Rotation Invariant Deep Embedding for Remote Sensing Images

Jian Kang, *Member, IEEE,* Ruben Fernandez-Beltran, *Senior Member, IEEE,* Zhirui Wang, Xian Sun, *Senior Member, IEEE,* Jingen Ni, *Senior Member, IEEE,* and Antonio Plaza, *Fellow, IEEE*

***Abstract*—Endowing convolutional neural networks (CNNs) with the rotation-invariant capability is important for characterizing the semantic contents of remote sensing (RS) images, since they do not have typical orientations. Most of the existing deep methods for learning rotation-invariant CNN models are based on the design of proper convolutional or pooling layers, which aim at predicting the correct category labels of the rotated RS images equivalently. However, few works have focused on learning rotation-invariant embeddings in the framework of deep metric learning for modeling the fine-grained semantic relationships among RS images in the embedding space. To fill this gap, we first propose a rule that the deep embeddings of rotated images should be closer to each other than those of any other images (including the images belonging to the same class). Then, we propose to maximize the joint probability of the leave-one-out image classification and rotational image identification. With the assumption of independence, such optimization leads to the minimization of a novel loss function composed of two terms: 1) a class-discrimination term, and 2) a rotation-invariant term. Further, we introduce a penalty parameter that balances these two terms, and further propose a final loss to learn Rotation invariant Deep embedding for RS images, termed as `RiDe`. Extensive experiments conducted on two benchmark RS datasets validate the effectiveness of the proposed approach and demonstrate its superior performance when compared to other state-of-the-art methods. The codes of this paper will be publicly available from https://github.com/jiankang1991/TGRS_RiDe.**

***Index Terms*—Rotation invariant, convolutional neural networks (CNNs), scene classification, image retrieval, deep learning, deep metric learning, remote sensing (RS).**

I. Introduction

REMOTE sensing (RS) images have been widely used in multiple applications related to earth observation, such as object detection and recognition [1]–[5], land-use or land-cover classification [6]–[12], disaster monitoring and management of natural resources [13], [14], among others. All these tasks require an accurate characterization of RS scenes, from which semantic concepts should be precisely captured. Therefore, extensive methods for RS scene interpretation have been developed in recent years [15]–[17].

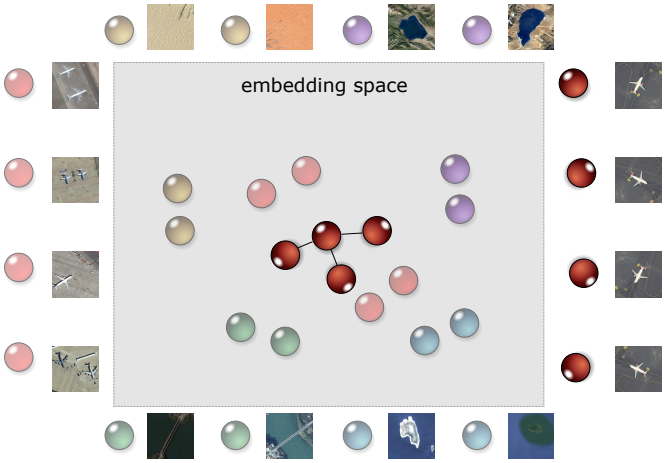


Fig. 1. A simplified illustration of the motivation of this work. To achieve rotation-invariant deep embedding, we aim at generating a hierarchical structure in the deep embedding space, which satisfies the following conditions: 1) the intraclass embeddings are grouped, and the interclass ones are separated; and 2) given the embedding of a source image, its nearest neighbors should be the embeddings from its rotated images.

Generally, existing RS scene characterization methods can be categorized into two main groups: 1) handcrafted feature-based methods [18], [19], and 2) data-driven feature-based methods [20]. Compared to the data-driven ones, handcrafted features are mainly constructed by applying color, texture or histograms of oriented gradients (HOG) descriptors on the RS images [21]–[24]. Although they demonstrate prominent performance in scene interpretation tasks (e.g., classification), it is still possible to improve their performance, especially for RS scenes with complex semantic contents. Data-driven feature-based methods aim at automatically learning or discovering the image descriptors by optimizing certain objective functions based on the training data, e.g., sparse coding [25], [26]. With the rapid development of deep learning methods, convolutional neural networks (CNNs) have been widely exploited for capturing the high-level semantic information of RS scenes in an end-to-end manner [27]. By enforcing intraclass compactness and interclass separation, deep metric

This work was in part supported by 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: Jingen Ni)*

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

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).

Z. Wang and X. Sun are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, also with the Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China. X. Sun is also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: sunxian@mail.ie.ac.cn).

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).

learning has been recently adopted for accurately capturing the complex semantics of RS scenes into low-dimensional vectors, termed *embeddings* [28]–[31]. This approach has been used successfully for RS scene classification and image retrieval tasks. In order to train CNN models under deep metric learning assumptions, most deep metric learning methods require supervised information (such as image annotations) for constructing image pairs or triplets with semantic relationships, where the images sharing the same label are semantically similar and dissimilar images have different labels. Then, loss functions are designed to pull together the intraclass deep embeddings and separate away the ones from different classes. Similar ideas of deep embedding even succeed in 3D point cloud processing via a dimensionality reduction of point features that originally encoded by a neural network [32]. However, these methods may not discover a fine-grained embedding space for RS scene characterization, since grouping intraclass deep embeddings may not be beneficial for accurately modeling their local relationships. Consequently, these may lead to image retrieval systems that cannot accurately rank the retrieved images based on their actual visual semantics with respect to the query. Moreover, unlike other kinds of images, RS scenes do not have typical orientations, because they are captured by airborne and spaceborne sensors. That is, the same semantic scene may appear at different geo-locations, being the only difference the orientations.

In the context of characterizing RS scenes via deep metric learning, all the images belonging to the same category are expected to produce similar representations (no matter whether they have been rotated or not). Therefore, the possibility of robustly learning the corresponding deep embeddings is highly beneficial to efficiently exploit these RS image characterizations in different downstream applications (e.g., classification, retrieval, etc.). Although equivariance and invariance are certainly two important aspects in this type of representation learning [33], the rotation-invariant alternative is more convenient to guarantee the locality structure among RS images in the metric space. Therefore, CNN models trained based on deep metric losses should be rotation-invariant. In order to solve the above issues, we introduce a novel loss function for learning a hierarchical structure of the deep embedding space (as shown in Figure 1), which satisfies the following conditions:

- Rotational invariance: given the embedding of a source image, its nearest neighbors should be the embeddings from its rotated images.
- Class discrimination: the intraclass embeddings are grouped together and the interclass ones are separated away.

To simultaneously achieve these two conditions, we maximize the joint probability of the leave-one-out image classification and rotational image identification. With the assumption of independence, we develop a new loss function composed of two terms: 1) class-discrimination, and 2) rotation-invariance. To balance these terms, we introduce a penalty parameter and further propose a final loss function to learn Rotation invariant Deep embeddings for RS images, termed as `RiDe`. To this end, the main contributions of this paper can be summarized as follows:

1) To our best knowledge, this work is the first one that focuses on analyzing the rotation-invariant capability of CNN models when extracting deep embeddings for RS images.
2) We introduce a novel metric learning loss function, `RiDe`, which can endow CNN models with both rotation-invariant and class-discrimination capabilities.
3) Our newly developed `RiDe` can be adapted to guide the training of any CNN model in a plug&play manner.
4) Based on our extensive experiments, we conclude that `RiDe` exhibits prominent performance in the generation of rotation-invariant deep embeddings as compared to other state-of-the-art losses.

The remainder of this paper is organized as follows. Section II introduces some related works. Section III thoroughly describes the proposed rotation-invariant deep embedding for RS images. Section IV presents the conducted experiments and discusses the obtained results. Section V concludes the paper with some remarks and hints at plausible future research lines.

