Towards Tightness of Scalable Neighborhood Component Analysis for Remote Sensing Image Characterization

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

***Abstract*—Deep metric learning methods have recently drawn significant attention in the field of remote sensing (RS), owing to their prominent capabilities for modeling relations among RS images based on their semantic contents. In the context of scene classification and large-scale image retrieval, one of the most prominent deep metric learning methods is the scalable neighborhood component analysis (SNCA), which has demonstrated excellent performance on the locality neighborhood structure in the metric space. However, the standard SNCA has important constraints on separating the hard positive and other negative images in the metric space, and this may become a major limitation when dealing with the large-scale variance problem inherent to RS data. To address this issue, we propose a novel deep metric learning formulation that introduces a new margin parameter to enforce the compactness of the within-class feature embeddings. Based on this innovative scheme, we propose two novel loss functions: 1) T-SNCA-c, where the parameter is based on the cosine similarity, and 2) T-SNCA-a, where the parameter is based on the angular distance. Besides, we exploit memory bank optimization to further enhance the semantic diversity during training. Our experimental results, conducted using three downstream applications (K-NN classification, clustering, and image retrieval) and two large-scale RS benchmark datasets, demonstrate that the proposed approach is able to achieve superior performance when compared to current state-of-the-art deep metric learning methods. The codes of this work will be made available online[1].**

***Index Terms*—Deep metric learning, deep learning, remote sensing scene classification, neighborhood component analysis, contrastive learning.**

I. Introduction

REMOTE sensing (RS) data have recently attracted significant attention due to their multiple applications in different human activities, such as urban planning [1], target identification and monitoring [2], [3], emergency response [4] as well as natural resource preservation and analysis [5]–[7]. In these, the task of characterizing RS images plays a fundamental role for the accurate semantic understanding and visual recognition of aerial scenes [8], [9]. The RS image characterization process consists of extracting discriminating features in order to generate precise representations useful for classifying, clustering, or even retrieving RS scenes based on their visual content. Different from standard imagery, airborne and spaceborne optical data comprise a wide variety of spatial structures and relationships that lead to high intra-class and low inter-class variations which make the RS scene characterization problem particularly challenging.

Over the past decade, different deep learning-based methods have been specifically developed to characterize RS imagery. Among all the conducted research, deep metric learning has recently shown to be one of the most prominent paradigms [10]–[13]. In particular, deep metric learning is concerned with projecting semantically similar images to nearby locations in the corresponding characterization space, which becomes highly convenient for RS scenes since the similarity measurements between images can preserve their complex semantic relationships and hence facilitate the corresponding downstream applications. Multiple works exemplify this growing trend. For example, Cheng *et al.* present in [11] the discriminative convolutional neural network (D-CNN) method for effectively characterizing aerial scenes. Their deep metric learning approach aims at uncovering highly discriminative CNN-based features by imposing a metric learning regularization term on the corresponding CNN model, according to the contrastive loss formulation [14]. Following this idea, Yan *et al.* propose in [15] a cross-domain extension to further reduce the feature distribution bias and spectral shift inherent to RS data. Despite the success of these techniques, other authors suggest the use of alternative formulations. For instance, Cao *et al.* define in [16] a content-based image retrieval system that employs the triplet loss formulation [17], which takes advantage of positive and negative samples to build the corresponding deep metric embedding. Nonetheless, sampling informative triplets (or even pairs) from scalable RS data becomes a challenging task, since the probability of finding representative samples within a randomly sampled and limited batch logically decreases with

This work was in part supported by the National Key R&D Program of China under Grant 2018YFB0505000, in part by the Natural Science Foundation of China under Grants 62101371 and 42071324, in part by the Shanghai Rising-Star Program (21QA1409100), in part by Jiangsu Province Science Foundation for Youths under Grant BK20210707, in part by the Ministry of Science, Innovation and Universities of Spain (PID2019-110315RB-I00, APRISA), in part by FEDER-Junta de Extremadura (Ref. GR18060) and in part by the European Union under the H2020 EOXPOSURE project (No. 734541). *(Corresponding author: Sicong Liu)*

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

R. Fernandez-Beltran is with the Department of Computer Science and Systems, University of Murcia, 30100 Murcia, Spain. (e-mail: rufernan@um.es).

