:0.1, RailwayStation:0.1, StorageTanks:0.1]	BareLand →	[BareLand:0.5, Desert:0.3, Mountain:0.2]
BaseballField →	[BaseballField:0.5, Meadow:0.2, SparseResidential:0.1, Farmland:0.2]	Beach →	[Beach:0.5, Bridge:0.1, Pond:0.2, Port:0.2]
Bridge →	[Bridge:0.5, Beach:0.1, Pond:0.2, Port:0.2]	Center →	[Center:0.5, Church:0.3, Commercial:0.1, Square:0.1]
Church →	[Church:0.5, Center:0.3, Commercial:0.1, Square:0.1]	Commercial →	[Commercial:0.5, Center:0.2, Church:0.1, Square:0.1, DenseResidential:0.1]
DenseResidential →	[DenseResidential:0.5, Center:0.2, Church:0.1, Square:0.1, Resort:0.1]	Desert →	[Desert:0.5, BareLand:0.3, Mountain:0.2]
Farmland →	[Farmland:0.5, BaseballField:0.2, Forest:0.2, Meadow:0.1]	Forest →	[Forest:0.5, BaseballField:0.2, Farmland:0.2, Meadow:0.1]
Industrial →	[Industrial:0.5, Airport:0.1, Square:0.1, Viaduct:0.1, StorageTanks:0.2]	Meadow →	[Meadow:0.5, BaseballField:0.1, Farmland:0.1, Forest:0.2, River:0.1]
MediumResidential →	[MediumResidential:0.5, SparseResidential:0.3, DenseResidential:0.2]	Mountain →	[Mountain:0.5, BareLand:0.2, Meadow:0.2, Farmland:0.1]
Park →	[Park:0.5, School:0.2, SparseResidential:0.2, Square:0.1]	Parking →	[Parking:0.5, Airport:0.2, RailwayStation:0.2, BareLand:0.1]
Playground →	[Playground:0.5, Stadium:0.4, Park:0.1]	Pond →	[Pond:0.5, Bridge:0.2, Port:0.1, River:0.2]
Port →	[Port:0.5, Pond:0.2, Beach:0.1, River:0.2]	Resort →	[Resort:0.5, Center:0.2, Square:0.1, Church:0.2]
River →	[River:0.5, Bridge:0.2, Forest:0.1, Meadow:0.1, Pond:0.1]	School →	[School:0.5, Resort:0.2, Stadium:0.1, Square:0.1, Center:0.1]
SparseResidential →	[SparseResidential:0.5, MediumResidential:0.3, DenseResidential:0.2]	Square →	[Square:0.5, Viaduct:0.2, Airport:0.2, Parking:0.1]
Stadium →	[Stadium:0.5, Playground:0.4, School:0.1]	StorageTanks →	[StorageTanks:0.5, Industrial:0.2, Airport:0.1, Parking:0.2]
Viaduct →	[Viaduct:0.5, Square:0.3, RailwayStation:0.2]	RailwayStation →	[RailwayStation:0.5, Square:0.3, Viaduct:0.2]

[2] Z. Wang, L. Du, P. Zhang, L. Li, F. Wang, S. Xu, and H. Su, "Visual attention-based target detection and discrimination for high-resolution sar images in complex scenes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 1855–1872, 2017.

[3] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 44–59, 2018.

[4] 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 Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[5] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in sar images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2020.

[6] R. Tao, X. Zhao, W. Li, H.-C. Li, and Q. Du, "Hyperspectral anomaly detection by fractional fourier entropy," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 12, pp. 4920–4929, 2019.

[7] Z. Xue, P. Du, and L. Feng, "Phenology-driven land cover classification and trend analysis based on long-term remote sensing image series," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 4, pp. 1142–1156, 2014.

