

# Graph Relation Network: Modeling Relations Between Scenes for Multilabel Remote-Sensing Image Classification and Retrieval

Jian Kang<sup>1</sup>, Member, IEEE, Ruben Fernandez-Beltran<sup>2</sup>, Senior Member, IEEE,  
Danfeng Hong<sup>3</sup>, Member, IEEE, Jocelyn Chanussot<sup>4</sup>, Fellow, IEEE,  
and Antonio Plaza<sup>5</sup>, Fellow, IEEE

**Abstract**—Due to the proliferation of large-scale remote-sensing (RS) archives with multiple annotations, multilabel 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 multilabel 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 multilabel 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 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 multilabel 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 multilabel scenario. An extensive experimental comparison, conducted on three multilabel RS scene data archives, validates the effectiveness of the proposed GRN in terms of  $KNN$  classification and image retrieval. The codes of this article will be made publicly available for reproducible research in the community.

**Index Terms**—Deep learning, loss function, metric learning, multilabel scene categorization, neighbor embedding, remote sensing (RS).

## 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 with 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 the visual semantics of RS images. In general, the multilabel 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 toward investigating the multilabel scheme [33]–[38]. For example, one of the primal multilabel methods proposed within the RS field was presented

Manuscript received May 2, 2020; revised July 14, 2020 and August 3, 2020; accepted August 9, 2020. This work was supported in part by the Spanish Ministry of Economy under Grant RTI2018-098651-B-C54, in part by the FEDER-Junta de Extremadura under Grant GR18060, in part by the European Union through the H2020 EOXP0SURE Project under Grant 734541, and in part by the AXA Research Fund. (Corresponding author: Danfeng Hong.)

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

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

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

Jocelyn Chanussot is with the Univ. Grenoble Alpes, INRIA, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China (e-mail: jocelyn@hi.is).

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

Color versions of one or more of the figures in this article are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TGRS.2020.3016020

0196-2892 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

in [39] where the authors define a multilabel support vector machine (SVM) for multilabel active learning. To simultaneously exploit the spatial-contextual information and the correlation among the labels, Zeggada *et al.* [40] presented a conditional random field (CRF) framework for multilabel 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 multilabel RS scene categorization due to its excellent feature extraction capabilities. Different works in the RS literature exemplify this fact. For instance, Karalas *et al.* [45] developed a sparse autoencoder framework to extract the underlying semantic features from satellite images, to effectively retrieve multilabel land-cover categories. Zeggada *et al.* [46] proposed a deep learning model for predicting multilabels 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]. It is noted that the availability of sufficient multilabeled 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 multilabel RS training sets. Nonetheless, other authors opt for different alternatives instead. It is the case of Hua *et al.* [49], who proposed an end-to-end network for multilabel 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) subnetwork. Driven by multiattention techniques, Sumbul and Demir [50] also designed a CNN-based deep learning system for RS images with multiple annotations. Alshehri *et al.* [51] presented a multilabel categorization approach based on an encoder–decoder neural network with pretrained CNN features and channel-spatial attention. Additionally, Shao *et al.* [52] proposed a multilabel 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 multiscale features.

Most of the existing deep learning methods in multilabel 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 multilabel 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 multilabel RS image analysis perspective.

In this article, we deal with the multilabel 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 using a new loss function called scalable neighborhood discriminative loss (SNDL). Inspired by the scalable neighborhood component analysis (SNCA) [54], the proposed SNDL provides a novel perspective on the multilabel 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 multilabel 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 multilabel information characterizes the contributions of the nearest neighbors associated with each image on its semantic class decision. To further improve the multilabel 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 multilabel RS scene classification and retrieval. In summary, the main contributions of this article can be highlighted as follows.

- 1) We develop a new GRN for multilabel 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 multilabel RS scene images in an end-to-end deep neural network and adapts the SNCA to the multilabel 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 multilabel RS scene classification and retrieval tasks.
- 4) The proposed GRN also shows a higher effectiveness and robustness when considering different benchmark RS data sets and backbone CNN architectures. The related codes of this article will be made publicly available for reproducible research inside the community<sup>1</sup>.

<sup>1</sup><https://github.com/jian kang1991>

The rest of this article 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 this article with some remarks and hints at plausible future research lines.

## II. METHODOLOGY

Let  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  be a set of  $N$  RS images and  $\mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$  be the associated set of label vectors, where each label vector  $\mathbf{y}_i$  is represented by a multiclass hot encoding vector, that is  $\mathbf{y}_i \in \{-1, 1\}^C$ . 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  $\theta$ , 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, that is  $\|\mathbf{f}_i\|_2 = 1$ . A training set  $\mathcal{T}$  (extracted from  $\mathcal{X}$ ) is built 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 multilabel RS scene image classification and retrieval.

### A. Scalable Neighborhood Component Analysis

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

$$s_{ij} = \mathbf{f}_i^T \mathbf{f}_j \quad (1)$$

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

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

where  $\sigma$  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

$$p_i = \sum_{j \in \Omega_i} p_{ij} \quad (3)$$

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

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

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

Given  $\mathbf{x}_i$ , its similarities with respect to the other images in the data set should be calculated for optimizing (4). Therefore, to stochastically train a CNN model by  $L_{\text{SNCA}}$ , 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.,  $\mathcal{B} = \{\mathbf{f}_1, \dots, \mathbf{f}_M\}$ .  $\mathcal{B}$  is updated in each iteration during the training phase.