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

[1] https://github.com/jiankang1991/GRSL_TSNCA

the data volume and complexity [18].

In this paper, we tackle the RS image characterization task from a deep metric learning perspective by taking the large-scale variance problem into account. Specifically, the proposed approach pursues to learn, via CNNs, a low-dimensional metric space that is able to model more complex semantic relationships among RS scenes than other existing techniques. To reach this goal, we initially consider scalable neighborhood component analysis (SNCA) [19] (one of the most successful characterization methods) and investigate its main limitations when dealing with the large-scale variance problems inherent to RS data. Then, we formulate a novel deep metric learning scheme based on the locality neighborhood structure of SNCA by introducing a new margin parameter in the tightness term, with the goal of improving the compactness of the feature embeddings sharing the same semantic label. In this way, rather different RS scenes that belong to the same class are enforced to be consistently closer in the embedding space (and also distant from other negative images with a security margin). This guarantees a better separability for high intra-class and low inter-class variations. In order to test this innovative scheme under different types of angular margins, we propose two novel loss functions: 1) T-SNCA-c, where the margin parameter is based on the cosine similarity, and 2) T-SNCA-a, where the parameter is based on the angular distance. We also exploit memory bank optimization to generate a richer semantic variability during training. To validate the effectiveness of our contribution, we conduct a comprehensive experimental comparison. The fundamental contributions of this paper can be summarized as follows: 1) We propose a new deep metric learning scheme for RS image characterization by introducing a new margin parameter in the tightness term of SNCA in order to enhance the intra-class compactness and inter-class separability of the feature embeddings; and 2) Under this novel scheme, we develop two new loss functions: a) T-SNCA-c, and b) T-SNCA-a, where the margin penalty is based on the cosine similarities and angular distances among the images sharing the same class.

II. Towards Tightness of Scalable Neighborhood Component Analysis

Let $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_M\}$ denote an RS image dataset composed of M images with class annotations, and $\mathcal{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_M\}$ be the associated set of label vectors, where each vector $\mathbf{y}_i$ is described by the one-hot vector, i.e., $\mathbf{y}_i \in \{0,1\}^C$, and C denotes the total number of classes. If the image $\mathbf{x}_i$ is annotated by class c, 1 is assigned to the c-th element of $\mathbf{y}_i$ and the other elements are 0. $\mathbf{v}_i \in \mathbb{R}^D$ denotes the feature extracted by the CNN model from the image $\mathbf{x}_i$, and D is the dimension. Based on the L_2 normalization, we define $\mathbf{f}_i$ as the normalized feature on a unit hypersphere (i.e., $\mathbf{f}_i = \mathbf{v}_i / \|\mathbf{v}_i\|_2$). $\mathcal{T}$ refers to the training set of the CNN model.

The SNCA [19] aims at embedding high-dimensional features into a low-dimensional space through a CNN by the stochastic maximization of the leave-one-out K-NN score. To achieve this, it minimizes the following loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{SNCA}} &= -\frac{1}{|\mathcal{T}|}\sum_{i}\log\big(\sum_{j\in\Omega_i}\frac{\exp(s_{ij}/\sigma)}{\sum_{k\neq i}\exp(s_{ik}/\sigma)}\big)\\ &= -\frac{1}{|\mathcal{T}|}\sum_{i}\Big(\log\sum_{j\in\Omega_i}\exp(s_{ij}/\sigma)\\ &\quad -\log\big(\sum_{j\in\Omega_i}\exp(s_{ij}/\sigma)+\sum_{k\notin\Omega_i}\exp(s_{ik}/\sigma)\big)\Big).\end{aligned} \tag{1}$$