II. Related Work

A. Deep Learning-based RS Scene Classification and Retrieval

Recently, deep learning techniques have drawn significant attention in the RS field, and extensive research has been carried out with the goal of characterizing the semantic contents of RS scenes. For example, Zheng *et al.* exploited pretrained CNN features, multi-scale pooling and Fisher vectors to generate invariant CNN features while enhancing their discriminative capability, and proposed a new deep scene classification method in [34]. Li *et al.* presented a multi-layer feature fusion method based on different pretrained CNN models for RS scene characterization [35]. To classify complex RS scenes, spatial pyramid pooling (combined with multi-scale CNN features) was exploited in [36]. Aiming at modeling the semantic relationships among RS images in the embedding space, deep metric learning has become an important trend to effectively capture the semantic contents of RS scenes. Cheng *et al.* exploited a pairwise loss as a regularizer, together with the cross-entropy loss to improve the class-discrimination capability of CNN models [29]. Yan *et al.* adopted a deep metric learning strategy for reducing the data distribution bias in the embedding space, and further proposed a domain-adaptation method for RS scene classification [30]. In order to obtain a robust image retrieval system against variations of RS images, Yun *et al.* introduced a novel triangular loss function within a coarse-to-fine training framework in [37]. Xu *et al.* developed a sketch-based RS image retrieval (SBRSIR) [38] framework for searching images in a scalable RS database based on hand-drawn sketches. Li *et al.* proposed a meta learning-based method for few-shot RS scene classification, where the balanced loss (which maximizes the distance between different categories) was developed in [39]. For more details about deep learning-based RS scene characterization methods, we refer readers to the comprehensive reviews in [16], [40], [41].

B. Rotation-invariant CNNs

Rotation-invariant CNN models aim at equivalently categorizing the original images and their rotated versions. In other words, the inference of the labels based on those models are not sensitive to image rotations. An efficient approach to learn such transformation-equivalent CNNs is data augmentation [42]. Its basic principle is to improve the rotation-invariant capability of CNN models by generating abundant rotated training images. Marcos *et al.* presented a shallow CNN where the rotational invariance is directly encoded by tying the weights of groups of filters to several rotated versions of the canonical filter in the group [43]. Cheng *et al.* proposed an effective method to train rotation-invariant and Fisher discriminative CNN models for object detection purposes [44]. Laptev *et al.* introduced a transformation-invariant pooling operator (Ti-pooling), where a siamese network is first employed to extract features (from multiple rotated images) which are then fed through a pooling to the first fully-connected layer [45]. Chen *et al.* developed a recurrent transformer network (RTN) for learning transformation-invariant regions based on a so-called "transformer mechanism," so that the semantic gap of the subordinate level feature representations could be reduced [46]. Yang *et al.* introduced a novel object detection method, termed SCRDet, for effectively detecting small, cluttered and rotated objects in RS images [47]. In [48], He *et al.* presented the skip-connected covariance network (SCCov), which jointly exploits skip connections and covariance pooling to achieve highly representative feature learning. By applying channel attention into group convolution, Chung *et al.* proposed a rotation-invariant RS scene image retrieval method with group convolutional metric learning [49]. To learn discriminative and invariant features of RS images, Wang *et al.* adopted a siamese network to transfer the input images (using a finite transformation group, consisting of multiple confounding orthogonal matrices) into a representation space, where an invariant representation can be derived [50].

C. Novelty and Advantages of the Proposed Method

From data augmentation techniques [42] to transformed convolutional and pooling layers [43], [45], different rotation-invariant CNN-based models have shown to be effective to relieve the large-scale variance problem inherent to the presence of rotated images. However, many of these methods rely on regular CNN classification schemes, where transformed images are projected onto their corresponding label spaces without accounting for the structure of the generated deep embeddings. In this way, rotated RS images belonging to the same class may have different internal characterizations, which eventually constrains the usability of such data representations to down-stream applications beyond classification. Although some authors have been able to extend this rotation-invariant scheme to other tasks, e.g. retrieval [49], the invariance process is often implemented as part of the CNN design, which may still limit the model generality due to the need for some sort of pre-training and arbitrary weight sharing schemes, that eventually undermine the end-to-end nature of the models.

With the objective of generating more meaningful RS image representations (with higher discrimination and generalization ability), we propose dealing with the rotation-invariant problem from a novel deep metric learning perspective. That is, instead of modifying the CNN model design, we aim at developing a new deep metric learning loss to produce rotation-invariant RS image characterizations, regardless of the considered feature extraction architecture. In this way, our newly developed RiDe pursues to advance the development of rotation-invariant RS scene representations from a loss function perspective.

III. RiDe

The proposed RS image characterization method consists of two main components: 1) a backbone CNN model for extracting the deep embeddings; and 2) a new deep metric learning loss for training such model in a rotation-invariant way. Figure 2 provides a graphical illustration of the proposed framework. As it is possible to see, the backbone architecture is independent from the proposed loss, since it is only used as feature extractor. Besides, a memory bank mechanism is employed to compute all the elements required by the proposed loss. In the following, we describe all these components in detail.

A. Notations

Let $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ denote a RS image dataset containing N images, along with their category labels $\mathcal{Y}^C = \{y_1^C, \cdots, y_N^C\}$, where $y_i^C \in \{1, \cdots, C\}$. To achieve the rotation-invariant capability of CNN models, we create a rotation-augmented dataset $\tilde{\mathcal{X}}$ based on $\mathcal{X}$, where each image in $\tilde{\mathcal{X}}$ is a rotated version of the original image in $\mathcal{X}$, i.e.,

$$\tilde{\mathcal{X}} = \left\{ \begin{matrix} \mathbf{x}_1, & \text{rot90}(\mathbf{x}_1), & \text{rot180}(\mathbf{x}_1), & \text{rot270}(\mathbf{x}_1) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_N, & \text{rot90}(\mathbf{x}_N), & \text{rot180}(\mathbf{x}_N), & \text{rot270}(\mathbf{x}_N) \end{matrix} \right\} = \{\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_{4N}\}. \tag{1}$$

Here, $\text{rot90}(\cdot)$ refers to the clockwise 90 degree rotation operator. Taking the NWPU-RESISC45 [15] dataset as an example, we show some images from the original dataset and their rotation-augmented ones in Figure 3. As an extension, the category label set of $\tilde{\mathcal{X}}$ is symbolized as $\tilde{\mathcal{Y}}^C$. In addition, we denote another label set $\tilde{\mathcal{Y}}^R$, termed as *image label set*, indicating the original image index that generates the rotated versions in $\tilde{\mathcal{X}}$, i.e., $\tilde{\mathcal{Y}}^R = \{1, 1, 1, 1, \cdots, N, N, N, N\}$. $\mathbf{f}_i \in \mathbb{R}^D$ is the normalized deep embedding (i.e., $\|\mathbf{f}_i\|_2 = 1$) of image $\mathbf{x}_i$, generated via a CNN model $\mathcal{F}(\cdot)$, i.e., $\mathcal{F}(\mathbf{x}_i) = \mathbf{f}_i$.

B. Neighborhood Component Analysis (NCA)

NCA [51] is a supervised dimension reduction method that maximizes the performance of K-nearest neighbour (KNN) classification in the embedding space. Given a training set: $\{(\mathbf{x}_1, y_1^C), \cdots, (\mathbf{x}_N, y_N^C)\}$, the purpose of NCA is to learn a linear function $\mathbf{A}$ which maps the input data into a new embedding space such that each point is more likely to select

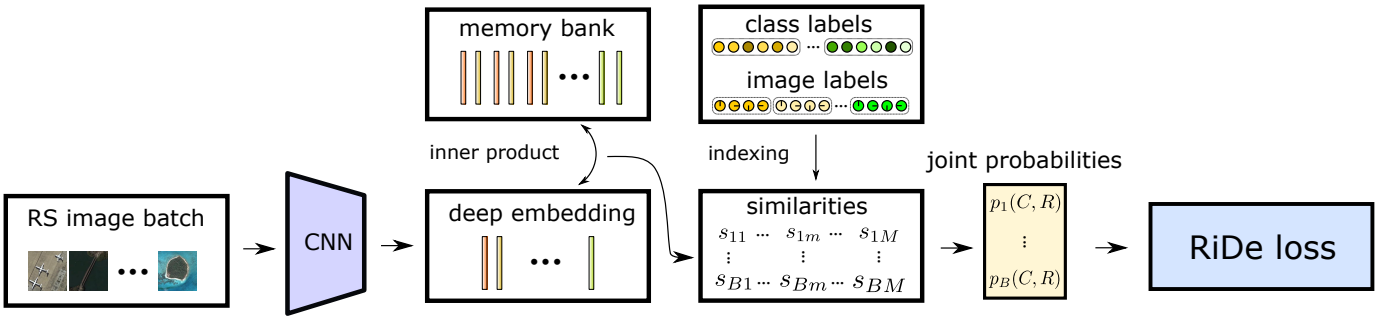


Fig. 2. Graphical illustration of the proposed rotation-invariant deep embedding method for RS images. Based on the proposed `RiDe` loss, optimized CNN models can capture the embedding space for RS images with a hierarchical structure. Not only the intraclass compactness and interclass separability can be preserved, but also the embeddings of the rotated RS images are closer to the embedding of the source image than any other images in the embedding space.

