Graph Relation Network: Modeling Relations between Scenes for Multi-Label Remote Sensing Image Classification and Retrieval

Jian Kang, *Member, IEEE,* Ruben Fernandez-Beltran, *Senior Member, IEEE,* Danfeng Hong, *Member, IEEE,* Jocelyn Chanussot, *Fellow, IEEE* and Antonio Plaza, *Fellow, IEEE*

Abstract

Owing to the proliferation of large-scale remote sensing (RS) archives with multiple annotations, multi-label RS scene classification and retrieval are becoming increasingly popular. Although some recent deep learning-based methods are able to achieve promising results in this context, the lack of research on how to learn embedding spaces under the multi-label assumption often makes these models unable to preserve complex semantic relations pervading aerial scenes, which is an important limitation in RS applications. To fill this gap, we propose a new graph relation network (GRN) for multi-label RS scene categorization. Our GRN is able to model the relations between samples (or scenes) by making use of a graph structure which is fed into network learning. For this purpose, we define a new loss function called scalable neighbor discriminative loss with binary cross entropy (SNDL-BCE), that is able

This work has been supported by the Spanish Ministry of Economy (RTI2018-098651-B-C54), FEDER-Junta de Extremadura under Grant GR18060 and the European Union under the H2020 EOXPOSURE project Grant 734541 as well as with the support of the AXA Research Fund. *(Corresponding author: Danfeng Hong)*

J. Kang is with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany (e-mail: jian.kang@tu-berlin.de).

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).

D. Hong is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: danfeng.hong@dlr.de).

J. Chanussot is with the Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France, and also with the Faculty of Electrical and Computer Engineering, University of Iceland, 101 Reykjavik, Iceland (e-mail: jocelyn@hi.is).

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).

to embed the graph structures through the networks more effectively. The proposed approach can guide deep learning techniques (such as convolutional neural networks) to a more discriminative metric space, where semantically similar RS scenes are closely embedded and dissimilar images are separated from a novel multi-label viewpoint. To achieve this goal, our GRN jointly maximizes a weighted leave-one-out K-nearest neighbors (KNN) score in the training set, where the weight matrix describes the contributions of the nearest neighbors associated with each RS image on its class decision, and the likelihood of the class discrimination in the multi-label scenario. An extensive experimental comparison, conducted on three multi-label RS scene data archives, validates the effectiveness of the proposed GRN in terms of KNN classification and image retrieval. The codes of this paper will be made publicly available for reproducible research in the community[1].

Index Terms

Remote sensing, deep learning, metric learning, loss function, neighbor embedding, multi-label scene categorization.

I. Introduction

With the constant development of satellite sensor technology, remote sensing (RS) images are widely employed in numerous applications, such as urban mapping [1]–[5], object detection and recognition [6]–[10], image processing and analysis [11]–[14], and spectral unmixing [15]–[17]. RS scene classification and retrieval [18], [19] play a crucial role in the aforementioned tasks, because they focus on predicting the semantic content and visual understanding associated to a given aerial scene [20].

During the last decades, extensive research has been conducted on the development of RS scene categorization models [18], [21]–[29]. For example, in [30], the proposed method can well integrate spatial information and efficiently extract nonlinear features, and shows state-of-the-art classification performance when there are limited training samples. The majority of the presented methods aim at providing a single interpretation of RS scenes, which are assumed to contain only one land-use or land-cover semantic class [31]. However, such hypothesis may not hold in RS problems, since it may not be sufficient to characterize the high semantic complexity of the RS image domain, especially when considering high-resolution remotely sensed images [32]. To better describe the objects within an aerial scene, multiple labels may be required to represent

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

the visual semantics of RS images. In general, the multi-label image classification and retrieval problem consists of predicting (or searching) semantically related visual contents that contain multiple annotations, providing a substantially richer semantic description of the corresponding scenes. As a result, extensive efforts have been recently directed towards investigating the multi-label scheme [33]–[38]. For example, one of the primal multi-label methods proposed within the RS field was presented in [39] where the authors define a multi-label support vector machine (SVM) for multi-label active learning. To simultaneously exploit the spatial-contextual information and the correlation among the labels, Zeggada *et al.* presented in [40] a conditional random field (CRF) framework for multi-label classification of images collected by unmanned aerial vehicles (UAVs).

Fostered by the fast proliferation of large-scale RS archives [41]–[44], deep learning has also been applied to multi-label RS scene categorization owing to its excellent feature extraction capabilities. Different works in the RS literature exemplify this fact. For instance, Karalas *et al.* developed in [45] a sparse auto-encoder framework to extract the underlying semantic features from satellite images, in order to effectively retrieve multi-label land-cover categories. The authors of [46] proposed a deep learning model for predicting multi-labels in UAV images via a radial basis function (RBF) network applied on the local image descriptors, which are then extracted using a convolutional neural network (CNN). Despite the effectiveness achieved by these and other relevant methods in the literature, the standard CNN architecture is generally unable to exhibit a salient performance in RS, due to the so-called Hughes phenomenon that arises when considering limited amounts of labeled images [47]. Note that the availability of sufficient multi-labeled images is a major problem in RS, because obtaining (fine-grained) ground-truth annotations is very expensive (as well as time-consuming). To overcome this important constraint, a data augmentation technique was recently proposed in [48] to enlarge available multi-label RS training sets. Nonetheless, other authors opt for different alternatives instead. It is the case of Hua *et al.*, who proposed in [49] an end-to-end network for multi-label aerial image classification which is based on three components: a CNN-based feature extraction module, a class-wise attention mechanism, and a bidirectional long short-term memory (LSTM) sub-network. Driven by multi-attention techniques, the authors of [50] also designed a CNN-based deep learning system for RS images with multiple annotations. Alshehri *et al.* presented in [51] a multi-label categorization approach based on an encoder-decoder neural network with pre-trained CNN features and channel-spatial attention. Additionally, Shao *et al.* proposed in [52] a multi-label

RS image retrieval system that employs a fully convolutional network which is first trained to predict the corresponding segmentation maps and then used to characterize each individual region with multi-scale features.

Most of the existing deep learning methods in multi-label RS scene classification and retrieval domains focus on designing suitable CNN architectures to improve the label assignment performance, given the high semantic complexity of the RS image domain. However, the learned feature embeddings for aerial images have not been fully investigated yet. Precisely, this is the gap that motivates this research work. In other words, despite the fact that some of the above-mentioned approaches already exhibit remarkable performances on multi-label categorization problems, their corresponding low-dimensional feature embeddings may not fully preserve the semantic relations pervading the objects in RS scenes, where semantically similar images are logically expected to be close in the uncovered feature space. Although one may think that such metric space could be produced by applying the standard contrastive loss or triplet loss [53], these techniques were initially designed for a single-label scene classification scheme, which may eventually constrain their performance from a multi-label RS image analysis perspective.

In this paper, we deal with the multi-label RS scene classification and retrieval problem by taking the characteristics of the learned CNN-based feature embeddings into account. Specifically, we propose a new graph relation network (GRN) for effectively classifying and retrieving RS scenes with multiple annotations by using a new loss function, called scalable neighbor discriminative loss (SNDL). Inspired by the scalable neighborhood component analysis (SNCA) [54], the proposed SNDL provides a novel perspective on the multi-label RS scene case through the ability to learn a metric space where semantically similar RS images are pulled closer (and dissimilar images are pushed away) based on their multi-label semantic contents. Specifically, we model the semantic proximity of the learned CNN-based feature embeddings using a stochastic process that maximizes a weighted leave-one-out k-nearest neighbors (KNN) [55] score in the training set, where the weight matrix obtained by the multi-label information characterizes the contributions of the nearest neighbors associated with each image on its semantic class decision. In order to further improve the multi-label discrimination capability over RS scenes, we also design a joint loss function, termed as SNDL-BCE, by combining SNDL with binary cross entropy (BCE). The experimental part of the work validates the performance of the proposed scheme by conducting a comprehensive experimental comparison, using three benchmark data archives and different state-of-the-art models in multi-label RS scene classification and retrieval.

In summary, the main contributions of our paper can be highlighted as follows:

1) We develop a new GRN for multi-label RS scene classification and retrieval by introducing an advanced scheme based on a new loss function (SNDL) and its corresponding joint version (SNDL-BCE). The new loss functions have been proven to be effective in guiding CNN models to produce a more discriminative metric space, both instantly and class-wisely.
2) To the best of our knowledge, this is the first work in the literature that considers graph-based neighborhood semantic relationships between multi-label RS scene images in an end-to-end deep neural network and adapts the SNCA to the multi-label scheme.
3) The proposed GRN demonstrates its superiority with respect to state-of-the-art loss functions, such as BCE and log-sum-exp pairwise (LSEP) [56], that have been widely used in multi-label RS scene classification and retrieval tasks.
4) The proposed GRN also shows a higher effectiveness and robustness when considering different benchmark RS datasets and backbone CNN architectures. The related codes of this paper will be made publicly available for reproducible research inside the community.

The rest of this paper is organized as follows. Section II provides the rationale and details of the proposed approach, and introduces our newly defined loss and optimization frameworks. Section III presents and discusses the quantitative and qualitative experimental results based on two different RS tasks: classification and image retrieval. Finally, Section IV concludes the paper with some remarks and hints at plausible future research lines.