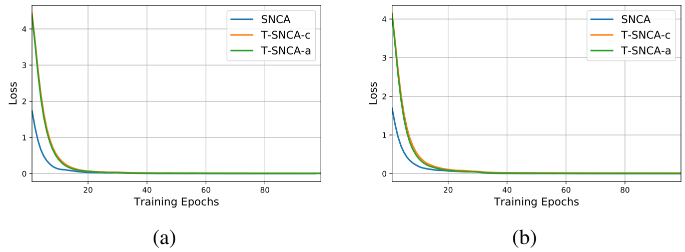
where s_{ij} is the cosine similarity described by the inner product of their normalized feature embeddings, i.e., $s_{ij} = \mathbf{f}_i^T \mathbf{f}_j$, and σ denotes a temperature parameter controlling the concentration level of the sample distribution. We focus on developing a novel loss formulation by enforcing the tightness of the feature embeddings of each class produced by SNCA. It can be seen that $\mathcal{L}_{\mathrm{SNCA}}$ is composed of two terms: $\sum_{j\in\Omega_i}\exp(s_{ij}/\sigma)$ and $\sum_{k\notin\Omega_i}\exp(s_{ik}/\sigma)$. Then, minimizing $\mathcal{L}_{\mathrm{SNCA}}$ will lead to the maximization of $\sum_{j\in\Omega_i}\exp(s_{ij}/\sigma)$. Consequently, an optimal solution will be reached when the following condition is satisfied:

$$s_{ij} = 1, \quad \forall j \in \Omega_i. \tag{2}$$

That is to say, all the feature embeddings of the images belonging to the same class should be aligned perfectly among each other. Therefore, $\sum_{j\in\Omega_i}\exp(s_{ij}/\sigma)$ controls the tightness of the feature embeddings for each class [20]. As a result, we introduce a penalty parameter m_c to represent the margin between the similarity of a positive pair and a negative pair. To be specific, let us assume that there are just three images $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ in the dataset, where $(\mathbf{x}_i, \mathbf{x}_j)$ represents the positive pair and $(\mathbf{x}_i, \mathbf{x}_k)$ represents the negative pair, with the cosine similarities s_{ij} and s_{ik}, respectively. Under the learning scheme of SNCA, $\mathbf{x}_i$ can be correctly classified if $s_{ij} > s_{ik}$. We can see that there is no margin incorporated in the metrics for the class decision. By introducing m_c, $\mathbf{x}_i$ should be correctly classified if $s_{ij} - m_c > s_{ik}$. Thus, the tightness of the feature embeddings from each class can be enforced when a margin is manually incorporated into the metrics for the class decision. To this end, we propose a novel loss by introducing a margin into the cosine similarity of the tightness term in SNCA, which is named as *T-SNCA-c*:

$$\begin{aligned}\mathcal{L}_{\mathrm{T-SNCA-c}} = -\frac{1}{|\mathcal{T}|}\sum_{i}\Big(&\log\sum_{j\in\Omega_i}\exp((s_{ij}-m_c)/\sigma)\\ &-\log\big(\sum_{j\in\Omega_i}\exp((s_{ij}-m_c)/\sigma)\\ &+\sum_{k\notin\Omega_i}\exp(s_{ik}/\sigma)\big)\Big).\end{aligned} \tag{3}$$

By comparing $\mathcal{L}_{\mathrm{T-SNCA-c}}$ and $\mathcal{L}_{\mathrm{SNCA}}$, we can observe that $\mathcal{L}_{\mathrm{T-SNCA-c}} > \mathcal{L}_{\mathrm{SNCA}}$. Precisely, this leads to a loss perspective for analyzing the effectiveness of the proposed $\mathcal{L}_{\mathrm{T-SNCA-c}}$ on the tightness enhancement. That is, $\mathcal{L}_{\mathrm{T-SNCA-c}}$ can be considered as an upper bound of $\mathcal{L}_{\mathrm{SNCA}}$. When both $\mathcal{L}_{\mathrm{T-SNCA-c}}$ and $\mathcal{L}_{\mathrm{SNCA}}$ are optimized to the same level, the within-class similarity values s_{ij} in $\mathcal{L}_{\mathrm{T-SNCA-c}}$ should be larger than the values in $\mathcal{L}_{\mathrm{SNCA}}$ (at least with a



(a) (b)

Fig. 1. The values of $\mathcal{L}_{\mathrm{SNCA}}$, $\mathcal{L}_{\mathrm{T-SNCA-c}}$ and $\mathcal{L}_{\mathrm{T-SNCA-a}}$ with respect to the training epochs on two benchmark datasets: (a) AID; (b) NWPU-RESISC45.