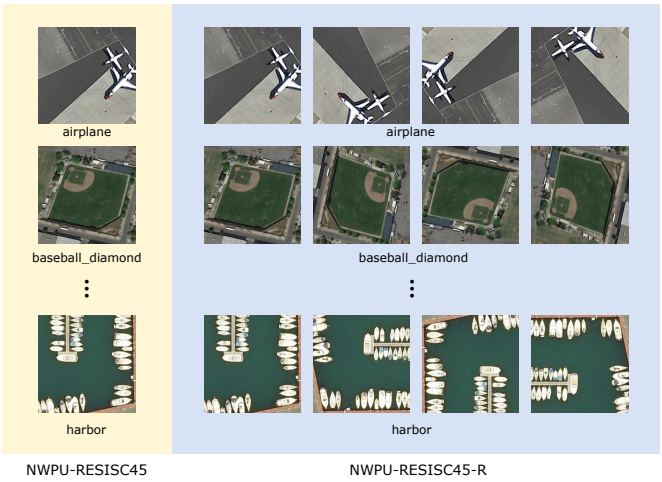


Fig. 3. The original NWPU-RESISC45 dataset and its rotation-augmented dataset (NWPU-RESISC45-R).

the ones sharing the same class as its neighbors. To achieve this, the probability that $\mathbf{x}_i$ selects $\mathbf{x}_j$ as its neighbor is defined as:

$$p_{ij} = \frac{\exp(-\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|^2)}{\sum_{k\neq i}\exp(-\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_k\|^2)}, \quad p_{ii} = 0. \tag{2}$$

Based on this, the probability that $\mathbf{x}_i$ can be correctly classified is:

$$p_i^C = \sum_{j\in\Omega_i} p_{ij}, \tag{3}$$

where $\Omega_i = \{j|y_i^C = y_j^C\}$ is the index set of the training images sharing the same class with $\mathbf{x}_i$. The goal of NCA is to maximize the log-likelihood that all images can be correctly classified, where the log-likelihood is defined as:

$$l(\mathbf{A}) = \sum_i \log(p_i^C). \tag{4}$$

Owing to its prominent capability for feature modeling, the embedding space can be better characterized by a CNN model than by a linear projection. Therefore, NCA can be extended to a deep version, the scalable neighborhood components analysis (SNCA) in [52], where the linear mapping $\mathbf{A}$ is replaced by a nonlinear mapping based on a CNN model $\mathcal{F}(\cdot)$. Moreover, the similarity measurement in the embedding space is realized by the *cosine* between the normalized embeddings of images $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e., $s_{ij} = \mathbf{f}_i^T\mathbf{f}_j$. Thus, p_{ij} is formulated as:

$$p_{ij} = \frac{\exp(\mathbf{f}_i^T\mathbf{f}_j/\sigma)}{\sum_{k\neq i}\exp(\mathbf{f}_i^T\mathbf{f}_k/\sigma)}, \quad p_{ii} = 0. \tag{5}$$

Then, SNCA aims at minimizing the following loss:

$$\begin{aligned} L_{\text{SNCA}} &= -\sum_i \log(p_i^C) \\ &= -\sum_i \log\Big(\sum_{j\in\Omega_i}\frac{\exp(\mathbf{f}_i^T\mathbf{f}_j/\sigma)}{\sum_{k\neq i}\exp(\mathbf{f}_i^T\mathbf{f}_k/\sigma)}\Big). \end{aligned} \tag{6}$$

In order to stochastically minimize the SNCA loss, a *memory bank* $\mathcal{B}$ is introduced to store the normalized embeddings of the training set serving for such contrastive learning.

C. `RiDe` Loss

1) Limitations of the SNCA loss: From Equation (6), it can be observed that an optimal solution can be reached when all the embeddings from the same class are the same, i.e., $\mathbf{f}_i^T\mathbf{f}_j = 1,\ \forall j \in \Omega_i$. In other words, the normalized embeddings from the same class are perfectly aligned with each other in the embedding space. However, such ideal case is hard to be achieved in practice due to the complex semantics of RS scenes. It indicates that there may be local structures in a set of images belonging to the same class that are represented by more than one point in the embedding space. On the other hand, as opposed to other kinds of images, RS scenes captured by earth observation sensors have no typical orientations, which means that any rotated RS image is meaningful in reality. The same scene may exhibit different locations while the only difference between them is their orientations. Therefore, an embedding space for characterizing RS images should satisfy the following structural condition: *given an anchor image in the embedding space, the distances between this image and its rotated versions should be closer than any other images, including the ones belonging to the same class.*

Since the SNCA loss only aims at grouping all the images that belong to the same class together in the embedding space, it cannot guarantee such a fine-grained structural condition.

This means that the trained CNN models do not possess rotation-invariant capabilities for generating deep embeddings. To address this issue, we develop a new loss function which is rotation invariant and preserves the class-discrimination capability.

2) Definition of the `RiDe` *loss:* In order to achieve the aforementioned goals, we make use of a rotation-augmented dataset $\tilde{\mathcal{X}}$. Note that the use of an augmented dataset logically brings some computational burden to the training stage. However, it is important to highlight that the asymptotic cost of processing $\tilde{\mathcal{X}}$ remains the same (compared with the original dataset) since the increase of the number of samples is limited by the amount of considered rotations, which is a constant value that does not depend on the database size. This situation is not exclusive to the proposed approach, but to any other method using augmented data. Besides, it does not affect the operational exploitation of the proposed model. Given a training set $\{(\tilde{\mathbf{x}}_1, \tilde{y}_1^C, \tilde{y}_1^R), \cdots, (\tilde{\mathbf{x}}_{4N}, \tilde{y}_{4N}^C, \tilde{y}_{4N}^R)\}$, p_{ij} denotes the probability that image $\mathbf{x}_i$ selects $\mathbf{x}_j$ as its neighbor, as defined in Equation (5). Likewise, p_i^C measures the probability that $\mathbf{x}_i$ is correctly classified:

$$p_i^C = \sum_{j \in \tilde{\Omega}_i} p_{ij}, \tag{7}$$

where $\tilde{\Omega}_i = \{j | \tilde{y}_i^C = \tilde{y}_j^C\}$. In addition, we define another probability p_i^R calculating the likelihood of its rotated counterparts lying nearby in the embedding space:

$$p_i^R = \sum_{j \in \tilde{\mathcal{R}}_i} p_{ij}, \tag{8}$$

where $\tilde{\mathcal{R}}_i = \{j | \tilde{y}_i^R = \tilde{y}_j^R\}$ is the index set of the rotated training images coming from the same source image. As discussed above, we aim at achieving both class discrimination and rotational invariance when extracting deep embeddings from RS images. Hereby, a joint probability $p_i(C, R)$ is introduced, which not only represents the likelihood that the image $\mathbf{x}_i$ is correctly classified, but also that its rotated images are located nearest to itself in the embedding space. To simplify the calculation of such joint probability, we assume that both cases are independent. Therefore, it can be formulated as follows:

$$p_i(C, R) = p_i^C p_i^R = \Big(\sum_{j \in \tilde{\Omega}_i} p_{ij} \Big) \Big(\sum_{j \in \tilde{\mathcal{R}}_i} p_{ij} \Big). \tag{9}$$