The SNCA loss in (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 intraclass variations. This is a highly desired scenario when dealing with the particular semantic complexity of aerial scenes. However, (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 multilabel deep metric learning approach, based on a newly defined GRN-SNDL concept, to effectively learn a metric space for RS images with multilabel information.

### B. Proposed Multilabel Deep Metric Learning Framework for RS Images

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

- 1) 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, that is ResNet18 [61], ResNet50 [61] and WideResNet50 [62].
- 2) A new loss function and its joint version, that is 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.
- 3) The corresponding optimization algorithm, which learns the proposed model parameters using a stochastic process based on an off-line memory bank.

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

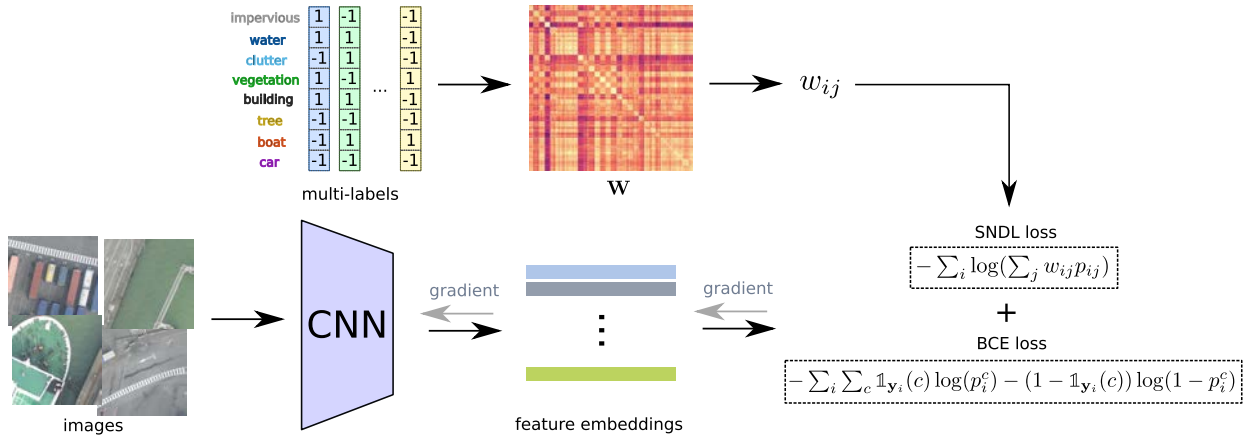


Fig. 1. Proposed framework for multilabel deep metric learning. The SNDL loss is targeted for pulling in the images that share more common labels and pushing away the images with less or no common labels. The BCE loss is integrated for further improving the class discrimination capability.

### 1) Loss Function: Scalable Neighbor Discriminative Loss:

To design our GRN-SNDL under a multilabel assumption, we first rewrite the probability  $p_i$  that  $\mathbf{x}_i$  can be correctly classified within the framework of SNCA [i.e., (3)] as

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

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

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

Given the index set ( $\Omega_i$ ) 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 data set. 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 multilabel annotations, the probability that  $\mathbf{x}_i$  is correctly classified can be determined by

$$p_i = \sum_j w_{ij} p_{ij} \quad (7)$$

where  $w_{ij}$  denotes the contribution weight associated with  $p_{ij}$ . Given an image  $\mathbf{x}_i$  and its multiple labels, we would like to pull in the images that share more common labels with regard to  $\mathbf{x}_i$  in the metric space, and push away the images with less or no common labels with regard 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 multilabel decision for  $\mathbf{x}_i$  through (7). For that purpose, we propose to calculate  $w_{ij}$  based on the multilabel information in the corresponding images as follows:

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

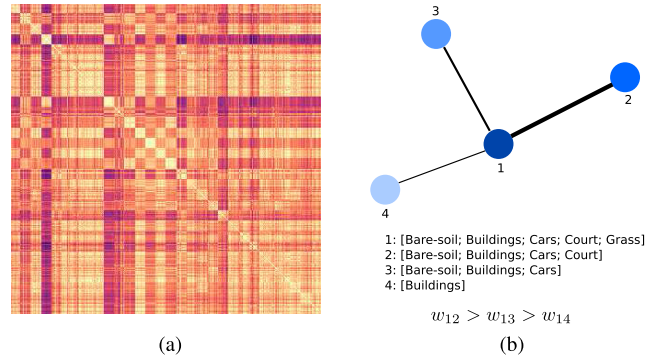


Fig. 2. (a) Weight matrix  $W$  of the aerial image data set (AID). Darker points indicate smaller weights assigned to image pairs (and vice versa). (b) Graph perspective view of the GRN-SNDL loss.