TABLE I

K-NN CLASSIFICATION ACCURACIES (%) OBTAINED BY ALL THE CONSIDERED METHODS WHEN $K = 1, 5, 10$.

	AID			**NWPU-RESISC45**		
	1	5	10	1	5	10
D-CNN	93.10	93.70	93.75	91.21	91.62	91.48
Triplet	92.85	93.10	93.25	90.83	91.46	91.43
NSL	93.25	93.30	93.55	91.75	91.78	91.84
SNCA	94.55	94.50	94.60	92.13	92.21	92.14
ArcFace	94.70	94.65	94.75	92.90	92.97	92.92
T-SNCA-c	**95.25**	**95.35**	**95.40**	**93.37**	93.35	**93.38**
T-SNCA-a	95.15	95.10	95.00	93.33	**93.40**	93.37

level of m_c). That leads to the generation of intra-class feature embeddings with more compactness, so that semantically-similar images can be grouped closer and inter-class feature embeddings can be pushed away with larger margins.

Since the cosine similarity tends to make nearly parallel vectors very similar, we also investigate another penalization scheme based on the angular distance [21] in order to better distinguish between even small differences depending on the considered downstream application and data. Hence, we also define the following loss, termed as *T-SNCA-a*:

$$\mathcal{L}_{\mathrm{T-SNCA-a}} = -\frac{1}{|\mathcal{T}|}\sum_{i}\Big(\log\sum_{j\in\Omega_i}\exp(\cos(\theta_{ij}+m_a)/\sigma) - \log\big(\sum_{j\in\Omega_i}\exp(\cos(\theta_{ij}+m_a)/\sigma) + \sum_{k\notin\Omega_i}\exp(s_{ik}/\sigma)\big)\Big), \tag{4}$$

where $\theta_{ij} = \arccos(s_{ij})$ denotes the angular distance between $\mathbf{f}_i$ and $\mathbf{f}_j$, and m_a represents the introduced angular margin. Likewise, given three images $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ in the dataset, where $(\mathbf{x}_i, \mathbf{x}_j)$ represents the positive pair and $(\mathbf{x}_i, \mathbf{x}_k)$ represents the negative pair with the cosine similarities $\cos(\theta_{ij})$ and $\cos(\theta_{ik})$, respectively. By adding the angular penalty, $\mathbf{x}_i$ should be correctly classified if $\cos(\theta_{ij}+m_a) > \cos(\theta_{ij})$, rather than $\cos(\theta_{ij}) > \cos(\theta_{ij})$ in the SNCA case. In other words, although m_a is added to the angle between the feature embeddings of the positive image pair, $\mathbf{x}_i$ can be also correctly distinguished. Thus, the tightness of the within-class feature embeddings is also enforced with T-SNCA-a. For $\mathcal{L}_{T-SNCA-c}$, the margin is directly enforced on the cosine similarities between two embeddings, which is *cosine margin penalty*. As a difference, the angular distance is investigated in $\mathcal{L}_{T-SNCA-a}$, whose margin directly penalizes the angle between the features in the embedding space. Compared with SNCA, the proposed T-SNCA-c and T-SNCA-a can both emphasize the tightness of the generated within-class feature embeddings during the learning progress of the CNN model. To optimize the proposed losses, we follow the same mechanism proposed in [19].

III. Experiments

A. Experimental Setup

In this work, we conduct extensive experiments based on the following RS benchmark datasets: **Aerial Image Dataset (AID)** [8] and **NWPU-RESISC45** [9]. These datasets have been randomly split into training, validation and test sets with the ratios of 70%, 10% and 20%, respectively. In order to sufficiently evaluate the effectiveness of the proposed approach over the considered datasets, following our previous work [12], we perform an experimental assessment based on the three different downstream applications including : K-NN classification, clustering, and image retrieval.