II. Methodology

Let $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ be a set of N RS images and $\mathcal{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$ be the associated set of label vectors, where each label vector $\mathbf{y}_i$ is represented by a multi-class hot encoding vector, i.e., $\mathbf{y}_i \in \{-1, 1\}^C$. Let C be the total number of RS classes. If an image scene is assigned to the class c, the c-th element of $\mathbf{y}_i$ is 1, and -1 otherwise. $\mathcal{F}(\cdot;\theta)$ is the nonlinear mapping function represented by a backbone CNN model with a parameter set θ, which can map the original RS image $\mathbf{x}_i$ into a corresponding feature embedding $\mathbf{f}_i \in \mathbb{R}^D$ on the unit sphere, i.e., $||\mathbf{f}_i||_2 = 1$. A training set $\mathcal{T}$ (extracted from $\mathcal{X}$) is built in order to train the proposed deep metric learning system. Based on this notation, we first analyze the SNCA in Section II-A. Then, in Section II-B we provide the technical details of our approach, which is specially designed for multi-label RS scene image classification and retrieval.

A. Scalable neighborhood component analysis (SNCA)

As a scalable version of the standard neighborhood component analysis [57], the SNCA [54] was introduced to effectively learn a metric space based on CNN models, where the neighborhood structure of original images can be preserved. In other words, semantically similar images are projected to the learned metric space with smaller distances, and dissimilar images are separated [58]. The similarity s_{ij} between an image pair $(\mathbf{x}_i, \mathbf{x}_j)$ from a training set $\mathcal{T}$ can be measured by the cosine similarity, based on their feature embeddings in the metric space:

$$s_{ij} = \mathbf{f}_i^T \mathbf{f}_j, \tag{1}$$

where s_{ij} ranges from -1 to 1. A larger value of s_{ij} indicates a higher similarity of the two images. Given the image $\mathbf{x}_i$, the probability p_{ij} that the image $\mathbf{x}_j$ is located around its neighborhood in the metric space can be defined as:

$$p_{ij} = \frac{\exp(s_{ij}/\sigma)}{\sum_{k \neq i} \exp(s_{ik}/\sigma)}, \quad p_{ii} = 0, \tag{2}$$

where σ is a temperature parameter controlling the concentration level of the sample distribution [59], [60]. If s_{ij} is larger, $\mathbf{x}_j$ can be chosen as the neighbor of $\mathbf{x}_i$ in the metric space at a higher chance than another image $\mathbf{x}_k$. $p_{ii} = 0$ indicates that each image cannot select itself as its neighbor. It is also termed as *leave-one-out* distribution on $\mathcal{T}$. Based on this, the probability that $\mathbf{x}_i$ can be correctly classified is:

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

where $\Omega_i = \{j | \mathbf{y}_i = \mathbf{y}_j\}$ is the index set of training images sharing the same class with $\mathbf{x}_i$. Basically, the more images $\mathbf{x}_j$ (sharing the same class with $\mathbf{x}_i$) that are positioned as neighbors around $\mathbf{x}_i$ in the metric space, the higher the probability p_i that $\mathbf{x}_i$ is correctly classified. To this end, the objective of SNCA is to minimize the expected negative log-likelihood over $\mathcal{T}$, represented as:

$$L_{\text{SNCA}} = -\frac{1}{|\mathcal{T}|} \sum_i \log(p_i), \tag{4}$$

where $|\mathcal{T}|$ represents the number of training images.

Given $\mathbf{x}_i$, its similarities with respect to the other images in the dataset should be calculated for optimizing Equation (4). Therefore, in order to stochastically train a CNN model by L_{SNCA}, an off-line memory bank $\mathcal{B}$ is constructed for conducting the look-up during the training phase, which ultimately stores the normalized features of $\mathcal{T}$, i.e., $\mathcal{B} = \{\mathbf{f}_i, \cdots, \mathbf{f}_M\}$. $\mathcal{B}$ is updated in each iteration during the training phase.

The SNCA loss in Equation (4) can be viewed as a way to learn the nearest neighbors of each image in the metric space in supervised fashion. Within the learned metric space, the inherent structures among the images can be discovered, especially when there are relevant intra-class variations. This is a highly desired scenario when dealing with the particular semantic complexity of aerial scenes. However, Equation (4) is specially designed for learning the feature embeddings of images with single-labels, which eventually becomes a very important constraint in the RS field. Although convenient, the SNCA approach cannot be applied to classify and retrieve RS images with multiple semantic annotations. To solve this issue, we present a novel multi-label deep metric learning approach, based on a newly defined GRN-SNDL concept, to effectively learn a metric space for RS images with multi-label information.

B. Proposed multi-label deep metric learning framework for RS images

Our newly proposed end-to-end multi-label deep metric learning model for RS scene classification and retrieval can be condensed into three main components:

- A backbone CNN model (used to generate the corresponding feature embedding space of the input RS scene images). In this work, we adopt three state-of-the-art backbone architectures to derive and validate the proposed approach under different conditions, i.e. ResNet18 [61], ResNet50 [61] and WideResNet50 [62].
- A new loss function and its joint version, i.e. the GRN-SNDL and GRN-SNDL-BCE, which model the semantic proximity of the learned feature embeddings by maximizing a weighted leave-one-out KNN score and preserves the capability of class discrimination.
- The corresponding optimization algorithm, which learns the proposed model parameters using a stochastic process based on an off-line memory bank.

Figure 1 provides a graphical illustration of our multi-label deep metric learning framework. In the following sections, our newly defined loss function and the considered optimization algorithm are described in detail.

1) Loss function: scalable neighbor discriminative loss (SNDL): In order to design our GRN-SNDL under a multi-label assumption, we first rewrite the probability p_i that $\mathbf{x}_i$ can be correctly classified within the framework of SNCA (i.e., Equation (3)) as:

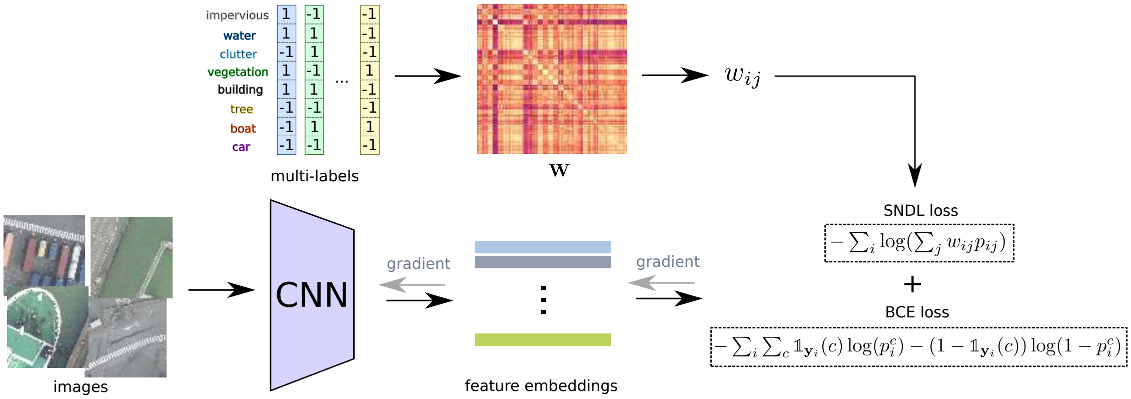


Fig. 1. The proposed framework for multi-label deep metric learning. The SNDL loss is targeted for pulling in the images that share more common labels and pushing away the images with less or no common labels. The BCE loss is integrated for further improving the class discrimination capability.

$$p_i = \sum_j \mathbb{1}_{\Omega_i}(j) p_{ij}, \tag{5}$$

where $\mathbb{1}_{\Omega_i}(j)$ is an indicator function given by:

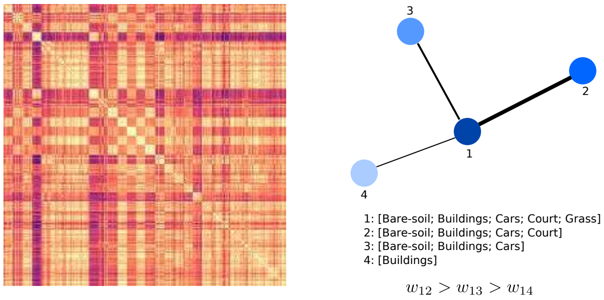
$$\mathbb{1}_{\Omega_i}(j) := \begin{cases} 1 & \text{if } j \in \Omega_i, \\ 0 & \text{if } j \notin \Omega_i. \end{cases} \tag{6}$$

Given the index set (Ω_i) of training images sharing the same class with respect to $\mathbf{x}_i$, the indicator function controls which images can be positioned as neighbors around $\mathbf{x}_i$ in the metric space. It can be observed that p_i is given by a weighted summation of p_{ij} over the whole dataset. If $\mathbf{x}_j$ shares the same class with $\mathbf{x}_i$, the associated weight is 1 (and 0 otherwise). In other words, all the contributions on the final class decision of $\mathbf{x}_i$ are dependent on the images that exhibit the same semantic annotation.