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 (8). As an example, based on the multilabel annotations of the AID data set [19], we utilize (8) to calculate the weight matrix  $W$ , and plot it in Fig. 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:

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

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 multilabels. In the example shown in Fig. 2(b), the connection between nodes 1 and 2 should be stronger than any other node linked with node 1,

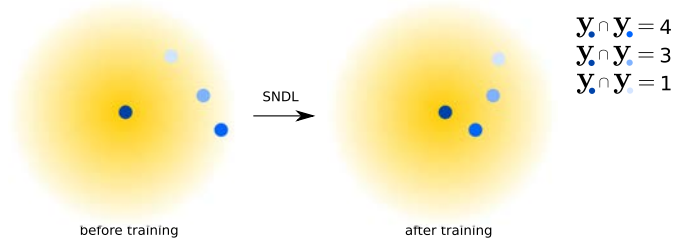


Fig. 3. Illustration of our learning scheme based on GRN-SNDL. Blue points represent features (associated with images) in the metric space. With respect to the center point, the other points have different numbers of identical class labels, and this determines their position in the metric space after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

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 Fig. 3. Blue points represent features (associated with 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 with images with more labels in common have been dragged closer than the points associated with 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:

$$L_{\text{BCE}} = - \sum_i \sum_c \mathbb{1}_{y_i}(c) \log(p_i^c) - (1 - \mathbb{1}_{y_i}(c)) \log(1 - p_i^c) \quad (10)$$

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

$$L = L_{\text{SNDL}} + L_{\text{BCE}}. \quad (11)$$

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:

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

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

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

Then,  $\theta$  can be learned by exploiting the back-propagation algorithm as follows:

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

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

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

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 (15), only the feature embeddings associated with the current mini-batch are updated within the current iteration. The optimization scheme is described in Algorithm 1.

---

#### Algorithm 1 Optimization Scheme for GRN

---

**Require:** Training images  $\mathbf{x}_i$ , the weight matrix  $\mathbf{W}$ , and the multilabel annotations  $\mathbf{y}_i$

- 1: 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  $m$ .
- 2: **for** The epoch number  $t = 0$  to maxEpoch **do**
- 3: Sample a mini-batch.
- 4: Obtain the normalized features  $\mathbf{f}_i^{(t)}$  based on the CNN model with  $\theta^{(t)}$ .
- 5: Calculate the similarities  $s_{ij}$  with reference to  $\mathcal{B}$ .
- 6: Calculate the weights  $w_{ij}$  based on Equation (8).
- 7: Calculate the gradients of SNDL based on Equation (13) (and the ones of BCE).
- 8: Back-propagate the gradients.
- 9: Update the feature embeddings of the current mini-batch stored in  $\mathcal{B}$  via Equation (15).

10: **end for**

**Ensure:**  $\theta$ ,  $\mathcal{B}$

---

### III. EXPERIMENTS

#### A. Data Set Description

In this article, three challenging multilabel RS image data sets are utilized to validate the performance of the proposed method. A detailed description of the considered data sets is provided below.

- 1) *UC Merced (UCM) Multilabel Data Set* [64]: This data set is recreated from the original UCM data set [65] by relabeling all the 2100 aerial images of  $256 \times 256$  pixels with multiple semantic annotations. The original UCM data set 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

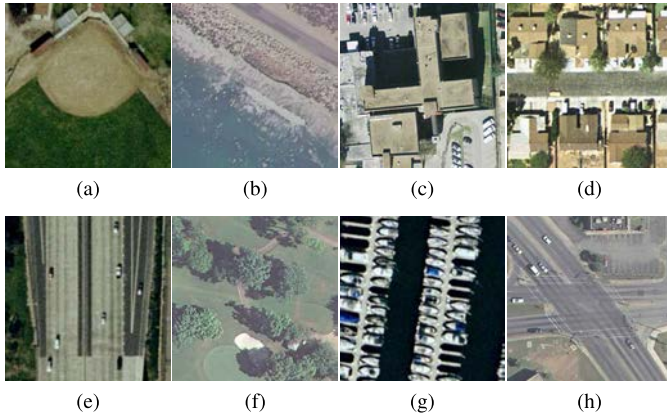


Fig. 4. Examples of the UCM multilabel data set. (a) Bare-soil, buildings, grass. (b) Pavement, sand, sea. (c) Buildings, cars, grass, pavement. (d) Bare-soil, buildings, cars, pavement, trees. (e) Cars, grass, pavement. (f) Bare-soil, grass, pavement, sand, trees. (g) Dock, ship, water. (h) Bare-soil, buildings, cars, grass, pavement, trees.

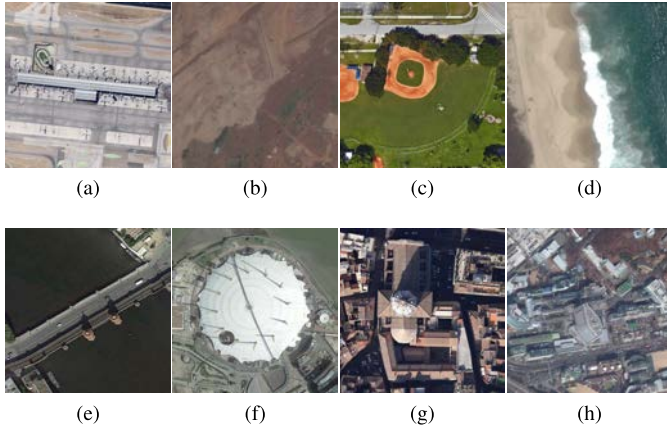


Fig. 5. Examples of the AID multilabel data set. (a) Airplane, bare-soil, buildings, cars, grass, pavement. (b) Bare-soil, buildings, cars, grass, pavement, trees. (c) Bare-soil, buildings, grass, pavement, trees. (d) Chaparral, sand, sea. (e) Buildings, cars, dock, pavement, ship, trees, water. (f) Bare-soil, buildings, car, grass, pavement, trees. (g) Buildings, cars, pavement. (h) Bare-soil, buildings, cars, grass, pavement, trees.

home, ship, bare soil, sea, and field. Fig. 4 illustrates some multilabel examples from this data set.