Regarding the general implementation details, the proposed method is implemented in PyTorch. Additionally, we select the ResNet18 [22] model as the backbone CNN architecture for the proposed approach as well as all the other considered methods. The data augmentation strategy is consistent with [12]. The parameters D, σ, λ, m_c, and m_a are set as 128, 0.1, 0.5, 0.1 and 0.2, respectively. To evaluate the effectiveness of the proposed losses for metric learning, we compare them with several state-of-the-art deep metric learning methods, including: 1) **D-CNN** [11]; 2) deep metric learning based on triplet loss [16], [17], simply termed as **Triplet** hereinafter; 3) **SNCA** [19]; 4) Normalized Softmax Loss (**NSL**) [23]; and 5) **ArcFace** [21].

B. Experimental Results

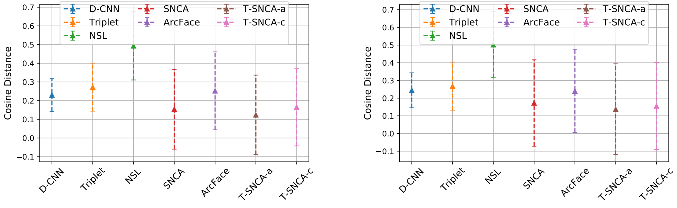
1) KNN Classification: Figure 1 presents the values of $\mathcal{L}_{\mathrm{SNCA}}$, $\mathcal{L}_{\mathrm{T-SNCA-c}}$ and $\mathcal{L}_{\mathrm{T-SNCA-a}}$ with respect to the training epochs on the two benchmark datasets. During the first several epochs, the values of $\mathcal{L}_{\mathrm{T-SNCA-c}}$ and $\mathcal{L}_{\mathrm{T-SNCA-a}}$ are much higher than $\mathcal{L}_{\mathrm{SNCA}}$, due to the incorporated margin parameter. As the training progresses, all the losses can converge to the same level. Thus, the within-class similarities produced by T-SNCA-a and T-SNCA-c are higher than those obtained by SNCA. Based on the K-NN classifier ($K = 1, 5, 10$), we can calculate the overall accuracies of all the considered methods on the test sets in Table I. It can be seen that the proposed losses can achieve the best performance on the two datasets compared with the other losses. For example, around 0.7% accuracy improvement can be achieved by T-SNCA-a in comparison with ArcFace. In ArcFace, all the within-class feature embeddings are optimized to be aligned with respect to

TABLE II
ACC SCORES (%) OF THE FEATURE EMBEDDINGS OF THE TEST SETS PRODUCED BY DIFFERENT LEARNING METHODS.

	AID	NWPU-RESISC45
D-CNN	84.50	87.44
Triplet	92.50	88.33
SNCA	94.65	92.13
NSL	88.25	88.03
ArcFace	94.85	92.90
T-SNCA-c	**95.45**	91.86
T-SNCA-a	95.20	**93.32**

TABLE III
MAP RESULTS (%) ON THE CONSIDERED BENCHMARK DATASETS WITH RESPECT TO 20, 50 AND 100 RETRIEVED IMAGES.

	AID			NWPU-RESISC45		
	20	50	100	20	50	100
D-CNN	93.72	91.47	89.39	92.09	89.64	87.83
Triplet	94.02	92.36	91.25	92.72	90.78	89.49
SNCA	99.16	98.76	98.49	98.94	98.53	98.28
NSL	94.17	91.74	90.02	93.47	91.48	89.98
ArcFace	98.84	98.15	97.79	98.88	98.44	98.04
T-SNCA-a	**99.51**	**99.35**	**99.09**	**99.50**	**99.22**	**99.11**
T-SNCA-c	99.33	99.08	98.77	99.38	99.20	99.00



(a) (b)

Fig. 2. Mean and standard deviation values of the cosine distances among the within-class feature embeddings on the considered benchmark datasets: (a) AID and (b) NWPU-RESISC45.

TABLE IV
SENSITIVITY ANALYSIS (%) OF PARAMETERS τ AND m ON THE NWPU-RESISC45 DATASET (T-SNCA-A).

parameters	$m = 0.1$	$m = 0.2$	$m = 0.3$	$m = 0.4$	$m = 0.5$
$\tau = 0.05$	93.67	93.73	92.79	92.29	92.11
$\tau = 0.1$	92.83	93.37	93.54	92.7	92.54
$\tau = 0.15$	92.83	92.75	92.6	91.84	91.84
$\tau = 0.2$	92.03	91.75	91.84	90.33	89.79
$\tau = 0.25$	91.76	90.7	91.22	89.59	86.13

the associated learned class prototypes. Such learning scheme may not be able to discover the locality structure among the images in the metric space, especially when the semantic variations exist within each class. Since the proposed losses align the within-class feature embeddings in a contrastive manner, it facilitates the CNN model to capture the semantic variations within each class. Thus, both T-SNCA-a and T-SNCA-c can outperform ArcFace in K-NN classification.