Inspired by this idea, for those images with multi-label annotations, the probability that $\mathbf{x}_i$ is correctly classified can be determined by:

$$p_i = \sum_j w_{ij} p_{ij}, \tag{7}$$

where w_{ij} denotes the contribution weight associated to p_{ij}. Given an image $\mathbf{x}_i$ and its multiple labels, we would like to pull in the images that share more common labels with regards to $\mathbf{x}_i$



(a) (b)

Fig. 2. (a) Weight matrix $\mathbf{W}$ of the aerial image dataset (AID). Darker points indicate smaller weights assigned to image pairs (and vice-versa). (b) A graph perspective view of the GRN-SNDL loss.

in the metric space, and push away the images with less or no common labels with regards to $\mathbf{x}_i$. To achieve this goal, a heavier weight w_{ij} should be allocated to an image pair (i, j) if the associated images have many labels in common, so that p_{ij} can contribute more to the multi-label decision for $\mathbf{x}_i$ through Equation (7). For that purpose, we propose to calculate w_{ij} based on the multi-label information in the corresponding images as follows:

$$w_{ij} = \frac{\langle \mathbf{y}_i, \mathbf{y}_j \rangle + C}{2C}, \quad w_{ij} \in [0, 1]. \tag{8}$$

Intuitively, w_{ij} depends on the inner product between $\mathbf{y}_i$ and $\mathbf{y}_j$, which is the *cosine* between $\mathbf{y}_i$ and $\mathbf{y}_j$. If $\mathbf{y}_i$ is more similar to $\mathbf{y}_j$, there will be a heavier weight assigned to the similarity term s_{ij} between $\mathbf{x}_i$ and $\mathbf{x}_j$. Since the original range of $\langle \mathbf{y}_i, \mathbf{y}_j \rangle$ is from $-C$ to C, we should normalize in the range from 0 to 1 via Equation (8). As an example, based on the multi-label annotations of the AID dataset [19], we utilize Equation (8) to calculate the weight matrix $\mathbf{W}$, and plot it in Figure 2(a), where the x and y axes represent the indexes of the images. The darker points indicate smaller weights assigned to image pairs (and vice-versa). To this end, the overall objective function is based on minimizing the expected negative log-likelihood through $\mathcal{T}$ with the following formulation, termed as *GRN-SNDL loss*:

$$L_{\text{SNDL}} = -\frac{1}{|\mathcal{T}|} \sum_i \log(p_i) = -\frac{1}{|\mathcal{T}|} \sum_i \log(\sum_j w_{ij} p_{ij}). \tag{9}$$

From a graph perspective, GRN-SNDL can be considered as a graph regularization, as it describes the relations between the scenes based on their semantic multi-labels. In the example shown in Figure 2(b), the connection between the node 1 and node 2 should be stronger than any

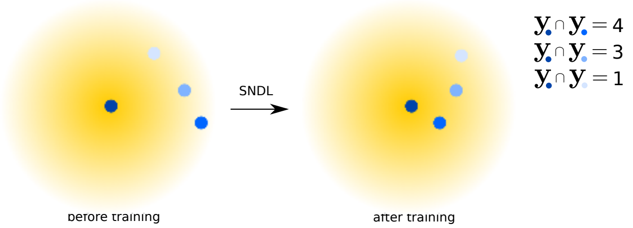

$\mathbf{y}_{\bullet} \cap \mathbf{y}_{\bullet} = 4$
$\mathbf{y}_{\bullet} \cap \mathbf{y}_{\bullet} = 3$
$\mathbf{y}_{\bullet} \cap \mathbf{y}_{\bullet} = 1$
SNDL
before training
after training

Fig. 3. An illustration of our learning scheme based on GRN-SNDL. Blue points represent features (associated to images) in the metric space. With respect to the center point, the other points have different numbers of identical class labels, and this determines their position in the metric space after training with GRN-SNDL. Specifically, the points associated to images with more labels in common have been dragged closer than the points associated to images with less common labels (with respect to the center point).

other node linked with node 1, since they share more common labels. By constructing such graph regularization based on their label information, the locality structure can be better discovered within the feature space.

An illustration of the learning scheme of the proposed GRN-SNDL is also given in Figure 3. Blue points represent features (associated to images) in the metric space. With respect to the center point, the other points have different numbers of identical class labels, which are indicated by different colors. After training with GRN-SNDL, the points associated to images with more labels in common have been dragged closer than the points associated to images with less common labels (with respect to the center point).

The proposed GRN-SNDL loss can be more beneficial to model the local geometry in the feature space, while the class-discrimination capability may not be well preserved. Following our previous work [63], we introduce another loss term based on BCE to further improve the performance of class discrimination. The definition of BCE loss is given by:

$$L_{\mathrm{BCE}} = -\sum_{i}\sum_{c} \mathbb{1}_{\mathbf{y}_i}(c)\log(p_i^c) - (1-\mathbb{1}_{\mathbf{y}_i}(c))\log(1-p_i^c), \tag{10}$$

where p_i^c measures the likelihood of the existence of label c, $\mathbb{1}_{\mathbf{y}_i}(c)$ indicates whether the class c is annotated or not. If the class c is annotated, the c-th element of $\mathbf{y}_i$ is set as 1 (and as 0 otherwise). To this end, we jointly optimize the following loss function:

$$L = L_{\mathrm{SNDL}} + L_{\mathrm{BCE}} \tag{11}$$

2) Optimization algorithm: The optimization of the BCE loss can be conducted by the standard back-propagation. For optimizing the GRN-SNDL loss, we first calculate the gradient with respect to $\mathbf{f}_i$ as indicated in the following equation based on the chain rule:

$$\frac{\partial L_{\text{SNDL}}}{\partial \mathbf{f}_i} = \frac{1}{\sigma}\sum_k p_{ik}\mathbf{f}_k - \frac{1}{\sigma}\sum_k w_{ik}\tilde{p}_{ik}\mathbf{f}_k, \tag{12}$$

where $\tilde{p}_{ik} = p_{ik}/\sum_j w_{ij}p_{ij}$ is the normalized distribution. It can be seen that the feature embeddings of the entire training set are required for the optimization. If we assume that $\mathcal{B}$ is up-to-date during training, the gradient of the loss function with respect to $\mathbf{f}_i$ at the $t+1$-th iteration is:

$$\frac{\partial L_{\text{SNDL}}}{\partial \mathbf{f}_i} = \frac{1}{\sigma}\sum_k p_{ik}\mathbf{f}_k^{(t)} - \frac{1}{\sigma}\sum_k w_{ik}\tilde{p}_{ik}\mathbf{f}_k^{(t)}. \tag{13}$$

Then, θ can be learned by exploiting the back-propagation algorithm as follows:

$$\frac{\partial L_{\text{SNDL}}}{\partial \theta} = \frac{\partial L_{\text{SNDL}}}{\partial \mathbf{f}_i} \times \frac{\partial \mathbf{f}_i}{\partial \theta}. \tag{14}$$

With the feature embeddings $\mathbf{f}_i$ obtained for the current mini-batch and $\mathcal{B}$, we can now update $\mathbf{f}_i$ as:

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

where $\mathbf{f}_i^{(t)}$ denotes the historical feature embeddings stored in $\mathcal{B}$, and m is a regularization parameter for updating $\mathbf{f}_i$ based on the empirical weighted average. As described in Equation (15), only the feature embeddings associated to the current mini-batch are updated within the current iteration. The optimization scheme is described in Algorithm 1.

III. Experiments

A. Dataset description

In this paper, three challenging multi-label RS image datasets are utilized to validate the performance of the proposed method. A detailed description of the considered datasets is provided below:

1) *UC Merced (UCM) multi-label dataset* [64]: This dataset is recreated from the original UCM dataset [65] by relabeling all the 2100 aerial images of 256×256 pixels with multiple

Algorithm 1 The optimization scheme for GRN

```
Require: Training images $\mathbf{x}_i$, the weight matrix $\mathbf{W}$, and the multi-label annotations $\mathbf{y}_i$
Randomly initialize the parameters $\theta$ of CNN model, and the memory bank $\mathcal{B}$, as well as the the temperature parameter $\sigma$, the dimensionality $D$, and the regularization parameter $m$.
for The epoch number $t = 0$ to $\mathrm{maxEpoch}$ do
    Sample a mini-batch.
    Obtain the normalized features $\mathbf{f}_i^{(t)}$ based on the CNN model with $\theta^{(t)}$.
    Calculate the similarities $s_{ij}$ with reference to $\mathcal{B}$.
    Calculate the weights $w_{ij}$ based on Equation (8).
    Calculate the gradients of SNDL based on Equation (13) (and the ones of BCE).
    Back-propagate the gradients.
    Update the feature embeddings of the current mini-batch stored in $\mathcal{B}$ via Equation (15).
end for
Ensure: $\theta$, $\mathcal{B}$
```

semantic annotations. The original UCM dataset consists of 21 scene classes, and each class contains 100 images. The newly defined labels are 17 object classes: airplane, sand, pavement, building, car chaparral, court, tree, dock, tank, water, grass, mobile home, ship, bare soil, sea, and field. Figure 4 illustrates some multi-label examples from this dataset.

2) *Aerial image database (AID) multi-label dataset* [66]: This dataset is built upon the original AID dataset [19], which is specially dedicated to aerial image classification. The original AID dataset consists of $10,000$ RGB images belonging to 30 scene classes. The number of images per class ranges from 220 to 420, and the spatial resolution varies from 0.5m to 8m. 3000 aerial images are selected to construct the AID multi-label dataset. The newly defined labels are the same as those in the UCM multi-label dataset. Some examples of multi-label annotations are given in Figure 5.