- 2) *Aerial Image Database (AID) Multilabel Data Set* [66]: This data set is built upon the original AID data set [19], which is specially dedicated to aerial image classification. The original AID data set 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 0.5 to 8 m; 3000 aerial images are selected to construct the AID multilabel data set. The newly defined labels are the same as those in the UCM multilabel data set. Some examples of multilabel annotations are given in Fig. 5.
- 3) *DFC15 Multilabel Data Set* [66]: This data set is created from a semantic segmentation data set called DFC15<sup>2</sup> and acquired over Zeebrugge, Belgium, using an airborne sensor with spatial resolution of 5 cm.

<sup>2</sup>2015 IEEE GRSS data fusion contest. <http://www.grss-ieee.org/community/technical-committees/data-fusion/2015-ieee-grss-data-fusion-contest/>

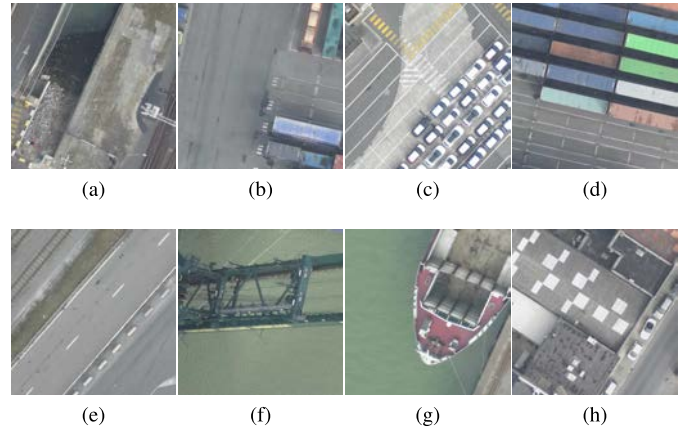


Fig. 6. Examples of the DFC15 multilabel data set. (a) Impervious, water, clutter. (b) Impervious, clutter. (c) Impervious, building, car. (d) Impervious, clutter. (e) Impervious, clutter, vegetation. (f) Water, clutter. (g) Impervious, water, clutter. (h) Impervious, building, car.

The DFC15 multilabel data set consists of 3342 images and there are eight object classes: impervious, water, clutter, vegetation, building, tree, boat, and car. Fig. 6 displays some images with the associated multilabels.

## B. Experimental Setup

The effectiveness of the proposed approach to categorize multilabel 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.

1) *Multilabel 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  $\theta$ . Its predicted label vector  $\mathbf{y}^*$  can be determined by thresholding the mean average of the label vectors of its  $K$ NNs 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

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

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

$$P_s = \frac{TP_s}{TP_s + FP_s}, \quad R_s = \frac{TP_s}{TP_s + FN_s} \quad (17)$$

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

2) *Multilabel RS Image Retrieval*: Image retrieval aims to find the most semantically similar images in the data set, 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 data set. Under a multilabel 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

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

where  $\mathcal{Q}$  denotes the query set,  $R$  represents the number of inspected images from the top-ranking,  $N_{\text{Rel}}(q)@R$  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 (ACGs) [70] score of the first  $r$  retrieved images, which is defined as

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

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

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

where

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

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

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

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 data sets. 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  $256 \times 256$  pixels, and three data augmentation strategies are adopted during training: 1) RandomGrayscale; 2) ColorJitter; and 3) RandomHorizontalFlip. The parameters  $D$ ,  $\sigma$ , and  $m$  are set to 128, 0.1, and 0.5, respectively. The 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 multilabel 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

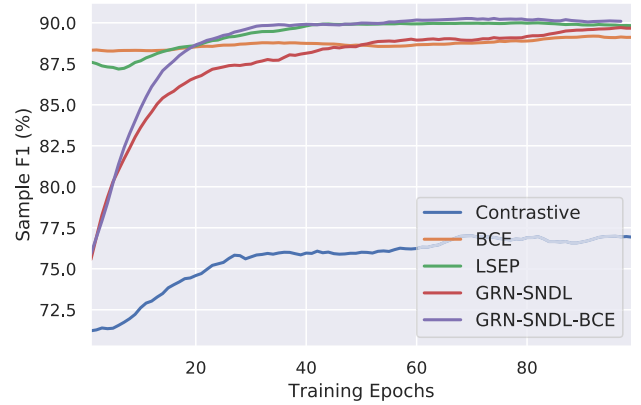


Fig. 7. Learning curves obtained after training ResNet18 with Contrastive, BCE, LSEP, GRN-SNDL, and GRN-SNDL-BCE losses on the AID multilabel data set. We display  $F_s^1$  (%) in the validation set as a function of the number of training epochs.

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. To construct image pairs with multilabel 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 multilabel information of the DFC15 data set is not appropriate to construct pairwise labels for the contrastive loss. Thus, the experiments of the contrastive loss on the DFC15 data set are omitted here. All the experiments have been conducted on an NVIDIA Tesla P100 GPU.