2) Clustering: Table II display the clustering accuracy (ACC) scores obtained on the feature embeddings of the test sets after applying K-means clustering. It can be observed that the proposed losses can achieve the best matching between the ground-truth labels and the pseudo-labels assigned based on the K-means clustering. For example, more than 10% performance gain can be achieved by the proposed methods in comparison with D-CNN. By enforcing the tightness of the within-class feature embeddings for each class, we can see T-SNCA-a can improve the score by a value around 0.5% to 1%. In order to illustrate the within-class similarities of the feature embeddings, we plot the associated histograms with respect to different similarity bins on the two benchmark datasets. Moreover, we calculate the mean and standard deviation scores of the cosine distance, i.e., $1 - s_{ij}$, for the within-class feature embeddings obtained by all the considered methods. As shown in Figure 2, the mean values of the cosine distances of the within-class feature embeddings for SNCA,T-SNCA-c and T-SNCA-a are lower than those obtained by the other methods. By introducing the margin parameter, the proposed T-SNCA-a and T-SNCA-c can decrease the distances among the images sharing the same semantic class in the metric space.

3) Image Retrieval: Table III presents the MAP values with respect to 20, 50, and 100 retrieved images. With an increasing number of retrieved images, the MAP values of the proposed methods do not degrade much, while the other methods (such as NSL) cannot exhibit comparable performances. Among all the considered losses, both the proposed T-SNCA-a and T-SNCA-c losses achieve the highest image retrieval accuracies. Figure 3 illustrates some retrieval examples based on SNCA, ArcFace and T-SNCA-a. Given two images from the AID and NWPU-RESISC45 test sets, we show the top-5 nearest neighbor images retrieved from the training sets. For example, Church and School images are confused in the retrieval result based on SNCA. For both SNCA and ArcFace, the semantic contents of Forest and Wetland cannot be well distinguished.

4) Parameter Sensitivity Analysis: For the proposed losses, two main parameters: τ and m may influence the effectiveness of the metric learning. Table IV and Table V present the K-NN classification accuracies with respect to different values of τ and m obtained by T-SNCA-a and T-SNCA-c on the NWPU-RESISC45 dataset, when $K = 10$. Darcker colors refer to better results and vice versa. For τ, optimal classification performance can be achieved in the range from 0.05 to 0.15. When the values of m are around 0.1, both T-SNCA-a and T-SNCA-c can obtain the best performance. It can be observed that the two proposed losses favor small values of the margin m, e.g., 0.1. Over-tightness, i.e., large values of m, may lead to decreased intra-class feature variations, which can degrade the generalization capability of the CNN models on test data. Moreover, some samples may easily satisfy the over-tightness condition, while other (hard) samples are difficult to be optimized under such condition. Thus, a sub-optimal solution may be obtained in this case.

IV. CONCLUSION

In this paper, we present a new deep metric learning model (called T-SNCA) able to accurately characterize RS scenes by

TABLE V
SENSITIVITY ANALYSIS (%) OF PARAMETERS τ AND m ON THE NWPU-RESISC45 DATASET (T-SNCA-C).

parameters	$m = 0.1$	$m = 0.2$	$m = 0.3$	$m = 0.4$	$m = 0.5$
$\tau = 0.05$	93.87	93	92.24	92.05	92.24
$\tau = 0.1$	93.38	93.24	93.27	92.94	92.84
$\tau = 0.15$	93.16	92.94	92.78	93.24	92.83
$\tau = 0.2$	92.81	92.78	93.08	92.6	92.52
$\tau = 0.25$	92.46	91.95	92.33	92.43	92.14