To maximize such joint probability over the whole training set, we equally minimize the following negative log-likelihood:

$$L = -\sum_i \log \big(p_i(C, R) \big). \tag{10}$$

According to the properties of logarithms, Equation (10) can be further expanded as:

$$\begin{aligned} L &= -\sum_i \log(p_i^C) - \sum_i \log(p_i^R) \\ &= \underbrace{-\sum_i \log \Big(\sum_{j \in \tilde{\Omega}_i} \frac{\exp(\mathbf{f}_i^T \mathbf{f}_j / \sigma)}{\sum_{k \neq i} \exp(\mathbf{f}_i^T \mathbf{f}_k / \sigma)} \Big)}_{\text{class discrimination}} \\ &\quad \underbrace{-\sum_i \log \Big(\sum_{j \in \tilde{\mathcal{R}}_i} \frac{\exp(\mathbf{f}_i^T \mathbf{f}_j / \sigma)}{\sum_{k \neq i} \exp(\mathbf{f}_i^T \mathbf{f}_k / \sigma)} \Big)}_{\text{rotational invariance}}. \end{aligned} \tag{11}$$

From a loss function perspective, there are two different terms in Equation (11): 1) class discrimination, which is optimized for pulling intraclass embeddings together while pushing interclass embeddings away; 2) rotational invariance, which is optimized for grouping together the embeddings of the rotated images obtained from the same source image. In order to better balance these two terms, a penalty parameter λ is introduced, and the final `RiDe` loss is formulated as:

$$L_{\text{RiDe}} = -\sum_i \log(p_i^C) - \lambda \sum_i \log(p_i^R). \tag{12}$$

In fact, `RiDe` can be considered as a rotation-invariant generalization of SNCA. When $\lambda = 0$, `RiDe` turns into SNCA.

D. Optimization

Based on the back-propagation technique, the gradients of $L_{\texttt{RiDe}}$ with respect to the parameters in CNN models can be obtained. To stochastically minimize $L_{\texttt{RiDe}}$, we exploit the memory bank $\mathcal{B}$ to store all the normalized embeddings of the training images. After each training iteration, $\mathcal{B}$ is updated in an empirical weighted averaging manner:

$$\mathbf{f}_i^{(t+1)} \leftarrow m \mathbf{f}_i^{(t)} + (1 - m) \mathbf{f}_i, \tag{13}$$

where m is a parameter controlling the balance between the two embeddings. The associated optimization scheme is described in Algorithm 1.

Algorithm 1 `RiDe`

Require: $\tilde{\mathcal{X}}$, $\tilde{\mathcal{Y}}^C$ and $\tilde{\mathcal{Y}}^R$

1: Initialize CNN parameters and $\mathcal{B}$ (randomly), along with σ, λ, D and m.
2: **for** $t = 0$ to maxEpoch **do**
3: Sample a mini-batch.
4: Obtain $\mathbf{f}_i^{(t)}$ based on $\mathcal{F}(\cdot)$.
5: Calculate the similarities s_{ij} based on the extracted mini-batch embeddings and those in $\mathcal{B}$.
6: Index the similarities based on $\tilde{\mathcal{Y}}^C$ and $\tilde{\mathcal{Y}}^R$.
7: Calculate the `RiDe` loss in Equation (12).
8: Back-propagate the gradients.
9: Update $\mathcal{B}$ via (13).
10: **end for**

Ensure: $\mathcal{F}(\cdot)$

E. Complexity Analysis

With an embedding size of D and a whole number of rotated images $4N$, the memory bank $\mathcal{B}$ requires $\mathcal{O}(DN)$ of memory. Suppose that the batch size is B. In this case, the similarity metric and the probability density require $\mathcal{O}(BN)$ of memory, and the other intermediate variables occupy $\mathcal{O}(BN)$ of memory.

IV. Experiments

A. Experimental Setup

1) Dataset configuration: In our experiments, we use two RS scene benchmark datasets: 1) Aerial Image Dataset (AID) [40] and 2) NWPU-RESISC45 [15]. For additional details about the datasets, we refer the readers to the associated papers. As introduced above, the proposed method exploits rotation-augmented datasets. Thus, we expand the original datasets by adding the rotated versions of each image with 90, 180 and 270 degrees, and create rotation-augmented (AID-R and NWPU-RESISC45-R) versions of the original datasets. From the original datasets, we first randomly select 70%, 10% and 20% of the available images for training, validation and testing, respectively. Then, we associate the rotated versions of each source image into the corresponding sets to construct the splitting of AID-R and NWPU-RESISC45-R. In order to evaluate the effectiveness of the proposed method, we carry out KNN classification and image retrieval tasks based on the extracted deep embeddings.

1) *KNN classification* aims at classifying the input image based on its k nearest neighbors in the embedding space, whose class is dependent on its neighbors' classes via a majority voting. To evaluate the classification performance, we adopt the overall accuracy and confusion matrix.
2) *Image retrieval* aims at effectively finding the most semantically similar images in a database given a query image, ranking them based on the similarities measured in the embedding space. The evaluation is based on mean average precision (MAP) and Recall@k (R@k). MAP is defined with the form:

$$\mathrm{AP} = \frac{1}{Q}\sum_{r=1}^{R} P(r)\delta(r), \tag{14}$$

where Q is the number of ground-truth RS images in the database 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. R@k is defined as the percentage of queries having at least one relevant image retrieved among the top k results.

Since `RiDe` aims to generate deep embeddings of images with rotational invariance and class-discrimination, we design two different scenarios for our experimental evaluation:

1) *Rotated image identification*: Given the test sets of AID-R and NWPU-RESISC45-R, we utilize the associated image label set $\tilde{\mathcal{Y}}^R$ to obtain the ground-truth labels. Based on the deep embedding of each test image, we check whether it can retrieve those of the other rotated images as its nearest neighbors in the embedding space and evaluate the associated performance. This scenario is designed for validating the rotation-invariant capability of the trained CNN models. For KNN classification, we split the deep embeddings of the test sets with 5 folds, where the training-to-test ratio is $3/1$. For image retrieval, each image from the test set is used for query purposes, and the other ones are used as the database.
2) *Class-wise image discrimination*: Given the test sets of AID and NWPU-RESISC45, we utilize the associated category label set $\mathcal{Y}^C$ to obtain the ground-truth labels. Based on the deep embedding of each test image, we evaluate whether its neighbors in the embedding space belong to the same class. This scenario is designed for testing the class-discrimination power of the trained CNN models. For KNN classification, the training sets are utilized for training the KNN classifier and the test sets are exploited for the prediction. For image retrieval, each image from the test set is used for query purposes, and the images in the training set are used as the database.

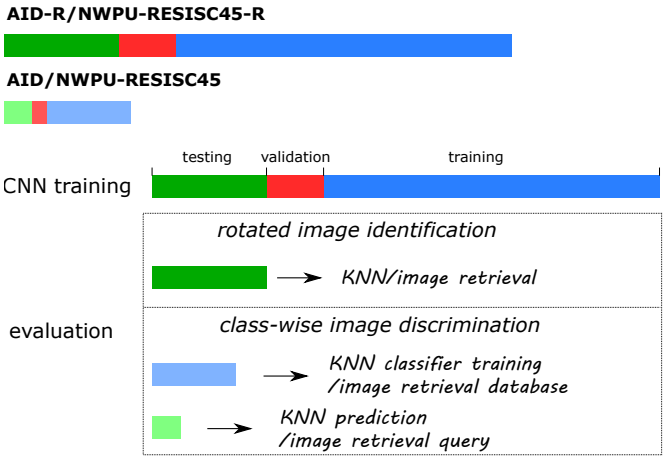


Fig. 4. A simplified graphical illustration describing the configuration of datasets in our experiments

A simplified graphical illustration describing the configuration of datasets in our experiments is displayed in Figure 4.

2) Implementation details: We utilize ResNet34 and ResNet50 [53] as the CNN backbones to extract the deep embeddings of the input images. The spatial size of the input images is 256×256, and they are augmented by 1) RandomGrayscale; 2) ColorJitter; and 3) RandomHorizontalFlip. The parameters D, σ, λ and m are set to 128, 0.1, 0.1 and 0.5, respectively. We utilize the Stochastic Gradient Descent (SGD) optimizer to train the CNN models with an initial learning rate of 0.1 and a decay rate of 0.5 every 30 epochs. We train the networks for a total of 100 epochs. Due to the limitations of memory in the graphical processing unit (GPU) used in experiments, the batch size is 256 for ResNet34 and 128 for ResNet50. To validate the effectiveness of the proposed method, we compare it to several state-of-the-art

deep embedding methods from two perspectives:

1) *Rotated image identification*: for evaluating the rotation-invariant capability, we compare RiDe to 1) **SNCA** [52], 2) **SNCA-aug**, where rotation-based data augmentation is used for training the CNN models, and 3) **SCCov** [48].
2) *Class-wise image discrimination*: for validating the preservation of the class-discrimination capability, we compare RiDe to 1) **Triplet** [54], 2) Normalized Softmax Loss (**NSL**) [55], 3) **ArcFace** [56] , 4) **SCCov** [48] and 5) **TI-POOLING** [45].