### C. Experimental Results

1) *Multilabel RS Image Classification*: Fig. 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 data set. 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 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 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.

To visualize the learned feature embeddings in the metric space, we exploit  $t$ -distributed stochastic neighbor embedding ( $t$ -SNE) to visualize their projections on a 2-D plane. Fig. 8(a)–(d) shows the  $t$ -SNE scatter plots of the feature embeddings in the UCM training set, obtained using BCE, GRN-SNDL, LSEP, and GRN-SNDL-BCE with WideResNet50, respectively. As we can observe, the proposed method

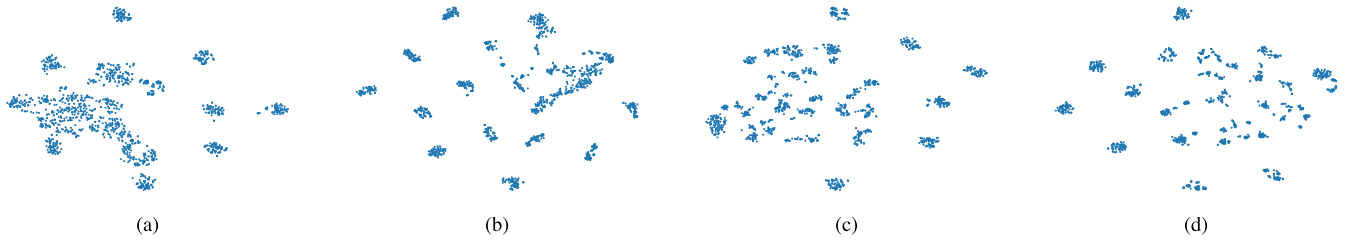


Fig. 8. 2-D projection of the feature embeddings on the UCM training set using  $t$ -SNE. (a) WideResNet50-BCE. (b) WideResNet50-GRN-SNDL. (c) WideResNet50-LSEP. (d) WideResNet50-GRN-SNDL-BCE.

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	<b>89.82</b>	<b>90.26</b>	<b>91.06</b>	<b>91.15</b>	<b>90.26</b>	<b>89.66</b>	<b>93.11</b>	<b>89.63</b>	<b>94.72</b>	<b>94.35</b>	<b>96.09</b>	<b>94.29</b>
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	<b>93.27</b>	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	<b>91.31</b>	<b>91.92</b>	<b>91.98</b>	<b>92.83</b>	<b>90.95</b>	<b>90.82</b>	92.79	<b>91.08</b>	<b>95.80</b>	<b>95.78</b>	<b>96.53</b>	<b>95.95</b>
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	<b>90.81</b>	<b>91.18</b>	<b>91.97</b>	<b>91.92</b>	<b>91.02</b>	<b>90.50</b>	<b>93.66</b>	<b>90.49</b>	<b>95.96</b>	<b>95.73</b>	<b>97.13</b>	<b>95.74</b>

is able to uncover a remarkably finer-grained neighborhood structure by comparing Fig. 8(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 Fig. 8(b) can be separated farther away in Fig. 8(d). Moreover, within some groups, the images are located closer in Fig. 8(d) than Fig. 8(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 data sets. 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 data sets 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 data set. Based on the ResNet50 model, the BCE loss can achieve the








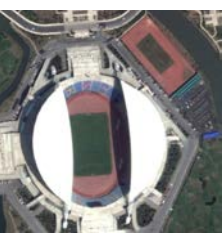


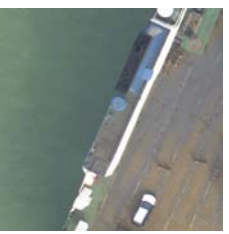

comparable classification performance with respect to the GRN-SNDL-BCE loss with the ResNet18 model.

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 multilabel 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



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				
Ground-truth labels	Impervious, Clutter	Impervious, Vegetation, Building, Tree, Car	Impervious, Water, Car	Impervious, Clutter, Vegetation, Building, Car
Predicted labels	Impervious, Clutter	Impervious, Tree, Vegetation, Building, Car	Impervious, Water	Impervious, Clutter, Vegetation, Building, Car

third image in the UCM data set, grass is a false positive (due to its analogous appearance with regard to court). Similarly, trees is also positively predicted in the third image of the

AID data set, 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 data set, since

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		
		WMAF	MAP(%)	HL	WMAF	MAP(%)	HL	WMAF	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	<b>2.75</b>	<b>99.79</b>	<b>0.09</b>	4.39	99.06	0.11	2.40	100.00	0.11
	GRN-SNDL-BCE	2.71	99.70	0.10	<b>4.47</b>	<b>99.29</b>	<b>0.09</b>	<b>2.51</b>	100.00	<b>0.07</b>
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	<b>99.67</b>	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	<b>2.80</b>	<b>99.92</b>	<b>0.08</b>	<b>4.60</b>	99.66	<b>0.07</b>	<b>2.58</b>	100.00	<b>0.06</b>
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	<b>100.00</b>	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	<b>99.87</b>	0.09	4.40	<b>99.75</b>	0.10	2.47	<b>100.00</b>	0.09
	GRN-SNDL-BCE	<b>2.80</b>	<b>99.87</b>	<b>0.08</b>	<b>4.59</b>	<b>99.75</b>	<b>0.07</b>	<b>2.57</b>	99.99	<b>0.06</b>

its RGB spectral values are close to those of grass in the same image.