3) *DFC15 multi-label dataset* [66]: This dataset is created from a semantic segmentation dataset called DFC15[2] and acquired over Zeebrugge, Belgium, using an airborne sensor with spatial resolution of 5 cm. The DFC15 multi-label dataset consists of 3342 images

[2] 2015 IEEE GRSS data fusion contest. http://www.grss-ieee.org/community/technical-committees/data-fusion/2015-ieee-grss-data-fusion-contest/

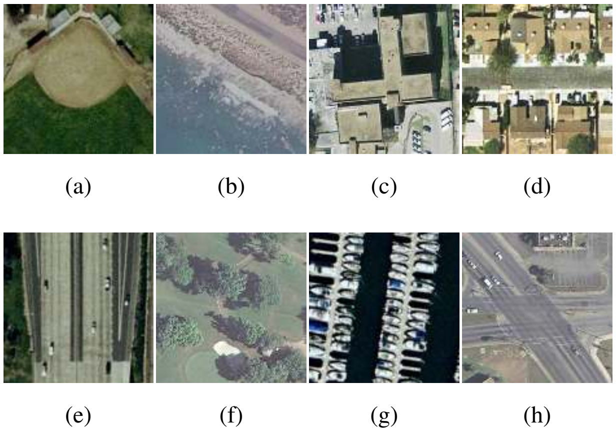


Fig. 4. Examples of the UCM multi-label dataset. (a) Bare-soil, Buildings, Grass, (b) Pavement, Sand, Sea. (c) Buildings, Cars, Grass, Pavement. (d) Bare-soil, Buildings, Cars, Pavement, Trees. (e) Cars, Grass, Pavement. (f) Bare-soil, Grass, Pavement, Sand, Trees. (g) Dock, Ship, Water. (h) Bare-soil, Buildings, Cars, Grass, Pavement, Trees.

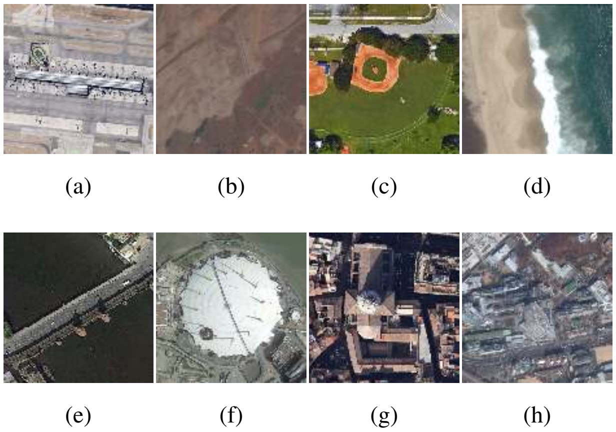


Fig. 5. Examples of the AID multi-label dataset. (a) Airplane, Bare-soil, Buildings, Cars, Grass, Pavement. (b) Bare-soil, Buildings, Cars, Grass, Pavement, Trees. (c) Bare-soil, Buildings, Grass, Pavement, Trees. (d) Chaparral, Sand, Sea. (e) Buildings, Cars, Dock, Pavement, Ship, Trees, Water. (f) Bare-soil, Buildings, Car, Grass, Pavement, Trees. (g) Buildings, Cars, Pavement. (h) Bare-soil, Buildings, Cars, Grass, Pavement, Trees.

and there are 8 object classes: impervious, water, clutter, vegetation, building, tree, boat and car. Figure 6 displays some images with the associated multi-labels.

B. *Experimental setup*

The effectiveness of the proposed approach to categorize multi-label RS scene images is evaluated on two different tasks: 1) image classification and 2) image retrieval. The following sections describe in detail the experimental setup considered for each task.

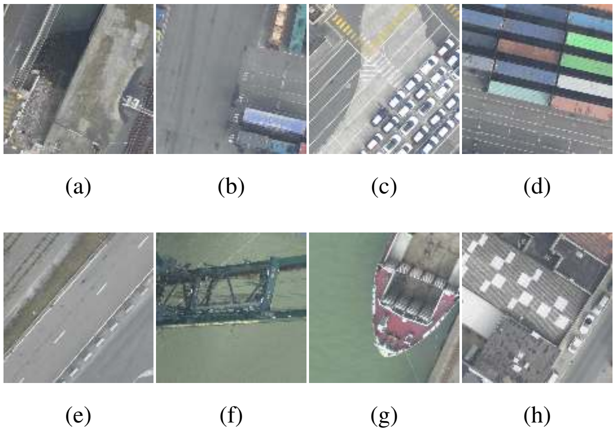


Fig. 6. Examples of the DFC15 multi-label dataset. (a) Impervious, Water, Clutter. (b) Impervious, Clutter. (c) Impervious, Building, Car. (d) Impervious, Clutter. (e) Impervious, Clutter, Vegetation. (f) Water, Clutter. (g) Impervious, Water, Clutter. (h) Impervious, Building, Car.

1) Multi-label RS image classification: For an out-of-sample image $\mathbf{x}^*$, its feature embedding $\mathbf{f}^*$ can be obtained by applying $\mathcal{F}(\cdot)$ with the learned parameter set θ. Its predicted label vector $\mathbf{y}^*$ can be determined by thresholding the mean average of the label vectors of its K nearest neighbors in $\mathcal{B}$ using the value 0.5. We exploit four metrics to evaluate the classification performance, including: 1) sample F1 score (F_s^1), 2) sample F2 score (F_s^2), 3) sample precision (P_s), and 4) sample recall (R_s). Specifically, the sample F1 and F2 scores are defined as:

$$F_s^b = (1+b^2)\frac{P_s R_s}{b^2 P_s + R_s}, \quad b = 1, 2, \tag{16}$$

where P_s and R_s are the sample-based precision and recall, respectively. They are calculated based on:

$$P_s = \frac{\mathrm{TP}_s}{\mathrm{TP}_s + \mathrm{FP}_s}, \quad R_s = \frac{\mathrm{TP}_s}{\mathrm{TP}_s + \mathrm{FN}_s}, \tag{17}$$

where TP_s, FP_s and FN_s are the sample-based true positives, false positives and false negatives, respectively.

2) Multi-label RS image retrieval: Image retrieval aims to find the most semantically similar images in the dataset, based on the distances calculated on their feature embeddings with respect to those of a query image. Given such query image, a more effective metric learning method can lead to more relevant images retrieved from the dataset. Under a multi-label RS scheme, we evaluate the image retrieval quality based on three metrics: 1) Weighted Mean Average Precision

(WMAP) [67], 2) Mean Average Precision (MAP) [68], [69], and 3) Hamming Loss (HL). To be specific, WMAP is calculated as:

$$\text{WMAP} = \frac{1}{|\mathcal{Q}|}\sum_{q=1}^{|\mathcal{Q}|}\Big(\frac{1}{N_{Rel}(q)@R}\sum_{r=1}^{R}\big(\delta(q,r)\times \text{ACG}@r\big)\Big), \tag{18}$$

where $\mathcal{Q}$ denotes the query set, R represents the number of inspected images from the top-ranking, $N_{Rel}(q)@R$ indicates the total number of relevant images (with respect to the query image $\mathbf{x}_q$) within the top R retrieved images, $\delta(q,r)$ is an indicator function that indicates whether the r-th retrieved image from the top-ranking is truly relevant to the query image $\mathbf{x}_q$, (i.e., if there is at least one common class annotated to both images $\mathbf{x}_q$ and $\mathbf{x}_r$, $\delta(q,r)$ is set to 1 [relevant] and 0 [non-relevant] otherwise) and $\text{ACG}@r$ denotes the Average Cumulative Gains (ACG) [70] score of the first r retrieved images, which is defined as:

$$\text{ACG}@r = \frac{1}{r}\sum_{i}^{r}\text{Sim}(q,i). \tag{19}$$

Here, $\text{Sim}(q,i)$ is the number of shared labels between image $\mathbf{x}_q$ and image $\mathbf{x}_i$, and MAP is the mean of the average precision for each query image, defined by:

$$\text{MAP} = \frac{1}{|\mathcal{Q}|}\sum_{q=1}^{|\mathcal{Q}|}\text{AP}(q), \tag{20}$$

where

$$\text{AP}(q) = \frac{1}{N_{Rel}(q)@R}\sum_{r=1}^{R}\big(\delta(q,r)\times\frac{N_{Rel}(q)@r}{r}\big). \tag{21}$$

HL evaluates the fraction of labels that are incorrectly predicted, which is given by:

$$\text{HL}(\mathbf{y},\hat{\mathbf{y}}) = \frac{1}{C}\sum_{c}\delta(\hat{y}_c \neq y_c), \tag{22}$$

where $\hat{\mathbf{y}}$ is the predicted label vector and $\hat{y}_c$ denotes its c-th element.