All the experiments are implemented in PyTorch [57] and carried out on an NVIDIA RTX3090 GPU.

B. Experimental Results

1) Rotated image identification: Table I presents the KNN classification results for the rotated images based on the deep embeddings extracted from the test sets. Since one source image generates four rotated versions, we report KNN classification results with $K = 1, 2, 3$. This experiment is intended to analyze whether all the deep embeddings of rotated images are located close to each other in the embedding space. From Table I, it can be observed that with the vanilla ResNets, RiDe achieves the best performance (with a value near 100%) on the two considered benchmark datasets. Compared to SNCA, the inclusion of the proposed rotation-invariant term in RiDe can significantly improve the rotation-invariant capability of the trained CNN models in the generation of deep embeddings. Without this term, the KNN classification performance drops more than 10% in SNCA. Although data augmentation is an efficient approach to improve the rotation-invariant capability, RiDe generally outperforms SNCA-aug by more than 5% in the KNN classification results. With the state-of-the-art CNN architecture (SCCov) for scene classification, the use of RiDe can greatly improve the performance of rotated image identification compared with the Cross Entropy (CE) loss utilized in [48]. It is worth noting that TI-POOLING is not considered in this experiment since all the rotated images have the same embedding produced by TI-POOLING. Thus, the accuracy of TI-POOLING will be 100%. Moreover, TI-POOLING is a kind of rotation-invariant CNN architecture, while the proposed RiDe is a novel loss targeted at learning the embeddings invariant to the rotation of the input images which can be combined with any CNN architecture. However, one disadvantage of TI-POOLING is that the computational cost will be increased due to the feature aggregation from multiple input images. As shown in Table II, we illustrate the computational cost study of both RiDe and TI-POOLING for both training and inference phases. It can be observed that TI-POOLING will spend more time for learning and extracting deep embeddings than RiDe. Table III displays the image retrieval results evaluated by both MAP and R@k when $R = 1, 2, 3$ and $k = 1, 2, 3$. Consistently with the KNN classification results, RiDe exhibits superior performance when compared to other methods. It can be seen from the obtained results that all the deep embeddings of the rotated images are located close to each other in the embedding space generated using the proposed method. To visually verify such observation, given some query images, we display their three nearest neighbors retrieved from the test sets of AID-R and NWPU-RESISC45-R in Figure 5. Without the penalty on learning rotation-invariant deep embeddings, the nearest neighbors retrieved by the SNCA are not always from the same source image (images marked in red color in Figure 5). Although their class labels are the same, their semantics may exhibit large divergences (this can be seen, for instance, in the first row of Figure 5). By utilizing a data augmentation strategy, SNCA-aug indeed increases the semantic similarities of the nearest neighbors associated to the query as compared to SNCA. However, SNCA-aug cannot perfectly retrieve all the rotated images from the test sets (marked in green color in Figure 5), given the query. In this regard, RiDe ranks all the rotated images with the highest similarity in the database with respect to the input query image. Similar performances can be also observed when SCCov is utilized as the CNN architecture. Therefore, we conclude that RiDe not only guides CNN models to learn rotation-invariant deep embeddings, but also models better the semantic similarities among the images.

2) Class-wise image discrimination: Table IV presents the KNN classification results for class-wise image discrimination based on the extracted deep embeddings from the test sets of AID and NWPU-RESISC45, when $K = 1, 5, 10$. Compared to the other losses, RiDe achieves the best accuracies with different values of K. When the CNN backbone changes from ResNet34 to a more powerful network, i.e., ResNet50, the other methods improve their classification performance on the NWPU-RESISC45 dataset. For example, Triplet, NSL and ArcFace exhibit about $1\% - 3\%$ increase in accuracy when the network is changed from ResNet34 to ResNet50. On the contrary, the associated accuracy differences for RiDe are less than 1%. This suggests that RiDe can generate high-quality deep embeddings based on both light-weight and powerful CNN architectures. Compared to the state-of-the-art deep embedding method, i.e., ArcFace, RiDe can better uncover the local neighborhood structure of the input images for class discrimination purposes. Compared with the CE loss exploited in SCCov [48] and TI-POOLING [45], the adoption of RiDe can lead to higher KNN-based classification accuracy. Moreover, we display the normalized confusion matrices obtained for the KNN classification via RiDe on the two test sets in Figure 8. It can be observed that most classes can be correctly discriminated in the two considered datasets. Nevertheless, there are several classes that are misclassified, e.g., *Resort* and *Park* in AID. In order to evaluate the image retrieval performance, we calculated the MAP scores based on the deep embeddings extracted from the test sets when $R = 20, 50, 100$ and display them in Table V. It can be observed that the semantic relations among the images can be best captured through RiDe, regardless of the exploited CNN architectures, and the retrieval performance can be still preserved as the number of retrieved images increases. Therefore, we conclude that our newly proposed method can be applied for accurately indexing large RS databases.

3) Other rotational angles: We conduct the previous experiments on the datasets which totally include four ro-

TABLE I
KNN-BASED CLASSIFICATION RESULTS FOR ROTATED IMAGE IDENTIFICATION, BASED ON THE EXTRACTED DEEP EMBEDDINGS OF THE TEST SETS WITH 5 FOLDS. ($K = 1, 2, 3$)

		AID-R			NWPU-RESISC45-R		
		$K=1$	$K=2$	$K=3$	$K=1$	$K=2$	$K=3$
ResNet34	SNCA [52]	87.72 ± 0.60	78.23 ± 0.16	78.17 ± 0.54	81.15 ± 0.41	71.04 ± 0.41	70.77 ± 0.64
	SNCA-aug	94.58 ± 0.19	88.16 ± 0.45	88.16 ± 0.41	90.85 ± 0.14	84.08 ± 0.26	83.87 ± 0.34
	RiDe	$\mathbf{99.54 \pm 0.10}$	$\mathbf{99.66 \pm 0.06}$	$\mathbf{99.67 \pm 0.10}$	$\mathbf{99.81 \pm 0.04}$	$\mathbf{99.55 \pm 0.05}$	$\mathbf{99.67 \pm 0.06}$
ResNet50	SNCA [52]	95.58 ± 0.22	91.56 ± 0.41	91.32 ± 0.18	94.13 ± 0.17	88.91 ± 0.22	88.38 ± 0.16
	SNCA-aug	98.65 ± 0.18	96.75 ± 0.24	96.76 ± 0.24	98.06 ± 0.08	95.51 ± 0.14	95.21 ± 0.14
	RiDe	$\mathbf{99.83 \pm 0.06}$	$\mathbf{99.86 \pm 0.04}$	$\mathbf{99.93 \pm 0.05}$	$\mathbf{99.90 \pm 0.02}$	$\mathbf{99.83 \pm 0.03}$	$\mathbf{99.91 \pm 0.04}$
SCCov[ResNet34]	CE [48]	54.45 ± 0.98	46.71 ± 1.29	46.50 ± 0.67	58.48 ± 0.24	49.99 ± 0.38	50.39 ± 0.46
	RiDe	$\mathbf{99.72 \pm 0.06}$	$\mathbf{99.78 \pm 0.05}$	$\mathbf{99.82 \pm 0.05}$	$\mathbf{99.93 \pm 0.02}$	$\mathbf{99.89 \pm 0.02}$	$\mathbf{99.96 \pm 0.02}$
SCCov[ResNet50]	CE [48]	42.99 ± 0.91	36.54 ± 0.62	36.63 ± 1.13	57.20 ± 0.25	49.65 ± 0.22	49.75 ± 0.34
	RiDe	$\mathbf{99.98 \pm 0.02}$	$\mathbf{99.95 \pm 0.03}$	$\mathbf{100 \pm 0.00}$	$\mathbf{99.98 \pm 0.02}$	$\mathbf{99.56 \pm 0.08}$	$\mathbf{99.68 \pm 0.05}$