2) *Multilabel 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 regard 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 to achieve higher retrieval performances on all the benchmark data sets. To improve multilabel classification accuracy, LSEP is targeted at minimizing the produced label confidence scores in a pairwise manner, where the scores of the true labels should be greater than those of the negative labels. However, the feature embeddings 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 GRN, 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.

Fig. 9 shows the top 5 retrieved images based on ResNet50-LSEP and ResNet50-GRN-SNDL-BCE with respect to the associated query images, where Fig. 9(a), (d), and (g) are the query images from the UCM, AID and DFC15 multilabel data sets, respectively, Fig. 9(b), (e), and (h) are the retrieved images based on ResNet50-LSEP, and Fig. 9(c), (f), and (i)

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

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  $\sigma$  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 data sets) 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 data sets. 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 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  $\sigma$ , with a range from

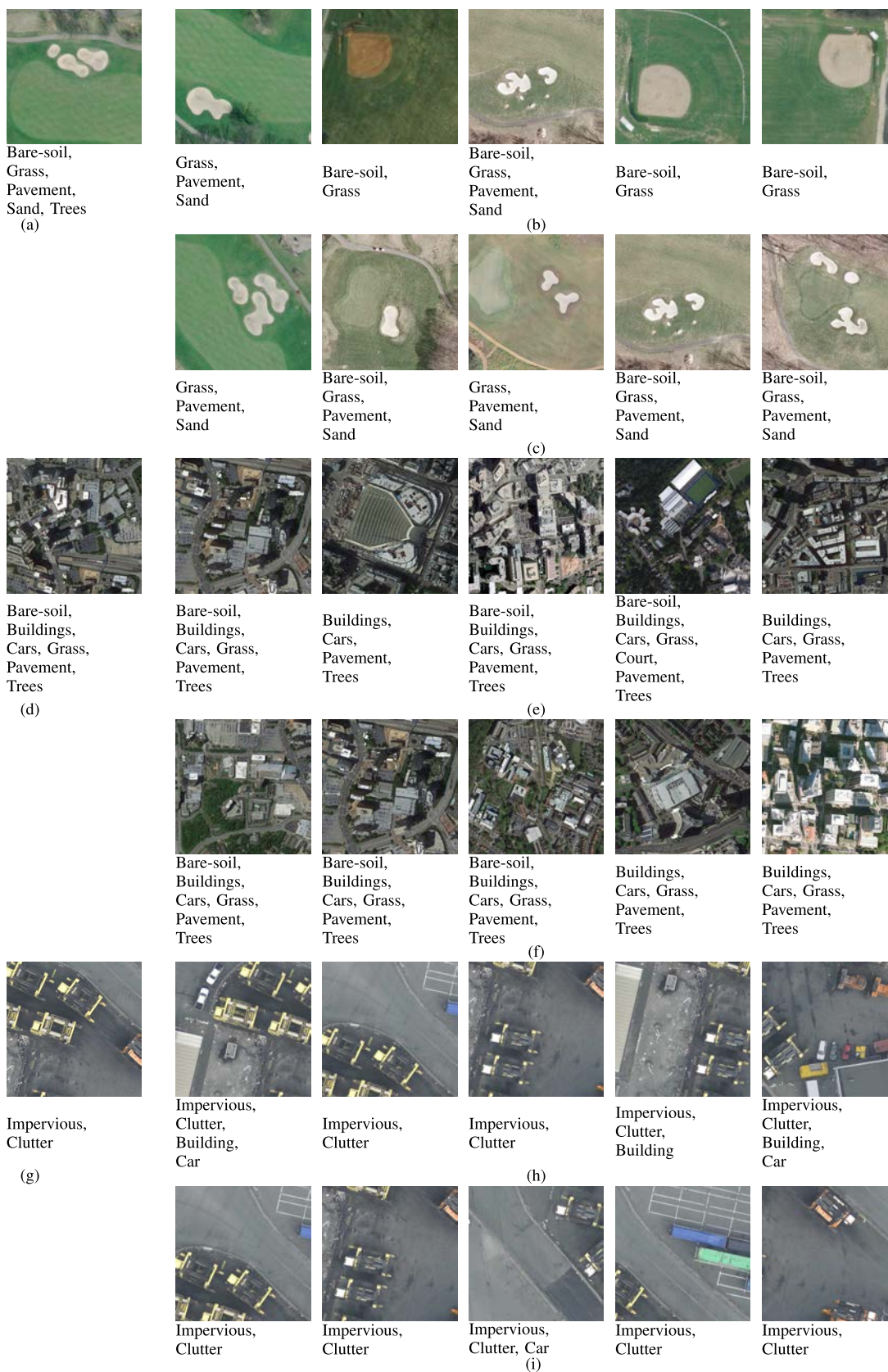


Fig. 9. Image retrieval examples based on ResNet50-LSEP and ResNet50-GRN-SNDL-BCE. (a), (d), and (g) Query images from UCM, AID, and DFC15 data sets, respectively. (b), (e), and (h) Top five nearest neighbors retrieved from the associated training sets, based on ResNet50-LSEP. (c), (f), and (i) Retrieved based on ResNet50-GRN-SNDL-BCE.