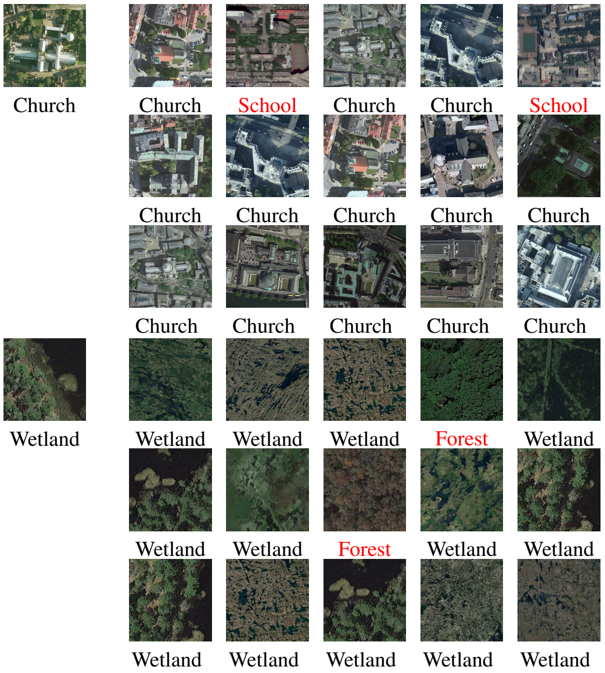


Fig. 3. Top-5 nearest neighbors retrieved with respect to the query images based on SNCA (first row), ArcFace (second row) and T-SNCA-a (third row), where the left-most query images are from the AID and NWPU-RESISC45 datasets, respectively.

relieving the large-scale variance problem. Specifically, a novel formulation that introduces a new margin parameter to enforce the compactness of the within-class feature embeddings and the inter-class separation is proposed, which leads to two loss functions: T-SNCA-c and T-SNCA-a. Extensive experiments validate the effectiveness of the proposed method. It can be observed that the proposed method outperforms other state-of-the-art metric learning methods on three different tasks. In addition, by enforcing the tightness, the proposed method also achieves superior performances than SNCA. By analyzing the performances between the two different losses, the angular version achieves slightly higher accuracy than the other one, which is suggested to be exploited in real cases. Moreover, we conclude that it is highly suitable for characterizing large-scale RS image archives. As future work, we plan to adapt our approach to other learning paradigms, including multiple labels, semi-supervised and unsupervised frameworks.

REFERENCES

[1] Q. Weng, D. Quattrochi, and P. E. Gamba, *Urban remote sensing*. CRC press, 2018.

[2] 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.

[3] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, “Picoco: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 10 548–10 559, 2021.

[4] 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,” *Proc. IEEE*, vol. 105, no. 10, pp. 1900–1912, 2017.

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

[6] R. Fernandez-Beltran, F. Pla, and A. Plaza, “Endmember extraction from hyperspectral imagery based on probabilistic tensor moments,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 12, pp. 2120–2124, 2020.

[7] F. Luo, Z. Zou, J. Liu, and Z. Lin, “Dimensionality reduction and classification of hyperspectral image via multi-structure unified discriminative embedding,” *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[8] 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 Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.

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

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

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

[12] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, “Deep metric learning based on scalable neighborhood components for remote sensing scene characterization,” *IEEE Trans. Geosci. Remote Sens.*, pp. 1–14, 2020.

[13] Y. Duan, H. Huang, and T. Wang, “Semisupervised feature extraction of hyperspectral image using nonlinear geodesic sparse hypergraphs,” *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[14] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. IEEE, 2006, pp. 1735–1742.

[15] L. Yan, R. Zhu, N. Mo, and Y. Liu, “Cross-domain distance metric learning framework with limited target samples for scene classification of aerial images,” *IEEE Trans. Geosci. Remote Sens.*, 2019.

[16] 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.

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

[18] 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],” *Proc. IEEE*, vol. 107, no. 12, pp. 2294–2301, 2019.

[19] 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.

[20] M. Boudiaf, J. Rony, I. M. Ziko, E. Granger, M. Pedersoli, P. Piantanida, and I. B. Ayed, “Metric learning: cross-entropy vs. pairwise losses,” *arXiv preprint arXiv:2003.08983*, 2020.

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

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

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