[8] R. Fernandez-Beltran, A. Plaza, J. Plaza, and F. Pla, "Hyperspectral unmixing based on dual-depth sparse probabilistic

TABLE VI

LABEL TRANSITION PROBABILITIES FOR CREATING LABEL-DEPENDENT NOISE OF NWPU-RESISC45 DATASET WHEN $\eta = 0.5$.

airplane →	[airplane:0.5, airport:0.2, freeway:0.1, industrial_area:0.1, runway:0.1]	airport →	[airport:0.5, airplane:0.1, freeway:0.1, industrial_area:0.1, runway:0.1, storage_tank:0.1]
baseball_diamond →	[baseball_diamond:0.5, basketball_court:0.1, circular_farmland:0.1, golf_course:0.2, rectangular_farmland:0.1]	basketball_court →	[basketball_court:0.5, baseball_diamond:0.1, ground_track_field: 0.3, tennis_court:0.1]
beach →	[beach:0. 5, bridge:0.1, harbor:0.1, island:0.1, lake:0.1, river:0.1]	bridge →	[bridge:0.5, beach:0.1, wetland:0.1, island:0.1, lake:0.1, river:0.1]
chaparral →	[chaparral:0.5, desert:0.4, terrace:0.1]	church →	[church:0.5, commercial_area:0.2, industrial_area:0.1, palace:0.2]
circular_farmland →	[circular_farmland:0.5, baseball_diamond:0.1, forest:0.1, meadow:0.1, rectangular_farmland:0.2]	cloud →	[cloud:0.5, island:0.1, sea_ice:0.2, snowberg:0.2]
commercial_area →	[commercial_area:0.5, church:0.1, dense_residential:0.2, industrial_area:0.1, roundabout:0.1]	dense_residential →	[dense_residential:0.5, mobile_home_park:0.1, commercial_area:0.2, thermal_power_station:0.1, intersection:0.1]
desert →	[desert:0.5, chaparral:0.3, terrace:0.2]	forest →	[forest:0.5, golf_course:0.1, meadow:0.2, wetland:0.1, rectangular_farmland:0.1]
freeway →	[freeway:0.5, bridge:0.1, intersection:0.1, overpass:0.1, railway:0.1, runway:0.1]	golf_course →	[golf_course:0.5, baseball_diamond:0.1, sparse_residential:0.1, forest:0.1, meadow:0.2]
ground_track_field →	[ground_track_field:0.5, tennis_court:0.1, stadium:0.1, basketball_court:0.1, baseball_diamond:0.2]	harbor →	[harbor:0.5, beach:0.1, island:0.1, mobile_home_park:0.2, ship:0.1]
industrial_area →	[industrial_area:0.5, storage_tank:0.1, thermal_power_station:0.1, railway_station:0.2, roundabout:0.1]	intersection →	[intersection:0.5, roundabout:0.1, railway_station:0.1, overpass:0.2, commercial_area:0.1]
island →	[island:0.5, beach:0.2, wetland:0.1, lake:0.2]	lake →	[lake:0.5, beach:0.2, wetland:0.2, island:0.1]
meadow →	[meadow:0.5, forest:0.2, wetland:0.3]	medium_residential →	[medium_residential:0.5, forest:0.2, ground_track_field:0.1, tennis_court:0.1, sparse_residential:0.1]
mobile_home_park →	[mobile_home_park:0.5, railway_station:0.2, commercial_area:0.1, dense_residential:0.1, industrial_area:0.1]	mountain →	[mountain:0.5, meadow:0.2, snowberg:0.1, terrace:0.2]
overpass →	[overpass:0.5, freeway:0.2, intersection:0.1, railway:0.1, runway:0.1]	palace →	[palace:0.5, church:0.3, commercial_area:0.1, dense_residential:0.1]
parking_lot →	[parking_lot:0.5, mobile_home_park:0.3, railway_station:0.1, runway:0.1]	railway →	[railway:0.5, freeway:0.2, intersection:0.1, runway:0.1, railway_station:0.1]
railway_station →	[railway_station:0.5, railway:0.2, industrial_area:0.1, airport:0.1, runway:0.1]	rectangular_farmland →	[rectangular_farmland:0.5, baseball_diamond:0.2, circular_farmland:0.1, meadow:0.1, wetland:0.1]
river →	[river:0.5, bridge:0.1, meadow:0.1, lake:0.1, wetland:0.2]	roundabout →	[roundabout:0.5, intersection:0.4, ground_track_field:0.1]
runway →	[runway:0.5, airport:0.1, freeway:0.2, intersection:0.1, overpass:0.1]	sea_ice →	[sea_ice:0.5, island:0.1, cloud:0.2, wetland:0.1, snowberg:0.1]
ship →	[ship:0.5, harbor:0.2, bridge:0.2, river:0.1]	snowberg →	[snowberg:0.5, cloud:0.3, sea$_ice : 0.2$]
sparse_residential →	[sparse_residential:0.5, baseball_diamond:0.2, golf_course:0.2, meadow:0.1]	stadium →	[stadium:0.5, tennis_court:0.2, basketball_court:0.3]
storage_tank →	[storage_tank:0.5, industrial_area:0.4, mobile_home_park:0.1]	tennis_court →	[tennis_court:0.5, baseball_diamond:0.2, basketball_court:0.1, stadium:0.2]
terrace →	[terrace:0.5, circular_farmland:0.2, chaparral:0.1, rectangular_farmland:0.2]	thermal_power_station →	[thermal_power_station:0.5, industrial_area:0.2, cloud:0.1, storage_tank:0.2]
wetland →	[wetland:0.5, bridge:0.1, forest:0.1, lake:0.2, river:0.1]		