We randomly select 70% of the images for training, 10% for validation and 20% for testing from the three benchmark datasets. For image retrieval purposes, the test set is utilized as the query set, and the relevant images are retrieved from the training set. The proposed method is implemented in PyTorch. All the images are resized to 256×256 pixels, and three data augmentation strategies are adopted during training: 1) *RandomGrayscale*, 2) *ColorJitter*, and 3) *RandomHorizontalFlip*. The parameters D, σ and m are set to 128, 0.1 and 0.5, respectively. The

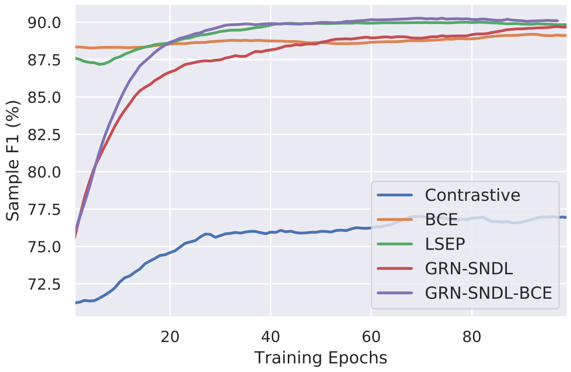
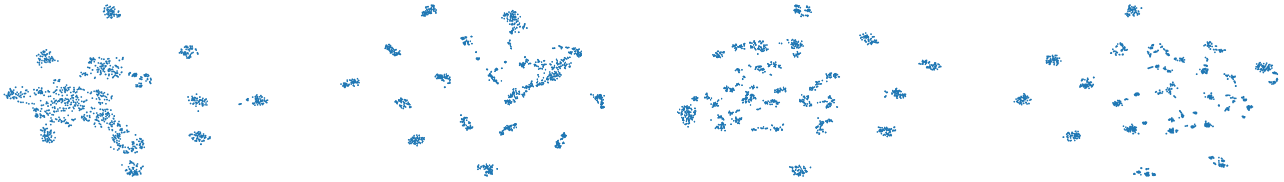


Fig. 7. Learning curves obtained after training ResNet18 with Contrastive, BCE, LSEP, GRN-SNDL and GRN-SNDL-BCE losses on the AID multi-label dataset. We display F_s^1 (%) in the validation set as a function of the number of training epochs.

stochastic gradient descent (SGD) optimizer is employed for training the CNN model with an initial learning rate set to 0.01, which is decayed by 0.5 every 30 epochs. The batch size is set to 256, and we train the CNN model for 100 epochs. To validate the effectiveness of the proposed framework for multi-label deep metric learning, we compare it with: 1) BCE loss [46], [48], [71], 2) contrastive loss [53], [72], and 3) LSEP loss [56]. Additionally, we test several prevalent backbone architectures in RS: 1) ResNet18 [61], 2) ResNet50 [61], and 3) WideResNet50 [62]. For optimizing other loss functions, the associated learning rates are selected based on cross-validation. In order to construct image pairs with multi-label annotations for the contrastive loss, we consider the image pairs sharing at least one common label as positive pairs, and the other pairs (without any labels in common) as negative pairs. It is worth noting that the multi-label information of the DFC15 dataset is not appropriate to construct pairwise labels for the contrastive loss. Thus, the experiments of the contrastive loss on the DFC15 dataset are omitted here. All the experiments have been conducted on an NVIDIA Tesla P100 GPU.

C. Experimental results

1) Multi-label RS image classification: Figure 7 shows the learning curves obtained for ResNet18, optimized with the considered losses (including Contrastive, BCE, LSEP, GRN-SNDL and GRN-SNDL-BCE) on the AID dataset. Using the KNN classifier with $K = 10$, we calculate the sample F1 scores $(\%)$ on the validation set and plot them versus the number of training epochs. It can be seen that, in the first 20 epochs, ResNet18 trained with the BCE and



(a) (b) (c) (d)

Fig. 8. 2D projection of the feature embeddings on the UCM training set using t-SNE: (a) WideResNet50-BCE; (b) WideResNet50-GRN-SNDL; (c) WideResNet50-LSEP; and (d) WideResNet50-GRN-SNDL-BCE.

TABLE I

KNN CLASSIFICATION PERFORMANCES OBTAINED BY DIFFERENT CNN MODELS OPTIMIZED WITH CONTRASTIVE, BCE, LSEP, GRN-SNDL AND GRN-SNDL-BCE LOSSES ON THE TEST SETS. THE PERFORMANCES ARE EVALUATED USING FOUR DIFFERENT METRICS: F_s^1, F_s^2, P_s AND R_s (%).

		UCM				**AID**				**DFC15**			
		F_s^1	F_s^2	P_s	R_s	F_s^1	F_s^2	P_s	R_s	F_s^1	F_s^2	P_s	R_s
ResNet18	Contrastive	64.67	64.63	69.49	65.73	75.43	74.75	80.78	75.67	–	–	–	–
	BCE	87.76	88.23	89.19	89.19	88.31	87.42	91.77	87.25	92.74	91.88	95.38	91.55
	GRN-SNDL	88.47	88.87	89.82	89.76	89.13	88.43	92.39	88.37	93.08	92.47	95.19	92.32
	LSEP	88.75	89.37	89.65	90.40	89.78	89.13	92.70	89.17	92.91	92.04	95.58	91.72
	GRN-SNDL-BCE	**89.82**	**90.26**	**91.06**	**91.15**	**90.26**	**89.66**	**93.11**	**89.63**	**94.72**	**94.35**	**96.09**	**94.29**
ResNet50	Contrastive	77.02	76.78	81.50	77.49	76.31	75.88	80.42	76.88	–	–	–	–
	BCE	89.73	90.61	90.10	91.76	89.18	88.45	92.32	88.37	93.95	93.39	95.88	93.24
	GRN-SNDL	89.68	90.11	90.71	90.91	90.43	89.88	**93.27**	89.95	94.53	94.27	95.79	94.31
	LSEP	90.36	91.26	90.57	92.43	89.43	88.65	92.74	88.58	93.52	93.03	95.49	92.96
	GRN-SNDL-BCE	**91.31**	**91.92**	**91.98**	**92.83**	**90.95**	**90.82**	92.79	**91.08**	**95.80**	**95.78**	**96.53**	**95.95**
WideResNet50	Contrastive	74.84	74.99	78.36	76.08	81.06	80.30	85.48	80.59	–	–	–	–
	BCE	88.45	88.89	89.75	89.76	89.36	88.67	92.41	88.63	93.39	92.94	95.29	92.88
	GRN-SNDL	90.31	90.81	91.21	91.68	90.55	89.93	93.39	89.89	94.81	94.46	96.31	94.44
	LSEP	90.22	90.81	91.13	91.79	89.40	88.56	92.87	88.44	93.68	92.92	96.11	92.67
	GRN-SNDL-BCE	**90.81**	**91.18**	**91.97**	**91.92**	**91.02**	**90.50**	**93.66**	**90.49**	**95.96**	**95.73**	**97.13**	**95.74**

LSEP losses achieve higher classification accuracies than both GRN-SNDL and GRN-SNDL-BCE. However, the performances of the BCE and LSEP losses are relatively stable during the whole training phase. This fact indicates that the effectiveness of the metric learning based on the these two losses is less obvious than the proposed losses. Moreover, as the learning curves converge, better KNN classification results can be obtained when we use the GRN-SNDL-BCE loss (instead of the other losses) for optimization.

TABLE II

Some KNN classification examples associated with the ground-truth and the predicted labels. The false positives are marked in red, and the false negatives are marked in blue.

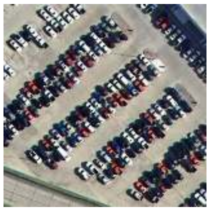
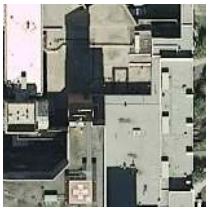
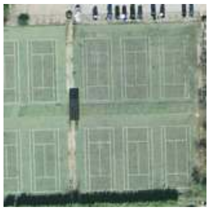
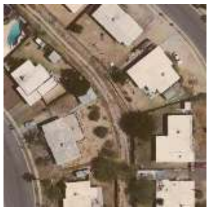
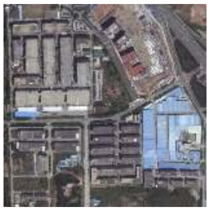
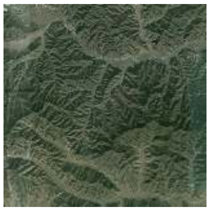
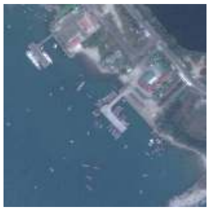
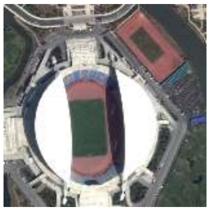
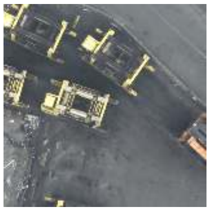
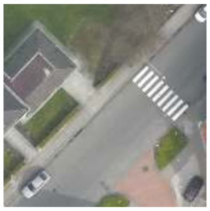
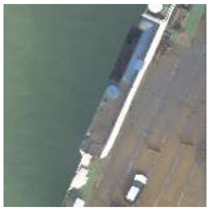
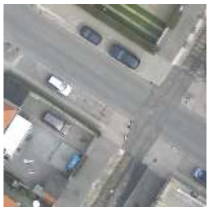
UCM images				
Ground-truth labels	Cars, Pavement	Buildings, Trees	Bare-soil, Cars, Court, Pavement, Trees	Bare-soil, Buildings, Cars, Grass, Pavement, Trees
Predicted labels	Cars, Trees, Pavement	Buildings, Trees	Bare-soil, Court, Grass, Pavement, Trees	Bare-soil, Buildings, Cars, Pavement, Trees
AID images				
Ground-truth labels	Bare-soil, Buildings, Cars, Grass, Pavement, Trees	Grass, Trees	Bare-soil, Buildings, Cars, Dock, Grass, Pavement, Sea, Ship	Bare-soil, Buildings, Cars, Court, Grass, Pavement, Trees, Water
Predicted labels	Bare-soil, Buildings, Cars, Grass, Pavement, Trees	Grass, Trees	Bare-soil, Buildings, Cars, Dock, Grass, Pavement, Sea, Ship, Trees	Buildings, Cars, Court, Grass, Pavement, Trees
DFC15 images				
		Impervious,		Impervious,

TABLE III

Image retrieval performances obtained by different CNN models optimized via the Contrastive, BCE, LSEP, GRN-SNDL and GRN-SNDL-BCE losses on the test sets. The performances are evaluated with the metrics: WMAP, MAP (%) and HL.

