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
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 ($K$NN) 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 $K$NN 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 ($K$NN) [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 \( \mathbf{X} = \{x_1, \cdots, x_N\} \) be a set of \( N \) RS images and \( \mathbf{Y} = \{y_1, \cdots, y_N\} \) be the associated set of label vectors, where each label vector \( y_i \) is represented by a multi-class hot encoding vector, i.e., \( y_i \in \{-1, 1\}^C \). Let \( C \) be the total number of RS classes. If an image scene is assigned to the class \( c \), the \( c \)-th element of \( y_i \) 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 \( x_i \) into a corresponding feature embedding \( f_i \in \mathbb{R}^D \) on the unit sphere, i.e., \( \|f_i\|_2 = 1 \). A training set \( \mathcal{T} \) (extracted from \( \mathbf{X} \)) 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 $(x_i, x_j)$ from a training set $\mathcal{T}$ can be measured by the cosine similarity, based on their feature embeddings in the metric space:

$$s_{ij} = f^T_i f_j,$$

(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 $x_i$, the probability $p_{ij}$ that the image $x_j$ is located around its neighborhood in the metric space can be defined as:

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

(2)

where $\sigma$ is a temperature parameter controlling the concentration level of the sample distribution [59], [60]. If $s_{ij}$ is larger, $x_j$ can be chosen as the neighbor of $x_i$ in the metric space at a higher chance than another image $x_k$. $p_{ii} = 0$ 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 $x_i$ can be correctly classified is:

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

(3)

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

$$L_{\text{SNCA}} = -\frac{1}{|\mathcal{T}|} \sum_{i} \log(p_i),$$

(4)

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

Given $x_i$, 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 $L_{\text{SNCA}},$
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., $\mathcal{B} = \{f_i, \cdots, f_M\}$. $\mathcal{B}$ 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 $K$NN 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 $x_i$ 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.

\[ p_i = \sum_j \mathbb{1}_{\Omega_i}(j) p_{ij}, \]  

(5)

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

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

(6)

Given the index set \( (\Omega_i) \) of training images sharing the same class with respect to \( x_i \), the indicator function controls which images can be positioned as neighbors around \( x_i \) 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 \( x_j \) shares the same class with \( x_i \), the associated weight is 1 (and 0 otherwise). In other words, all the contributions on the final class decision of \( x_i \) 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 \( x_i \) is correctly classified can be determined by:

\[ p_i = \sum_j w_{ij} p_{ij}, \]  

(7)

where \( w_{ij} \) denotes the contribution weight associated to \( p_{ij} \). Given an image \( x_i \) and its multiple labels, we would like to pull in the images that share more common labels with regards to \( x_i \).
in the metric space, and push away the images with less or no common labels with regards to \( x_i \). 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 \( x_i \) through Equation (7). For that purpose, we propose to calculate \( w_{ij} \) based on the multi-label information in the corresponding images as follows:

\[
    w_{ij} = \frac{\langle y_i, y_j \rangle + C}{2C}, \quad w_{ij} \in [0, 1]. \tag{8}
\]

Intuitively, \( w_{ij} \) depends on the inner product between \( y_i \) and \( y_j \), which is the cosine between \( y_i \) and \( y_j \). If \( y_i \) is more similar to \( y_j \), there will be a heavier weight assigned to the similarity term \( s_{ij} \) between \( x_i \) and \( x_j \). Since the original range of \( \langle y_i, y_j \rangle \) 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 \( W \), 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 \( T \) with the following formulation, termed as **GRN-SNDL loss**:

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

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
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:

\[
L_{\text{BCE}} = -\sum_i \sum_c \mathbb{1}_{y_i(c)} \log(p_i^c) - (1 - \mathbb{1}_{y_i(c)}) \log(1 - p_i^c),
\]

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:

\[
L = L_{\text{SNDL}} + L_{\text{BCE}}
\]
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 \( f_i \) as indicated in the following equation based on the chain rule:

\[
\frac{\partial L_{\text{SNDL}}}{\partial f_i} = \frac{1}{\sigma} \sum_k p_{ik} f_k - \frac{1}{\sigma} \sum_k w_{ik} \tilde{p}_{ik} f_k,
\]

where \( \tilde{p}_{ik} = \frac{p_{ik}}{\sum_j w_{ij} p_{ij}} \) 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 \( B \) is up-to-date during training, the gradient of the loss function with respect to \( f_i \) at the \( t+1 \)-th iteration is:

\[
\frac{\partial L_{\text{SNDL}}}{\partial f_i} = \frac{1}{\sigma} \sum_k p_{ik} f_k^{(t)} - \frac{1}{\sigma} \sum_k w_{ik} \tilde{p}_{ik} f_k^{(t)}.
\]

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

\[
\frac{\partial L_{\text{SNDL}}}{\partial \theta} = \frac{\partial L_{\text{SNDL}}}{\partial f_i} \times \frac{\partial f_i}{\partial \theta}.
\]

With the feature embeddings \( f_i \) obtained for the current mini-batch and \( B \), we can now update \( f_i \) as:

\[
f_i^{(t+1)} \leftarrow m f_i^{(t)} + (1 - m) f_i,
\]

where \( f_i^{(t)} \) denotes the historical feature embeddings stored in \( B \), and \( m \) is a regularization parameter for updating \( f_i \) 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 \( x_i \), the weight matrix \( W \), and the multi-label annotations \( y_i \)

1: Randomly initialize the parameters \( \theta \) of CNN model, and the memory bank \( B \), as well as the temperature parameter \( \sigma \), the dimensionality \( D \), and the regularization parameter \( m \).