latent semantic analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6344–6360, 2018.

[9] C. Paris, L. Bruzzone, and D. Fernández-Prieto, "A novel approach to the unsupervised update of land-cover maps by classification of time series of multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4259–4277, 2019.

[10] R. Fernandez-Beltran, F. Pla, and A. Plaza, "Endmember extraction from hyperspectral imagery based on probabilistic tensor moments," *IEEE Geoscience and Remote Sensing Letters*, 2020, dOI: 10.1109/LGRS.2019.2963114.

[11] 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.*, 2020, dOI: 10.1109/TNNLS.2020.2979546.

[12] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-lidar and hyperspectral data," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[13] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, D. Qian, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, 2020, dOI: 10.1109/TGRS.2020.3016820.

[14] 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.*, 2020, dOI: 10.1109/TGRS.2020.3015157.

[15] J. Haas and Y. Ban, "Mapping and monitoring urban ecosystem services using multitemporal high-resolution satellite data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 2, pp. 669–680, 2016.

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

[17] L. Fang, Z. Ye, S. Su, J. Kang, and X. Tong, "Glacier surface motion estimation from sar intensity images based on subpixel gradient correlation," *Sensors*, vol. 20, no. 16, p. 4396, 2020.

[18] 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 Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 359–372, 2020.

[19] J. Segarra, M. L. Buchaillot, J. L. Araus, and S. C. Kefauver, "Remote sensing for precision agriculture: Sentinel-2 improved features and applications," *Agronomy*, vol. 10, no. 5, p. 641, 2020.

[20] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote Sensing of Environment*, vol. 236, p. 111402, 2020.

[21] 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, 2019.

[22] S. M. Xie, N. Jean, M. Burke, D. B. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," in *AAAI*, 2016, pp. 3929–3935.

[23] Y. Ni, X. Li, Y. Ye, Y. Li, C. Li, and D. Chu, "An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[24] S. Liang, *Comprehensive Remote Sensing*. Elsevier, 2017.

[25] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 1, p. 193, 2011.

[26] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Single-frame super-resolution in remote sensing: a practical overview," *International journal of remote sensing*, vol. 38, no. 1, pp. 314–354, 2017.