		UCM			AID			DFC15		
		WMAP	MAP(%)	HL	WMAP	MAP(%)	HL	WMAP	MAP(%)	HL
ResNet18	Contrastive	1.97	86.77	0.19	3.66	93.31	0.18	–	–	–
	BCE	2.52	97.70	0.13	4.25	97.36	0.12	2.37	100.00	0.12
	GRN-SNDL	2.63	99.17	0.11	4.35	99.17	0.11	2.43	100.00	0.10
	LSEP	**2.75**	**99.79**	**0.09**	4.39	99.06	0.11	2.40	100.00	0.11
	GRN-SNDL-BCE	2.71	99.70	0.10	**4.47**	**99.29**	**0.09**	**2.51**	100.00	**0.07**
ResNet50	Contrastive	2.28	97.02	0.15	3.85	93.44	0.17	–	–	–
	BCE	2.64	98.99	0.11	4.33	98.31	0.11	2.45	100.00	0.09
	GRN-SNDL	2.71	99.64	0.10	4.47	**99.67**	0.09	2.51	99.99	0.08
	LSEP	2.77	99.81	0.09	4.40	99.52	0.10	2.46	100.00	0.09
	GRN-SNDL-BCE	**2.80**	**99.92**	**0.08**	**4.60**	99.66	**0.07**	**2.58**	100.00	**0.06**
WideResNet50	Contrastive	2.22	96.47	0.15	3.99	95.49	0.15	–	–	–
	BCE	2.62	99.37	0.11	4.39	98.93	0.10	2.45	**100.00**	0.09
	GRN-SNDL	2.73	99.44	0.10	4.48	99.49	0.09	2.53	99.95	0.07
	LSEP	2.76	**99.87**	0.09	4.40	**99.75**	0.10	2.47	**100.00**	0.09
	GRN-SNDL-BCE	**2.80**	**99.87**	**0.08**	**4.59**	**99.75**	**0.07**	**2.57**	99.99	**0.06**

In order to visualize the learned feature embeddings in the metric space, we exploit t-distributed stochastic neighbour embedding (t-SNE) to visualize their projections on a 2D plane. Figure 8 shows the t-SNE scatter plots of the feature embeddings in the UCM training set, obtained using: (a) BCE, (b) GRN-SNDL, (c) LSEP, and (d) GRN-SNDL-BCE with WideResNet50. As we can observe, the proposed method is able to uncover a remarkably finer-grained neighborhood structure by comparing (a) and (b). This is because, with the proposed GRN-SNDL, those images that are semantically similar tend to be grouped together, while dissimilar RS scenes are farther separated than with the BCE loss. By jointly using the SNDL and BCE losses, the class-discrimination capability can be further improved with respect to GRN-SNDL. It can be seen that the mixed group of images shown in (b) can be separated farther away in (d). Moreover, within some groups, the images are located closer in (d) than (c). That is to say, the proposed GRN-SNDL-BCE loss can both discover the locality structure of the images in the metric space and preserve the class-discrimination capability.

Table I illustrates the performance of all the CNN models (trained with all the considered losses) on the test sets of the three considered benchmark datasets. All the results are based on a KNN classifier with $K = 10$. It can be observed that the performance achieved by the proposed GRN-SNDL-BCE on the three datasets is generally better than the one achieved by the other compared losses. For example, the sample F1 score of ResNet18-GRN-SNDL-BCE exhibits around 1% and 2% performance improvements over ResNet18-LSEP and ResNet18-BCE, respectively, on the UCM dataset. Based on the ResNet50 model, the BCE loss can achieve the comparable classification performance with respect to the GRN-SNDL-BCE loss with the ResNet18 model.

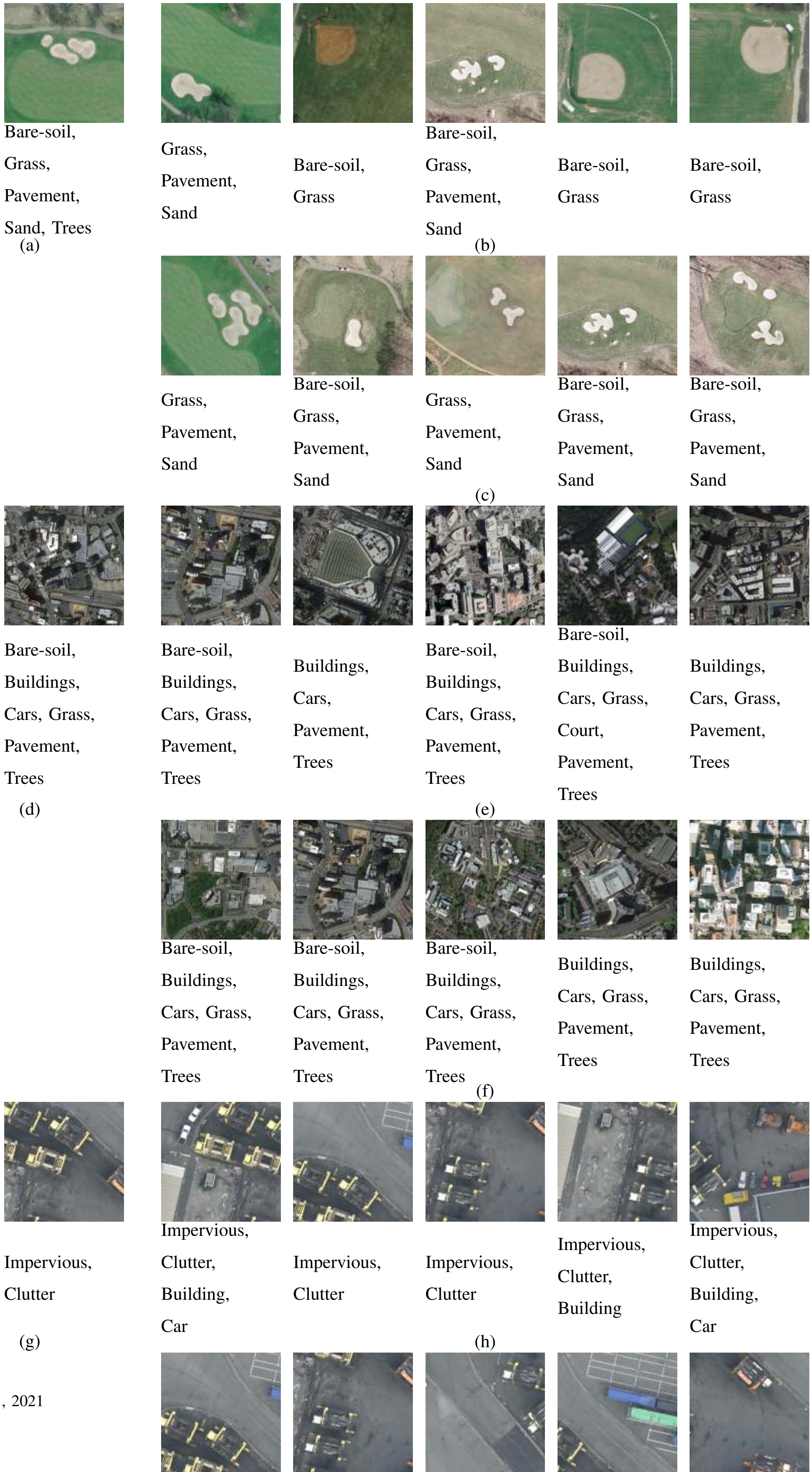
Moreover, as the CNN model becomes deeper and wider, the classification accuracies obtained by all the losses improve. As the BCE loss is optimized for aligning all the images from each category to each parameterized prototype, the ability to capture the relationships among the images is lacking. Thus, the BCE loss cannot sufficiently learn the metric space, where semantically similar images need to be grouped together. In contrast, the proposed method can effectively model the relationships among all the RS images by constructing a weight matrix based on their multi-label information. If two images have multiple classes in common, their similarity metric is granted with a heavier weight. By optimizing the associated GRN-SNDL loss, a metric space can be learned through training, and images with more common classes are pulled closer. Therefore, the proposed loss can better discover their inherent locality structures of the images within the metric space, which leads to better KNN classification performance.

Table II illustrates some predicted examples using the WideResNet50 model optimized by the GRN-SNDL-BCE loss. It can be seen that most classes can be correctly classified, while there are still some false positive and false negative predictions (marked in red and blue, respectively). For the third image in the UCM dataset, *grass* is a false positive (due to its analogous appearance with regards to *court*). Similarly, *Trees* is also positively predicted in the third image of the AID dataset, since the pattern of *grass* on its upper-leftmost corner is analogous with *trees*. *Water* is not successfully distinguished in the fourth image of the AID dataset, since its RGB spectral values are close to those of *grass* in the same image.