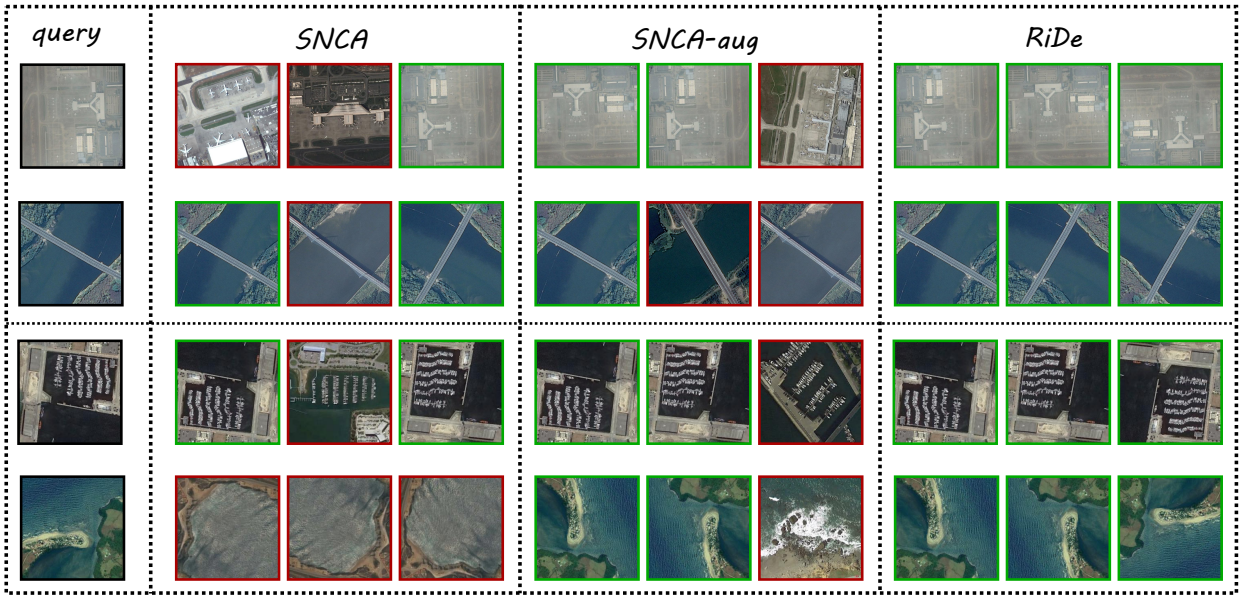


Fig. 5. For different query images, we display the three nearest neighbors retrieved from the test sets of AID-R (first two rows) and NWPU-RESISC45-R (last two rows), based on the deep embeddings extracted from the trained ResNet50 via SNCA, SNCA-aug and RiDe. The images not rotated from the same source image are marked in red color, while the images rotated from the same source image are marked in green color.

TABLE II
THE COMPUTATIONAL COST STUDY OF RiDe AND TI-POOLING FOR BOTH THE TRAINING AND INFERENCE PHASES (SECOND PER IMAGE).

		Train	Inference
ResNet34	RiDe	2.25×10^{-3}	2.00×10^{-3}
	TI-POOLING	8.43×10^{-3}	5.15×10^{-3}
ResNet50	RiDe	3.29×10^{-3}	2.15×10^{-3}
	TI-POOLING	13.87×10^{-3}	6.72×10^{-3}

tational angles, i.e., $[0°, 90°, 180°, 270°]$. In order to verify the performance of RiDe on other rotational angles, we rotate the original images by the angles of $[0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°]$ and construct the associated datasets. Figure 6 illustrates the KNN-based classification results for rotated image identification when $K = 1, 2, 3, 4, 5, 6, 7, 8$. It is observed that RiDe can be generalized to more rotational angles and the nearest neighbor performance is well preserved.

4) Hyperparameter analysis: Two parameters: λ and σ should be carefully tuned for achieving good performance using RiDe. λ controls the balance between the class-discrimination and rotation-invariant terms in RiDe, and $1/\sigma$ represents the radius of the hypersphere on which the embeddings are projected. With a larger value of $1/\sigma$, the embedding hyperspace can be more appropriate for class discrimination [58], [59]. Taking [52], [60] into account, we empirically set it as a small number, e.g., $\sigma = 0.1$, and keep it constant in our experiments. In order to analyze the influence of different values of λ on the performance of RiDe, we conduct KNN classification for both rotated image identification and class-wise image discrimination when λ is in the range $[0.05, 0.1, 0.5, 1]$. We adopt the ResNet50 architecture and plot the classification results in Figure 7. As λ increases,

TABLE III
IMAGE RETRIEVAL METRICS (MAP AND R@k) FOR ROTATED IMAGE IDENTIFICATION BASED ON THE EXTRACTED DEEP EMBEDDINGS OF THE TEST SETS ($R = 1, 2, 3$ FOR MAP AND $k = 1, 2, 3$ FOR R@k).

		MAP					
		AID-R			**NWPU-RESISC45-R**		
		$R = 1$	$R = 2$	$R = 3$	$R = 1$	$R = 2$	$R = 3$
ResNet34	SNCA [52]	86.62	89.06	88.73	79.29	82.30	82.24
	SNCA-aug	93.94	95.06	94.62	89.73	91.63	91.31
	RiDe	**99.58**	**99.75**	**99.75**	**99.72**	**99.83**	**99.78**
ResNet50	SNCA [52]	95.17	95.99	95.69	93.72	94.62	94.31
	SNCA-aug	98.56	98.91	98.74	97.73	98.13	97.93
	RiDe	**99.85**	**99.92**	**99.93**	**99.90**	**99.94**	**99.92**
SCCov[ResNet34]	CE [48]	51.25	56.74	58.06	55.11	60.43	61.50
	RiDe	**99.73**	**99.85**	**99.85**	**99.92**	**99.96**	**99.95**
SCCov[ResNet50]	CE [48]	39.22	45.10	46.80	53.94	59.75	60.98
	RiDe	**99.99**	**99.99**	**99.98**	**99.76**	**99.85**	**99.81**
		R@k					
		$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
ResNet34	SNCA [52]	86.62	91.51	93.41	79.29	85.31	88.30
	SNCA-aug	93.94	96.17	97.34	89.73	93.53	95.09
	RiDe	**99.58**	**99.91**	**100.00**	**99.72**	**99.93**	**99.97**
ResNet50	SNCA [52]	95.17	96.82	97.59	93.72	95.52	96.33
	SNCA-aug	98.56	99.27	99.54	97.73	98.53	98.88
	RiDe	**99.85**	**99.99**	**100.00**	**99.90**	**99.97**	**99.99**
SCCov[ResNet34]	CE [48]	51.25	62.23	68.72	55.12	65.76	71.86
	RiDe	**99.73**	**99.98**	**99.99**	**99.92**	**100**	**100**
SCCov[ResNet50]	CE [48]	39.22	45.10	46.80	53.94	59.75	60.99
	RiDe	**99.99**	**99.99**	**99.98**	**99.76**	**99.85**	**99.81**

TABLE IV
KNN RESULTS FOR CLASS-WISE IMAGE DISCRIMINATION BASED ON THE DEEP EMBEDDINGS EXTRACTED FROM THE TEST SETS ($K = 1, 5, 10$).

		AID			**NWPU-RESISC45**		
		$K = 1$	$K = 5$	$K = 10$	$K = 1$	$K = 5$	$K = 10$
ResNet34	Triplet [54]	94.24	94.24	94.49	92.79	93.02	92.79
	NSL [55]	92.70	92.50	92.65	92.37	92.17	92.25
	ArcFace [56]	95.63	95.53	95.58	92.94	92.97	92.92
	RiDe	**96.08**	**96.22**	**96.52**	**95.22**	**95.33**	**95.32**
ResNet50	Triplet [54]	94.68	94.83	94.93	93.68	94.05	93.97
	NSL [55]	94.98	95.18	95.08	94.63	94.49	94.43
	ArcFace [56]	95.98	95.78	95.88	95.46	95.46	95.43
	RiDe	**96.42**	**96.67**	**96.57**	**96.11**	**96.08**	**95.98**
SCCov[ResNet34]	CE [48]	94.24	94.49	94.39	94.13	94.33	94.32
	RiDe	**95.88**	**96.13**	**96.17**	**95.35**	**95.35**	**95.30**
SCCov[ResNet50]	CE [48]	94.09	94.34	94.44	95.14	95.11	95.05
	RiDe	**96.17**	**96.08**	**96.03**	**95.83**	**95.86**	**95.84**
TI-POOLING[ResNet34]	CE [45]	96.92	96.92	**97.17**	95.38	95.29	95.32
	RiDe	**97.02**	**97.07**	97.07	**96.02**	**96.05**	**96.05**
TI-POOLING[ResNet50]	CE [45]	**97.42**	**97.47**	**97.47**	96.05	96.06	96.11
	RiDe	97.32	97.32	97.42	**96.35**	**96.38**	**96.41**

TABLE V
MAP FOR CLASS-WISE IMAGE DISCRIMINATION BASED ON THE DEEP EMBEDDINGS EXTRACTED FROM THE TEST SETS ($R = 20, 50, 100$).