[27] C. Gómez, J. C. White, and M. A. Wulder, "Optical remotely sensed time series data for land cover classification: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 55–72, 2016.

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

[29] G. Thoonen, Z. Mahmood, S. Peeters, and P. Scheunders, "Multisource classification of color and hyperspectral images using color attribute profiles and composite decision fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 510–521, 2012.

[30] C. Chen, B. Zhang, H. Su, W. Li, and L. Wang, "Land-use scene classification using multi-scale completed local binary patterns," *Signal, image and video processing*, vol. 10, no. 4, pp. 745–752, 2016.

[31] Y. Yang and S. Newsam, "Comparing sift descriptors and gabor texture features for classification of remote sensed imagery," in *2008 15th IEEE international conference on image processing*. IEEE, 2008, pp. 1852–1855.

[32] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 439–451, 2014.

[33] 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 Geoscience and Remote Sensing Letters*, vol. 15, no. 9, pp. 1347–1351, 2018.

[34] E. Li, P. Du, A. Samat, Y. Meng, and M. Che, "Mid-level feature representation via sparse autoencoder for remotely sensed scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 3, pp. 1068–1081, 2017.

[35] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1793–1802, 2015.

[36] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.

[37] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[38] X. Zhao, R. Tao, W. Li, H.-C. Li, Q. Du, W. Liao, and W. Philips, "Joint classification of hyperspectral and lidar data using hierarchical random walk and deep cnn architecture," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[39] J. A. Benediktsson, J. Chanussot, and W. M. Moon, "Very high-resolution remote sensing: Challenges and opportunities [point of view]," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1907–1910, 2012.

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

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

[42] 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 transactions on geoscience and remote sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[43] 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 Transactions on Geoscience and Remote Sensing*, 2019.

[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 Transactions on Geoscience and Remote Sensing*, 2020, dOI:10.1109/TGRS.2020.2991657.

[45] 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 Transactions on Geoscience and Remote Sensing*, 2020, dOI:10.1109/TGRS.2020.3007029.

[46] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multi-label remote sensing image classification and retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, 2020, dOI:10.1109/TGRS.2020.3016020.

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

[48] B. Zhang, Z. Chen, D. Peng, J. A. Benediktsson, B. Liu, L. Zou, J. Li, and A. Plaza, "Remotely sensed big data: evolution in model development for information extraction [point of view]," *Proceedings of the IEEE*, vol. 107, no. 12, pp. 2294–2301, 2019.

[49] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "Dirs: On creating benchmark datasets for remote sensing image interpretation," *arXiv preprint arXiv:2006.12485*, 2020.

[50] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Häberle, Y. Hua, R. Huang *et al.*, "So2sat lcz42: A benchmark dataset for global local climate zones classification," *arXiv preprint arXiv:1912.12171*, 2019.

[51] G.-S. Xia, Z. Wang, C. Xiong, and L. Zhang, "Accurate annotation of remote sensing images via active spectral clustering with little expert knowledge," *Remote Sensing*, vol. 7, no. 11, pp. 15 014–15 045, 2015.

[52] Y. Li and D. Ye, "Greedy annotation of remote sensing image scenes based on automatic aggregation via hierarchical similarity diffusion," *IEEE Access*, vol. 6, pp. 57 376–57 388, 2018.

[53] H. Li, X. Dou, C. Tao, Z. Hou, J. Chen, J. Peng, M. Deng, and L. Zhao, "Rsi-cb: A large scale remote sensing image classification benchmark via crowdsource data," *arXiv preprint arXiv:1705.10450*, 2017.

[54] P. Jin, G.-S. Xia, F. Hu, Q. Lu, and L. Zhang, "Aid++: An updated version of aid on scene classification," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 4721–4724.

[55] J. Vargas, S. Srivastava, D. Tuia, and A. Falcao, "Openstreetmap: Challenges and opportunities in machine learning and remote sensing," *arXiv preprint arXiv:2007.06277*, 2020.