TABLE V  
SENSITIVITY ANALYSIS OF PARAMETER  $\sigma$  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

0.05 to 0.2. In this case, we can observe that the classification performances are better when  $\sigma$  equals 0.05 or 0.1. Therefore, we conclude that highly satisfactory results can be reached by the proposed approach function when  $\sigma$  is in the range from 0.05 to 0.1.

#### IV. CONCLUSION AND FUTURE LINES

In this article, we introduce a GRN 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 multilabel perspective. 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 multilabel 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. 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 multilabel benchmark data sets: UCM, AID, and DFC15. Compared with the state-of-the-art losses for multilabel RS scene categorization (including BCE and LSEP), the proposed losses exhibit better classification accuracy, with an improvement of around 2% and 1% with regard 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 data sets 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 multilabel information.

Due to the remarkable potential of the presented method for multilabel RS image classification and retrieval, our future work will be directed toward 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 multilabel annotations.

We are also interested in exploring further developments in terms of efficiency.

#### ACKNOWLEDGMENT

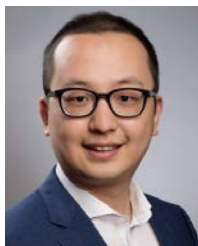
The authors would like to thank the authors for their efforts in creating the multilabel data sets based on UCM, AID and DFC15, and the reviewers for their valuable suggestions.

#### 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, Jun. 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, Jan. 2019.
- [3] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep (overview and toolbox)," *IEEE Geosci. Remote Sens. Mag.*, early access, Apr. 29, 2020, doi: 10.1109/MGRS.2020.2979764.
- [4] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. 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, Dec. 2019.
- [5] X. Xiang Zhu *et al.*, "So2Sat LCZ42: A benchmark dataset for global local climate zones classification," 2019, *arXiv:1912.12171*. [Online]. Available: <http://arxiv.org/abs/1912.12171>
- [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, Jul. 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, Jul. 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, Feb. 2020.
- [9] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 44–59, Nov. 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 Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.
- [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, Dec. 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, Oct. 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.*, early access, Apr. 9, 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.*, early access, Jun. 18, 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, Nov. 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, Apr. 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.*, early access, Jan. 13, 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, Oct. 2017.
- [19] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
- [20] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 193–204, Mar. 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. 14680–14707, Nov. 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, Apr. 2015.
- [23] D. Hong, N. Yokoya, and X. X. Zhu, "Learning a robust local manifold representation for hyperspectral dimensionality reduction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2960–2975, Jun. 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, Aug. 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, Nov. 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, Feb. 2019.
- [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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4982–4993, Dec. 2018.
- [28] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. 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, Sep. 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.*, early access, Feb. 6, 2020, doi: [10.1109/LGRS.2020.2969491](https://doi.org/10.1109/LGRS.2020.2969491).
- [30] H. Yu, L. Gao, W. Liao, B. Zhang, A. Pizurica, 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, Nov. 2017.
- [31] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. 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, Jun. 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, Jan. 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 coreentropy criterion," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1694–1707, Apr. 2017.
- [34] Y. Liu, B. Du, W. Tu, M. Gong, Y. Guo, and D. Tao, "LogDet metric-based domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 10, 2020, doi: [10.1109/TNNLS.2019.2957229](https://doi.org/10.1109/TNNLS.2019.2957229).
- [35] X. Li, B. Du, C. Xu, Y. Zhang, L. Zhang, and D. Tao, "Robust learning with imperfect privileged information," *Artif. Intell.*, vol. 282, May 2020, Art. no. 103246.
- [36] Y. Dong, B. Du, L. Zhang, and L. Zhang, "Dimensionality reduction and classification of hyperspectral images using ensemble discriminative local metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2509–2524, May 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, Jun. 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, Oct. 2018.
- [39] B. T. Zegeye and B. Demir, "A novel active learning technique for multi-label remote sensing image scene classification," *Proc. SPIE*, vol. 10789, Oct. 2018, Art. no. 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, Mar. 2018.
- [41] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.
- [42] X. Deng *et al.*, "Geospatial big data: New paradigm of remote sensing applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 10, pp. 3841–3851, Oct. 2019.
- [43] B. Zhang *et al.*, "Remotely sensed big data: Evolution in model development for information extraction [point of view]," *Proc. IEEE*, vol. 107, no. 12, pp. 2294–2301, Dec. 2019.
- [44] G. Sumbul *et al.*, "BigEarthNet dataset with a new class-nomenclature for remote sensing image understanding," 2020, *arXiv:2001.06372*. [Online]. Available: <http://arxiv.org/abs/2001.06372>
- [45] K. Karalas, G. Tsagkatakis, M. Zervakis, and P. Tsakalides, "Deep learning for multi-label land cover classification," *Proc. SPIE*, vol. 9643, Oct. 2015, Art. no. 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, May 2017.
- [47] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 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.*, vol. 16, no. 7, pp. 1031–1035, Jul. 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 J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.
- [50] G. Sumbul and B. Demir, "A novel multi-attention driven system for multi-label remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 5726–5729.
- [51] A. Alshehri, Y. Bazi, N. Ammour, H. Alzubair, and N. Alajlan, "Deep attention neural network for multi-label classification in unmanned aerial vehicle imagery," *IEEE Access*, vol. 7, pp. 119873–119880, 2019.
- [52] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 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, May 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, Mar. 2015.
- [56] Y. Li, Y. Song, and J. Luo, "Improving pairwise ranking for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 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, Jul. 2019.
- [59] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: <http://arxiv.org/abs/1503.02531>
- [60] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 14, 2020, doi: [10.1109/TGRS.2020.3007029](https://doi.org/10.1109/TGRS.2020.3007029).
- [61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [62] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*. [Online]. Available: <http://arxiv.org/abs/1605.07146>