		AID			NWPU-RESISC45		
		$R = 20$	$R = 50$	$R = 100$	$R = 20$	$R = 50$	$R = 100$
ResNet34	Triplet [54]	94.36	93.99	93.63	92.83	92.19	91.74
	NSL [55]	92.42	91.29	90.18	92.47	91.69	91.05
	ArcFace [56]	95.58	95.51	95.48	93.13	93.09	93.07
	RiDe	**96.34**	**96.02**	**95.62**	**95.42**	**95.22**	**95.06**
ResNet50	Triplet [54]	95.21	94.79	94.43	93.92	93.40	93.06
	NSL [55]	94.96	94.37	93.74	94.53	94.06	93.66
	ArcFace [56]	96.01	95.93	95.88	95.50	95.47	95.45
	RiDe	**96.57**	**96.31**	**96.04**	**96.22**	**96.00**	**95.84**
SCCov[ResNet34]	CE [48]	94.65	94.36	94.11	94.38	94.14	93.96
	RiDe	**96.18**	**95.92**	**95.59**	**95.46**	**95.19**	**95.03**
SCCov[ResNet50]	CE [48]	94.46	94.24	94.07	95.24	95.09	94.99
	RiDe	**96.17**	**95.93**	**95.68**	**95.90**	**95.70**	**95.54**
TI-POOLING[ResNet34]	CE [45]	96.91	96.55	96.21	95.51	95.15	94.90
	RiDe	**97.11**	**97.10**	**97.07**	**96.07**	**96.04**	**96.03**
TI-POOLING[ResNet50]	CE [45]	**97.49**	97.28	97.10	96.15	95.92	95.75
	RiDe	97.43	**97.42**	**97.41**	**96.41**	**96.40**	**96.38**

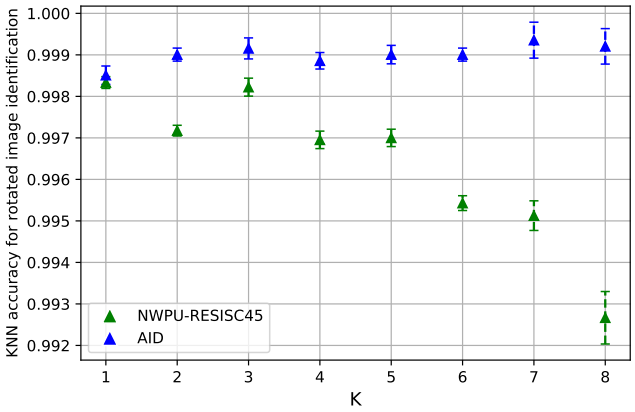


Fig. 6. KNN-based classification of rotated image identification when $K = 1, 2, 3, 4, 5, 6, 7, 8$ on the two rotation-augmented datasets with the angles of $[0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°]$.

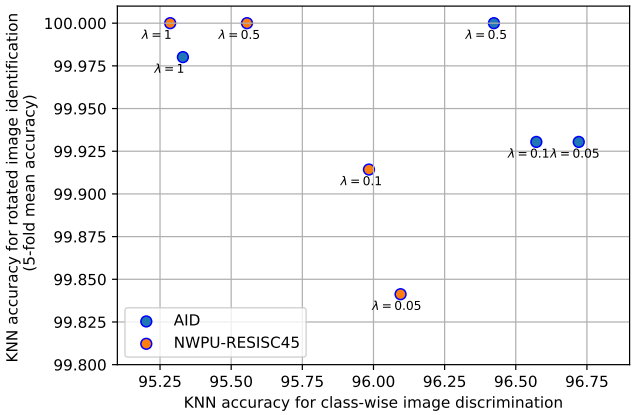


Fig. 7. Sensitivity analysis of parameter λ based on KNN classification of both rotated image identification and class-wise image discrimination, when $\lambda = 0.05, 0.1, 0.5, 1$.

the performance of rotated image identification also improves, since a larger penalty is on the rotation-invariant term of RiDe. On the contrary, when λ decreases, more emphasis is given to the class-discrimination term of RiDe, so the performance for class-wise image discrimination gets better. Therefore, to achieve a balanced performance in terms of both rotational invariance and class discrimination, λ should not be too small or too large. In our experiments, we empirically set λ to 0.1 and achieve good performance.

5) Discussion: We conducted extensive experiments from two perspectives (including rotated image identification and class-wise image discrimination) to validate the performance of RiDe. Our experiments are specifically designed to test the performance of the rotation-invariant and class-discrimination capabilities of the trained CNN models, respectively. Based on our experimental results, we can observe that RiDe significantly improves the rotation-invariant capability of the trained CNN models, since all the rotated images are located nearby each other in the embedding space. Moreover, differently to other rotation-invariant deep learning methods (e.g., TI-POOLING), RiDe is a loss function designed for learning rotation-invariant image embeddings, which can be applied to any CNN architectures. The experiments also indicate that RiDe can be applied with any rotation angles. Moreover, RiDe also achieves better class-discrimination results than the other losses. This indicates that the class-discrimination capability of the CNN models trained by RiDe can be also preserved. Therefore, RiDe can actually discover the hierarchical structure of the embedding space for RS images, where the nearest neighbors of query images are their rotated versions,

the next nearest neighbors are the images from the same class, and the images in different classes are well separated. Such hierarchical structure of the semantic relationships among the RS images is very important for the downstream task. For example, an image retrieval system requires to accurately and effectively find the most semantically similar images within the database given the query images. However, sometimes semantic similarities cannot be precisely modeled by category labels. Without the fine-grained semantic information required to categorize the images in the same class, the retrieved ranking order with respect to the input query image may not reflect the actual order of semantic similarities. In this case, by better modeling the similarities of rotated images, `RiDe` exploits an auxiliary task to model the fine-grained structure of the embedding space, which better fulfills the requirement of image retrieval systems.

V. Conclusions

In this paper, we introduce a new loss function for learning rotation-invariant deep embeddings of RS images. Specifically, we first review the limitations of the SNCA loss function when constructing the hierarchical structure of the images in the embedding space. Instead of just maximizing the leave-one-out classification probability, we introduce a joint probability for correctly classifying each image based on its neighbors and identifying its rotated images as the nearest ones. To maximize such probability, we assume that both cases are independent, which further leads to a loss function composed of two terms: 1) class-discrimination term, and 2) rotation-invariant term. To balance these two terms, we introduce a penalty parameter and finally propose the new `RiDe` loss function. We carry out extensive experiments on two RS benchmark datasets and compare `RiDe` to other state-of-the-art losses. Our experimental results validate the effectiveness of `RiDe` in the task of ensuring that CNN models exhibit both rotation-invariant and class-discrimination capabilities. As future work, we plan to reconsider the maximization of joint probabilities into a Bayesian framework, as well to extend the proposed approach to other kinds of transformations and data modalities.

Acknowledgement

The authors gratefully thank the Associate Editor and the two Anonymous Reviewers for their outstanding comments and suggestions, that greatly helped us to improve the technical quality and presentation of our work.

Appendix

In Figure 8, we demonstrate the normalized confusion matrices for the KNN classification ($K = 10$) of the test sets of the two benchmark datasets, based on the deep embeddings obtained by `RiDe`.

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

[2] 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 Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3377–3390, 2019.

[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] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.

[5] Y. Xu and U. Stilla, "Towards building and civil infrastructure reconstruction from point clouds: A review on data and key techniques," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2857–2885, 2021.