[56] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Learning to learn from noisy labeled data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5051–5059.

[57] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1944–1952.

[58] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5552–5560.

[59] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. Change Loy, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 765–780.

[60] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Transactions on Cybernetics*, 2020.

[61] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[62] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6712–6722, 2018.

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

[64] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 117–126, 2017.

[65] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4799–4809, 2019.

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

[67] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[68] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[69] R. Cao, Q. Zhang, J. Zhu, Q. Li, Q. Li, B. Liu, and G. Qiu, "Enhancing remote sensing image retrieval with triplet deep metric learning network," *arXiv preprint arXiv:1902.05818*, 2019.

[70] Z. Wu, A. A. Efros, and S. X. Yu, "Improving generalization via scalable neighborhood component analysis," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 685–701.

[71] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in neural information processing systems*, 2005, pp. 513–520.

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

[73] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[74] T. Yuan, W. Deng, J. Tang, Y. Tang, and B. Chen, "Signal-to-noise ratio: A robust distance metric for deep metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4815–4824.

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

[76] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8778–8788.

[77] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," *arXiv preprint arXiv:1406.2080*, 2014.

[78] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[79] H. Schütze, C. D. Manning, and P. Raghavan, "Introduction to information retrieval," in *Proceedings of the international communication of association for computing machinery conference*, 2008, p. 260.

[80] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[81] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[82] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[83] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "High-rankness regularized semi-supervised deep metric learning for remote sensing imagery," *Remote Sensing*, vol. 12, no. 16, p. 2603, 2020.

Jian Kang (S'16-M'19) received B.S. and M.E. degrees in electronic engineering from Harbin Institute of Technology (HIT), Harbin, China, in 2013 and 2015, respectively, and 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 at Institute of Computer Graphics and Vision (ICG), TU Graz, Graz, Austria. From 2019 to 2020, he was with 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 215006, China. His research focuses on signal processing and machine learning techniques, and their applications in remote sensing. In particular, he is interested in multi-dimensional data analysis, geophysical parameter estimation based on InSAR data, SAR denoising and deep learning based techniques for remote sensing image analysis. He obtained first place of the best student paper award in EUSAR 2018, Aachen, Germany. His joint work was selected as one of the 10 Student Paper Competition Finalists in IGARSS 2020.

Ruben Fernandez-Beltran (M'20) earned a B.Sc. degree in Computer Science, a M.Sc. in Intelligent Systems and a Ph.D. degree in Computer Science, from Universitat Jaume I (Castellon de la Plana, Spain) in 2007, 2011 and 2016, respectively. He is currently a postdoctoral researcher within the Computer Vision Group of the University Jaume I, as a member of the Institute of New Imaging Technologies. He has been visiting researcher at the University of Bristol (UK), University of Cáceres (Spain) and Technische Universität Berlin (Germany). 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). His research interests lie in multimedia retrieval, spatio-spectral image analysis, pattern recognition techniques applied to image processing and remote sensing. He was awarded with the Outstanding Ph.D. Dissertation Award at Universitat Jaume I in 2017.

Puhong Duan (S'17) received the B.Sc degree from Suzhou University, Suzhou, China, in 2014, and the M.S. degree from Hefei University of Technology, Hefei, China, in 2017. He is currently working toward the Ph.D. degree in the Laboratory of Vision and Image Processing, Hunan University, Changsha, China. His research interests include image classification, visualization, object detection and image fusion.

Xudong Kang (S'13-M'15-SM'17) 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 of Hunan University, Changsha, China. His research interest includes 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.

Antonio Plaza (M'05-SM'07-F'15) received the M.Sc. degree and the Ph.D. degree 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 TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS). He is also a fellow of IEEE for contributions to hyperspectral data processing and parallel computing of earth observation data. 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 TOPICS 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. Additional information: http://www.umbc.edu/rssipl/people/aplaza