- [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 Trans. Geosci. Remote Sens.*, pp. 1–14, 2020, doi: [10.1109/TGRS.2020.2991657](https://doi.org/10.1109/TGRS.2020.2991657).
- [64] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multi-label remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.
- [65] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL*, 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*, Jul. 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*, Jun. 2015, pp. 1556–1564.
- [68] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, vol. 463. New York, NY, USA: ACM Press, 1999.
- [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*, vol. 22, no. 2, pp. 540–553, Feb. 2020.
- [70] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proc. ACM SIGIR*, 2000, pp. 41–48.
- [71] D. Gardner and D. Nichols, "Multi-label classification of satellite images with deep learning," Stanford Univ., Stanford, CA, USA, Tech. Rep. 908, 2017. [Online]. Available: <http://cs231n.stanford.edu/reports/2017/pdfs/908.pdf>
- [72] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 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 J. Biomed. Health Informat.*, vol. 24, no. 8, pp. 2292–2302, Aug. 2020.



**Jian Kang** (Member, IEEE) received the B.S. and M.E. degrees in electronic engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2013 and 2015, respectively, and the Dr.Eng. degree from Signal Processing in Earth Observation (SIPEO), Technical University of Munich (TUM), Munich, Germany, in 2019.

In August of 2018, he was a Guest Researcher with the Institute of Computer Graphics and Vision (ICG), Graz University of Technology, Graz, Austria. He is with the Research Institute of Electronic Engineering Technology, Harbin Institute of Technology, and the 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 multidimensional data analysis, geophysical parameter estimation based on InSAR data, SAR denoising, and deep learning-based techniques for remote sensing image analysis.

Dr. Kang received the First Place of the Best Student Paper Award at EUSAR 2018, Aachen, Germany.



**Ruben Fernandez-Beltran** (Senior Member, IEEE) received the B.Sc. degree in computer science, the M.Sc. degree in intelligent systems, and the Ph.D. degree in computer science from Universitat Jaume I, Castellón de la Plana, Spain, in 2007, 2011, and 2016, respectively.

He has been a Visiting Researcher with the University of Bristol, Bristol, U.K., the University of Cáceres, Cáceres, Spain, and the Technische Universität Berlin, Berlin, Germany. He is a Post-Doctoral Researcher with the Computer Vision Group, Universitat Jaume I, where he is also a member of the Institute of New Imaging Technologies. His research interests include multimedia retrieval, spatio-spectral image analysis, pattern recognition techniques applied to image processing, and remote sensing.

Dr. Fernandez-Beltran is also a member of the Spanish Association for Pattern Recognition and Image Analysis (AERFAI), which is part of the International Association for Pattern Recognition (IAPR). He received the Outstanding Ph.D. Dissertation Award at Universitat Jaume I in 2017.



**Danfeng Hong** (Member, IEEE) received the M.Sc. degree (*summa cum laude*) in computer vision from the College of Information Engineering, Qingdao University, Qingdao, China, in 2015, and the Dr.Eng degree (*summa cum laude*) from Signal Processing in Earth Observation (SIPEO), Technical University of Munich (TUM), Munich, Germany, in 2019.

Since 2015, he has been a Research Associate with Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. He is a Research Scientist and leads Spectral Vision Working Group at IMF, DLR, and also an Adjunct Scientist at the GIPSA-lab, CNRS, Grenoble Institute of Technology (Grenoble INP), Université 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** (Fellow, IEEE) 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 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 with Stanford University, Stanford, CA, USA; the KTH Royal Institute of Technology, Stockholm, Sweden; and the National University of Singapore, Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland. From 2015 to 2017, he was a Visiting Professor with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. He holds the AXA Chair in remote sensing and is also an Adjunct Professor with the Chinese Academy of Sciences, Aerospace Information Research Institute, Beijing, China.

Dr. Chanussot was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008 and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He is a member of the Institut Universitaire de France from 2012 to 2017 and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters) from 2018 to 2019. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia, from 2017 to 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 and the Co-Chair of the GRS Data Fusion Technical Committee from 2009 to 2011 and 2005 to 2008, respectively. He is the Founding President of the IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010, which received the 2010 IEEE GRS-S Chapter Excellence Award. He has received multiple outstanding paper awards. He is an Associate Editor of 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 from 2011 to 2015. In 2014, he has served as a Guest Editor for the *IEEE Signal Processing Magazine*.



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

He is 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 the 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 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 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 the President of the Spanish Chapter of the 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 for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING from 2013 to 2017. He has guest-edited ten special issues on hyperspectral remote sensing for different journals. He is also an Associate Editor of IEEE ACCESS (received the recognition as an Outstanding Associate Editor of the journal in 2017). Additional information: <http://www.umbc.edu/rssi/pl/people/aplaza>