[6] X. Sun, A. Shi, H. Huang, and H. Mayer, "Basnet: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5398–5413, 2020.

[7] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang *et al.*, "So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 3, pp. 76–89, 2020.

[8] J. Li, Z. He, J. Plaza, S. Li, J. Chen, H. Wu, Y. Wang, and Y. Liu, "Social media: New perspectives to improve remote sensing for emergency response," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1900–1912, 2017.

[9] R. Hang, F. Zhou, Q. Liu, and P. Ghamisi, "Classification of hyperspectral images via multitask generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[10] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided cnns," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[11] A. C. Braun, "More accurate less meaningful? a critical physical geographer's reflection on interpreting remote sensing land-use analyses," *Progress in Physical Geography: Earth and Environment*, p. 0309133321991814, 2021.

[12] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[13] N. Yokoya, K. Yamanoi, W. He, G. Baier, B. Adriano, H. Miura, and S. Oishi, "Breaking limits of remote sensing by deep learning from simulated data for flood and debris-flow mapping," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[14] R. Fernandez-Beltran, F. Pla, J. Kang, J. Moreno, and A. Plaza, "Sentinel-3/FLEX Biophysical Product Confidence using Sentinel-2 Land Cover Spatial Distributions," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.

[15] 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.

[16] 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," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

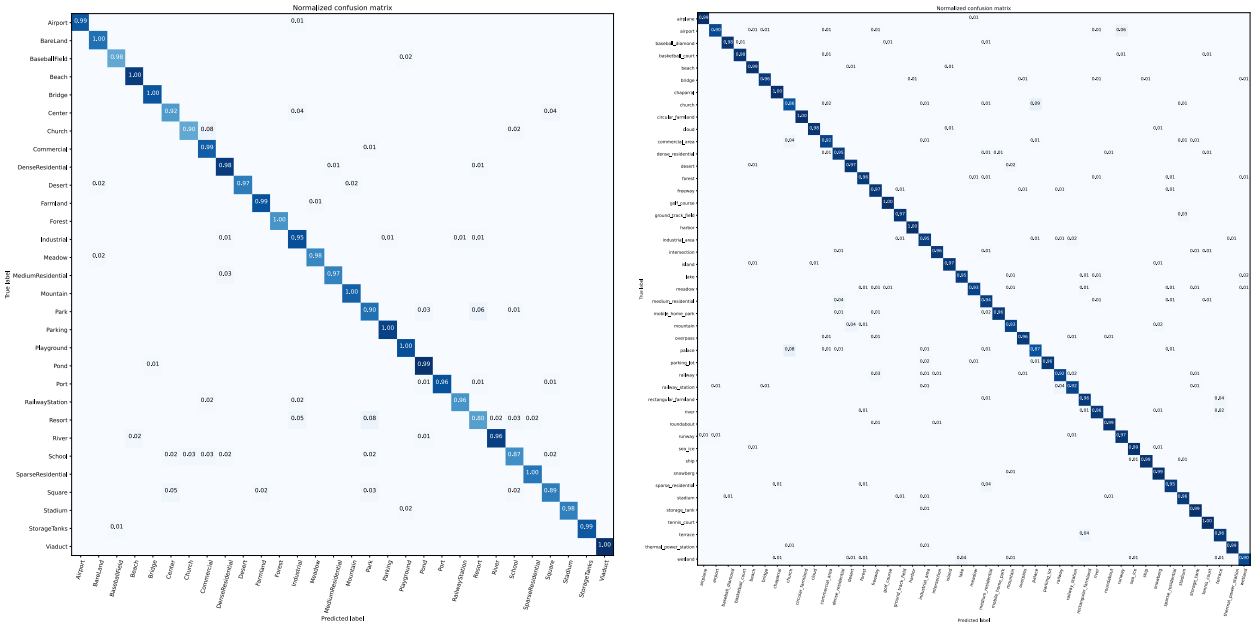
[17] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Information Fusion*, vol. 67, pp. 94–115, 2021.

[18] 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," *International journal of remote sensing*, vol. 31, no. 6, pp. 1453–1470, 2010.

[19] 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.

[20] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[21] X. Huang, L. Zhang, and L. Wang, "Evaluation of morphological texture features for mangrove forest mapping and species discrimination using multispectral ikonos imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 3, pp. 393–397, 2009.



(a) (b)

Fig. 8. Normalized confusion matrices for the KNN classification ($K = 10$) of the test sets of (a) AID and (b) NWPU-RESISC45, based on the deep embeddings obtained by `RiDe`.

[22] L. Huang, C. Chen, W. Li, and Q. Du, "Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors," *Remote Sensing*, vol. 8, no. 6, p. 483, 2016.

[23] G. Cheng, J. Han, L. Guo, X. Qian, P. Zhou, X. Yao, and X. Hu, "Object detection in remote sensing imagery using a discriminatively trained mixture model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 85, pp. 32–43, 2013.

[24] 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.

[25] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *International journal of remote sensing*, vol. 33, no. 8, pp. 2395–2412, 2012.

[26] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5148–5157, 2017.

[27] 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.

[28] 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, 2017.

[29] 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.

[30] 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*, vol. 57, no. 6, pp. 3840–3857, 2019.

[31] 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.

[32] 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 Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 62–81, 2020.

[33] F. Anselmi, G. Evangelopoulos, L. Rosasco, and T. Poggio, "Symmetry-adapted representation learning," *Pattern Recognition*, vol. 86, pp. 201–208, 2019.

[34] 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.

[35] 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.

[36] 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.

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

[38] F. Xu, W. Yang, T. Jiang, S. Lin, H. Luo, and G.-S. Xia, "Mental retrieval of remote sensing images via adversarial sketch-image feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7801–7814, 2020.

[39] H. Li, Z. Cui, Z. Zhu, L. Chen, J. Zhu, H. Huang, and C. Tao, "Rs-metanet: Deep meta metric learning for few-shot remote sensing scene classification," *arXiv preprint arXiv:2009.13364*, 2020.

[40] 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.

[41] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *National Science Review*, vol. 6, no. 6, pp. 1082–1086, 2019.

[42] P. Y. Simard, D. Steinkraus, J. C. Platt *et al.*, "Best practices for convolutional neural networks applied to visual document analysis." in *Icdar*, vol. 3, no. 2003. Citeseer, 2003.

[43] D. Marcos, M. Volpi, and D. Tuia, "Learning rotation invariant convolutional filters for texture classification," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2012–2017.

[44] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 265–278, 2018.

[45] D. Laptev, N. Savinov, J. M. Buhmann, and M. Pollefeys, "Ti-pooling: transformation-invariant pooling for feature learning in convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 289–297.

[46] Z. Chen, S. Wang, X. Hou, L. Shao, and A. Dhabi, "Recurrent transformer network for remote sensing scene categorisation." in *BMVC*, vol. 266, 2018.

[47] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8232–8241.

[48] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 5, pp. 1461–1474, 2019.

[49] H. Chung, W.-J. Nam, and S.-W. Lee, "Rotation invariant aerial image retrieval with group convolutional metric learning," *arXiv preprint arXiv:2010.09202*, 2020.

[50] S. Wang, Y. Ren, G. Parr, Y. Guan, and L. Shao, "Invariant deep compressible covariance pooling for aerial scene categorization," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[51] J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," *Advances in neural information processing systems*, vol. 17, pp. 513–520, 2004.

[52] 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.

[53] 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.

[54] 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.

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

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

[57] 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," *arXiv preprint arXiv:1912.01703*, 2019.

[58] J. Kang, R. Fernandez-Beltran, P. Duan, X. Kang, and A. J. Plaza, "Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[59] 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.

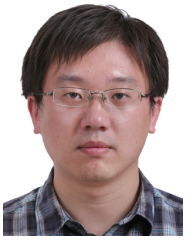
[60] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8905–8918, 2020.

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.

Zhirui Wang received the B.Sc. degree from the Harbin Institute of Technology, Harbin, China, in 2013, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2018.

He is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include SAR terrain classification, SAR target detection and recognition.



Xian Sun (Senior Member, IEEE) received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Jingen Ni (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, Jiangsu, 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, 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. His research interests include adaptive signal processing, distributed optimization and learning, remote sensing image processing, and artificial neural networks.

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