2) Multi-label RS image retrieval: Table III presents the quantitative retrieval results obtained by different CNN models, trained with all the losses. Consistently with the KNN classification results, our GRN-SNDL demonstrates its superiority over the BCE loss on all the considered CNN models. For example, with ResNet18, the MAP score obtained using the GRN-SNDL loss is higher than that obtained by the BCE loss, with an improvement of more than 1%. This fact indicates that, in the learned metric space based on the proposed GRN-SNDL, more relevant images (or images with more common labels with regards to the query image) can be retrieved (as compared to the metric space produced by the BCE). When focusing on LSEP, GRN-SNDL-BCE is also able achieve higher retrieval performances on all the benchmark datasets. In order to improve multi-label classification accuracy, LSEP is targeted at minimizing the produced label confidence scores in a pairwise manner, where the the scores of the true labels should be greater than those of the negative labels. However, the feature emebddings from images with multiple annotations are not directly considered in the LSEP loss. In other words, the feature embeddings of the images sharing more common annotations should be logically closer than the others in the feature space, however this aspect is not directly optimized in LSEP. In contrast, the proposed loss functions are able to exploit this property throughout a novel graph relation network, which is eventually able to provide superior retrieval results than LSEP. Moreover, the GRN-SNDL-BCE loss can generally achieve the best performance in terms of image retrieval with all the considered CNN models.

Figure 9 shows the top 5 retrieved images based on ResNet50-LSEP and ResNet50-GRN-SNDL-BCE with respect to the associated query images, where (a), (d) and (g) are the query images from the UCM, AID and DFC15 multi-label datasets, respectively, (b), (e) and (h) are the retrieved images based on ResNet50-LSEP, and (c), (f) and (i) are the retrieved images based on ResNet50-GRN-SNDL-BCE. Although there are some common classes between the retrieved images and the query images in all the results, ResNet50-GRN-SNDL-BCE can capture the images with more relevant classes as the nearest neighbors to the query image. Moreover, by measuring the relationship among the images during the training, ResNet50-GRN-SNDL-BCE can order the nearest neighbors with respect to the query image better than ResNet18-LSEP, where the images sharing more identical classes with the query image have the higher priority to be retrieved first.

3) Parameter Sensitivity Analysis: D and σ are the two main parameters of the proposed framework. With ResNet18, in Table IV we calculate the F_s^1 (%) of the KNN classification results on the test sets (for the three considered datasets) with respect to varying values of D, setting $K = 10$. It can be observed that the performances obtained using different values of D are stable on all the considered datasets. In other words, the proposed GRN-SNDL loss is robust to the use of different dimensional sizes of the learned feature embeddings. This characteristic is


Bare-soil, Grass, Pavement, Sand, Trees
Grass, Pavement, Sand
Bare-soil, Grass
Bare-soil, Grass, Pavement, Sand
Bare-soil, Grass
Bare-soil, Grass

(a) (b)

Grass, Pavement, Sand
Bare-soil, Grass, Pavement, Sand
Grass, Pavement, Sand
Bare-soil, Grass, Pavement, Sand
Bare-soil, Grass, Pavement, Sand

(c)

Bare-soil, Buildings, Cars, Grass, Pavement, Trees
Bare-soil, Buildings, Cars, Grass, Pavement, Trees
Buildings, Cars, Pavement, Trees
Bare-soil, Buildings, Cars, Grass, Pavement, Trees
Bare-soil, Buildings, Cars, Grass, Court, Pavement, Trees
Buildings, Cars, Grass, Pavement, Trees

(d) (e)

Bare-soil, Buildings, Cars, Grass, Pavement, Trees
Bare-soil, Buildings, Cars, Grass, Pavement, Trees
Bare-soil, Buildings, Cars, Grass, Pavement, Trees
Buildings, Cars, Grass, Pavement, Trees
Buildings, Cars, Grass, Pavement, Trees

(f)

Impervious, Clutter
Impervious, Clutter, Building, Car
Impervious, Clutter
Impervious, Clutter
Impervious, Clutter, Building
Impervious, Clutter, Building, Car

(g) (h)

TABLE IV

SENSITIVITY ANALYSIS OF PARAMETER D IN THE PROPOSED MODEL (GRN-SNDL) BASED ON THE F_s^1 (%) OF THE KNN CLASSIFICATION.

	UCM	**AID**	**DFC15**
$D = 32$	87.68	89.33	92.72
$D = 64$	87.52	88.56	92.98
$D = 128$	88.47	89.13	93.08

TABLE V

SENSITIVITY ANALYSIS OF PARAMETER σ IN THE PROPOSED MODEL (GRN-SNDL) BASED ON THE F_s^1 (%) OF THE KNN CLASSIFICATION.

	UCM	**AID**	**DFC15**
$\sigma = 0.05$	87.83	89.60	93.97
$\sigma = 0.1$	88.47	89.13	93.08
$\sigma = 0.15$	87.04	86.90	91.60
$\sigma = 0.2$	85.71	85.76	92.34

greatly beneficial for developing image classification or retrieval systems on scalable RS archives, where the storage space of the feature embeddings needs to be optimized.

Using the same settings adopted to report the results in Table IV, Table V shows a sensitivity analysis of GRN-SNDL in terms of parameter σ, with a range from 0.05 to 0.2. In this case, we can observe that the classification performances are better when σ equals 0.05 or 0.1. Therefore, we conclude that highly satisfactory results can be reached by the proposed approach function when σ is in the range from 0.05 to 0.1.

IV. CONCLUSIONS AND FUTURE LINES

In this paper, we introduce a graph relation network based on a newly developed loss function (GRN-SNDL) which has been specially designed to classify and retrieve RS scene images considering multiple semantic annotations. The proposed approach pursues to pull the most semantically similar RS images closer in the metric space when they share more classes in common, from a multi-label perspective. In order to achieve this goal, we stochastically maximize a weighted leave-one-out KNN score of the training set, where the corresponding weight matrix is obtained from the multi-label semantic information that describes the contributions of the nearest neighbors associated with each image on its class decision. To further preserve the class-discrimination capability, we also propose a joint loss function by combining SNDL and BCE. In order to validate the effectiveness of the proposed scheme, we conduct extensive experiments on two different RS processing tasks, i.e. image classification and image retrieval, using three multi-label benchmark datasets: UCM, AID and DFC15. Compared with the state-of-the-art losses for multi-label RS scene categorization (including BCE and LSEP), the proposed losses exhibit better classification accuracy, with an improvement of around 2% and 1% with regards to the BCE and LSEP losses, respectively. Moreover, the learned feature embeddings based on our approach manifest a very promising performance on the RS image retrieval task. With the ResNet18 model, the MAP scores on the three benchmark datasets can be improved in around 2% with respect to the use of BCE. In summary, the proposed model is able to provide not only superior performance for RS image classification, but also to preserve the neighborhood structures among the RS images in the learned metric space, which is guided by the multi-label information.

Due to the remarkable potential of the presented method for multi-label RS image classification and retrieval, our future work will be directed towards adapting our framework to other relevant RS tasks, such as dimensionality reduction or fine-grained land-use categorization. Moreover, we plan to investigate the graph CNN (GCN) [73] for deep metric learning of RS images with the guidance of the semantic information among the word embeddings of the multi-label annotations. We are also interested in exploring further developments in terms of efficiency.

Acknowledgment

The authors would like to thank the authors for their efforts in creating the multi-label datasets based on UCM, AID and DFC15, and the reviewers for their valuable suggestions.

References

[1] L. Gao, B. Zhao, X. Jia, W. Liao, and B. Zhang, “Optimized kernel minimum noise fraction transformation for hyperspectral image classification,” *Remote Sens.*, vol. 9, no. 6, p. 548, 2017.

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

[3] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. A. Benediktsson, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep (overview and toolbox)," *IEEE Geosci. Remote Sens. Mag.*, 2020, dOI: 10.1109/MGRS.2020.2979764.

[4] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 35–49, 2019.

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

[6] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, 2016.

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

[8] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, 2020.

[9] 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.

[10] 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*, 2019.

[11] T. R. Martha, N. Kerle, C. J. van Westen, V. Jetten, and K. V. Kumar, "Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4928–4943, 2011.

[12] A. Stumpf and N. Kerle, "Object-oriented mapping of landslides using random forests," *Remote Sens. Environ.*, vol. 115, no. 10, pp. 2564–2577, 2011.

[13] J. Kang, D. Hong, J. Liu, G. Baier, N. Yokoya, and B. Demir, "Learning convolutional sparse coding on complex domain for interferometric phase restoration," *IEEE Trans. Neural Netw. Learn. Syst.*, 2020, dOI: 10.1109/TNNLS.2020.2979546.

[14] L. Gao, D. Hong, J. Yao, B. Zhang, P. Gamba, and J. Chanussot, "Spectral superresolution of multispectral imagery with joint sparse and low-rank learning," *IEEE Trans. Geosci. Remote Sens.*, 2020, dOI:10.1109/TGRS.2020.3000684.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 4, no. 1, pp. 193–204, 2011.

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

[22] F. Zhang, B. Du, and L. Zhang, “Saliency-guided unsupervised feature learning for scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, 2014.

[23] D. Hong, N. Yokoya, and X. X. Zhu, “Learning a robust local manifold representation for hyperspectral dimensionality reduction,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 6, pp. 2960–2975, 2017.

[24] S. Chaib, H. Liu, Y. Gu, and H. Yao, “Deep feature fusion for vhr remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, 2017.

[25] W. Han, R. Feng, L. Wang, and Y. Cheng, “A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 23–43, 2018.

[26] Q. Wang, S. Liu, J. Chanussot, and X. Li, “Scene classification with recurrent attention of VHR remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, 2018.

[27] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, “Remote sensing image fusion using hierarchical multimodal probabilistic latent semantic analysis,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 12, pp. 4982–4993, 2018.

[28] D. Hong, N. Yokoya, G. Xia, J. Chanussot, and X. Zhu, “X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data,” *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, 2020.

[29] R. Fernandez-Beltran, B. Demir, F. Pla, and A. Plaza, “Unsupervised remote sensing image retrieval using probabilistic latent semantic hashing,” *IEEE Geosci. Remote Sens. Lett.*, 2020, dOI: 10.1109/LGRS.2020.2969491.

[30] H. Yu, L. Gao, W. Liao, B. Zhang, A. Pižurica, and W. Philips, “Multiscale superpixel-level subspace-based support vector machines for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2142–2146, 2017.

[31] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. Zhu, “Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, 2020.

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

[33] B. Du, Z. Wang, L. Zhang, L. Zhang, and D. Tao, “Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion,” *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1694–1707, 2017.

[34] Y. Liu, B. Du, W. Tu, M. Gong, Y. Guo, and D. Tao, “Logdet metric-based domain adaptation,” *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[35] X. Li, B. Du, C. Xu, Y. Zhang, L. Zhang, and D. Tao, “Robust learning with imperfect privileged information,” *Artificial Intelligence*, vol. 282, p. 103246, 2020.

[36] Y. Dong, B. Du, L. Zhang, and L. Zhang, “Dimensionality reduction and classification of hyperspectral images using ensemble discriminative local metric learning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2509–2524, 2017.

[37] K. Karalas, G. Tsagkatakis, M. Zervakis, and P. Tsakalides, “Land classification using remotely sensed data: Going multilabel,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3548–3563, 2016.

[38] S. Koda, A. Zeggada, F. Melgani, and R. Nishii, “Spatial and structured SVM for multilabel image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5948–5960, 2018.

[39] B. T. Zegeye and B. Demir, “A novel active learning technique for multi-label remote sensing image scene classification,” in *Image and Signal Processing for Remote Sensing XXIV*, vol. 10789. International Society for Optics and Photonics, 2018, p. 107890B.

[40] A. Zeggada, S. Benbraika, F. Melgani, and Z. Mokhtari, “Multilabel conditional random field classification for UAV images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 399–403, 2018.

[41] W. Zhou, S. Newsam, C. Li, and Z. Shao, “Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 197–209, 2018.

[42] X. Deng, P. Liu, X. Liu, R. Wang, Y. Zhang, J. He, and Y. Yao, “Geospatial big data: New paradigm of remote sensing applications,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 10, pp. 3841–3851, 2019.

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

[44] G. Sumbul, J. Kang, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, and B. Demir, “BigEarthNet deep learning models with a new class-nomenclature for remote sensing image understanding,” *arXiv preprint arXiv:2001.06372*, 2020.

[45] K. Karalas, G. Tsagkatakis, M. Zervakis, and P. Tsakalides, “Deep learning for multi-label land cover classification,” in *Image and Signal Processing for Remote Sensing XXI*, vol. 9643. International Society for Optics and Photonics, 2015, p. 96430Q.

[46] A. Zeggada, F. Melgani, and Y. Bazi, “A deep learning approach to UAV image multilabeling,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 694–698, 2017.

[47] D. Landgrebe, “Hyperspectral image data analysis,” *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, 2002.

[48] R. Stivaktakis, G. Tsagkatakis, and P. Tsakalides, “Deep learning for multilabel land cover scene categorization using data augmentation,” *IEEE Geosci. Remote Sens. Lett.*, 2019.

[49] Y. Hua, L. Mou, and X. X. Zhu, “Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification,” *ISPRS journal of photogrammetry and remote sensing*, vol. 149, pp. 188–199, 2019.

[50] G. Sumbul and B. Demir, “A novel multi-attention driven system for multi-label remote sensing image classification,” in *Proc. IGARSS*. IEEE, 2019, pp. 5726–5729.

[51] A. Alshehri, Y. Bazi, N. Ammour, H. Almubarak, and N. Alajlan, “Deep attention neural network for multi-label classification in unmanned aerial vehicle imagery,” *IEEE Access*, vol. 7, pp. 119 873–119 880, 2019.

[52] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, “Multilabel remote sensing image retrieval based on fully convolutional network,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 318–328, 2020.

[53] 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.

[54] Z. Wu, A. A. Efros, and S. X. Yu, “Improving generalization via scalable neighborhood component analysis,” in *Proc. ECCV*, 2018, pp. 685–701.

[55] D. Hong, W. Liu, J. Su, Z. Pan, and G. Wang, “A novel hierarchical approach for multispectral palmprint recognition,” *Neurocomputing*, vol. 151, pp. 511–521, 2015.

[56] Y. Li, Y. Song, and J. Luo, “Improving pairwise ranking for multi-label image classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3617–3625.

[57] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, “Neighbourhood components analysis,” in *Proc. NIPS*, 2005, pp. 513–520.

[58] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, “Cospace: Common subspace learning from hyperspectral-multispectral correspondences,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, 2019.

[59] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[60] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. Plaza, “Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast,” *IEEE Transactions on Geoscience and Remote Sensing*, 2020, dOI:10.1109/TGRS.2020.3007029.

[61] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016, pp. 770–778.

[62] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.

[63] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, “Deep metric learning based on scalable neighborhood components for remote sensing scene characterization,” *IEEE Transactions on Geoscience and Remote Sensing*, 2020, dOI:10.1109/TGRS.2020.2991657.

[64] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, “Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, 2017.

[65] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. ACM SIGSPATIAL*. ACM, 2010, pp. 270–279.

[66] Y. Hua, L. Mou, and X. X. Zhu, “Label relation inference for multi-label aerial image classification,” in *Proc. IGARSS*. IEEE, 2019, pp. 5244–5247.

[67] F. Zhao, Y. Huang, L. Wang, and T. Tan, “Deep semantic ranking based hashing for multi-label image retrieval,” in *Proc. CVPR*, 2015, pp. 1556–1564.

[68] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[69] Z. Zhang, Q. Zou, Y. Lin, L. Chen, and S. Wang, “Improved deep hashing with soft pairwise similarity for multi-label image retrieval,” *IEEE Trans. Multimedia*, 2019.

[70] K. Järvelin and J. Kekäläinen, “IR evaluation methods for retrieving highly relevant documents,” in *Proc. ACM SIGIR*. ACM, 2000, pp. 41–48.

[71] D. Gardner and D. Nichols, “Multi-label classification of satellite images with deep learning,” in *Stanford University Report*, 2017.

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

[73] B. Chen, J. Li, G. Lu, H. Yu, and D. Zhang, “Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification,” *IEEE Journal of Biomedical and Health Informatics*, 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. He is currently with Research Institute of Electronic Engineering Technology, Harbin Institute of Technology, Harbin, China and with Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin (TU Berlin), Berlin, Germany. His research focuses on signal processing and machine learning, and their applications in remote sensing. In particular, he is interested in 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.

Ruben Fernandez-Beltran (SM'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.

Danfeng Hong (S'16–M'19) received the M.Sc. degree (summa cum laude) in computer vision, College of Information Engineering, Qingdao University, Qingdao, China, in 2015, the Dr. -Ing degree (summa cum laude) in Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, in 2019.

Since 2015, he worked as a Research Associate at the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. Currently, he is a research scientist and leads a Spectral Vision working group at IMF, DLR, and also an adjunct scientist in GIPSA-lab, Grenoble INP, CNRS, Univ. Grenoble Alpes, Grenoble, France.

His research interests include signal / image processing and analysis, hyperspectral remote sensing, machine / deep learning, artificial intelligence and their applications in Earth Vision.

Jocelyn Chanussot (M'04–SM'04–F'12) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998. Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning and artificial intelligence. He has been a visiting scholar at Stanford University (USA), KTH (Sweden) and NUS (Singapore). Since 2013, he is an Adjunct Professor of the University of Iceland. In 2015-2017, he was a visiting professor at the University of California, Los Angeles (UCLA). He holds the AXA chair in remote sensing and is an Adjunct professor at the Chinese Academy of Sciences, Aerospace Information research Institute, Beijing.

Dr. Chanussot is the founding President of IEEE Geoscience and Remote Sensing French chapter (2007-2010) which received the 2010 IEEE GRS-S Chapter Excellence Award. He has received multiple outstanding paper awards. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia (2017-2019). He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing (WHISPERS). He was the Chair (2009-2011) and Cochair of the GRS Data Fusion Technical Committee (2005-2008). He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006-2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Transactions on Image Processing and the Proceedings of the IEEE. He was the Editor-in-Chief of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2011-2015). In 2014 he served as a Guest Editor for the IEEE Signal Processing Magazine. He is a Fellow of the IEEE, a member of the Institut Universitaire de France (2012-2017) and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters, 2018-2019).

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 over 200 JCR journal articles (over 160 in IEEE journals), 23 book chapters, and around 300 peer-reviewed conference proceeding papers. His research interests include hyperspectral data processing and parallel computing of remote sensing data.

Dr. Plaza 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 guestedited 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). Additional information: http://www.umbc.edu/rssipl/people/aplaza