2: for The epoch number \( t = 0 \) to \( \text{maxEpoch} \) do

3: Sample a mini-batch.

4: Obtain the normalized features \( f_i^{(t)} \) based on the CNN model with \( \theta^{(t)} \).

5: Calculate the similarities \( s_{ij} \) with reference to \( B \).

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 \( B \) via Equation (15).

10: end for

Ensure: \( \theta, B \)

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

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 $x^*$, its feature embedding $f^*$ can be obtained by applying $\mathcal{F}(\cdot)$ with the learned parameter set $\theta$. Its predicted label vector $y^*$ can be determined by thresholding the mean average of the label vectors of its $K$ nearest neighbors in $B$ using the value 0.5. We exploit four metrics to evaluate the classification performance, including: 1) sample F1 score ($F_1^s$), 2) sample F2 score ($F_2^s$), 3) sample precision ($P_s$), and 4) sample recall ($R_s$). Specifically, the sample F1 and F2 scores are defined as:

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

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

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

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:

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

where $Q$ 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 $x_q$) 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 $x_q$, (i.e., if there is at least one common class annotated to both images $x_q$ and $x_r$, $\delta(q,r)$ is set to 1 [relevant] and 0 [non-relevant] otherwise) and ACG@$r$ denotes the Average Cumulative Gains (ACG) [70] score of the first $r$ retrieved images, which is defined as:

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

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

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

where

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

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

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

where $\hat{y}$ 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 $256 \times 256$ 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 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 $K$NN 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
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.

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 $K$NN classification results can be obtained when we use the GRN-SNDL-BCE loss (instead of the other losses) for optimization.
TABLE II
SOME $k$NN 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.

<table>
<thead>
<tr>
<th>UCM images</th>
<th>Ground-truth labels</th>
<th>Predicted labels</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="UCM image 1" /></td>
<td>Cars, Pavement</td>
<td>Bare-soil, Buildings, Trees</td>
</tr>
<tr>
<td><img src="image2" alt="UCM image 2" /></td>
<td>Bare-soil, Cars, Court, Pavement, Trees</td>
<td>Bare-soil, Buildings, Cars, Grass, Pavement, Trees</td>
</tr>
<tr>
<td><img src="image3" alt="UCM image 3" /></td>
<td>Bare-soil, Buildings, Trees</td>
<td>Bare-soil, Buildings, Cars, Grass, Pavement, Trees</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AID images</th>
<th>Ground-truth labels</th>
<th>Predicted labels</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image4" alt="AID image 1" /></td>
<td>Bare-soil, Buildings, Cars, Grass, Pavement, Trees</td>
<td>Bare-soil, Buildings, Cars, Grass, Pavement, Trees</td>
</tr>
<tr>
<td><img src="image5" alt="AID image 2" /></td>
<td>Bare-soil, Buildings, Grass, Dock, Grass, Pavement, Sea, Ship</td>
<td>Bare-soil, Buildings, Cars, Grass, Pavement, Trees</td>
</tr>
<tr>
<td><img src="image6" alt="AID image 3" /></td>
<td>Bare-soil, Buildings, Cars, Grass, Pavement, Trees</td>
<td>Bare-soil, Buildings, Cars, Grass, Pavement, Trees</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DFC15 images</th>
<th>Ground-truth labels</th>
<th>Predicted labels</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7" alt="DFC15 image 1" /></td>
<td>Bare-soil, Clutter, Vegetation, Building, Tree, Car</td>
<td>Bare-soil, Clutter, Vegetation, Building, Car</td>
</tr>
<tr>
<td><img src="image8" alt="DFC15 image 2" /></td>
<td>Bare-soil, Water</td>
<td>Bare-soil, Clutter, Vegetation, Building, Car</td>
</tr>
<tr>
<td><img src="image9" alt="DFC15 image 3" /></td>
<td>Bare-soil, Clutter, Vegetation, Building, Car</td>
<td>Bare-soil, Water</td>
</tr>
</tbody>
</table>

January 11, 2021 DRAFT
### 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.

<table>
<thead>
<tr>
<th></th>
<th>UCM</th>
<th>AID</th>
<th>DFC15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WMAP</td>
<td>MAP(%)</td>
<td>HL</td>
</tr>
<tr>
<td>ResNet18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrastive</td>
<td>1.97</td>
<td>86.77</td>
<td>0.19</td>
</tr>
<tr>
<td>BCE</td>
<td>2.52</td>
<td>97.70</td>
<td>0.13</td>
</tr>
<tr>
<td>GRN-SNDL</td>
<td>2.63</td>
<td>99.17</td>
<td>0.11</td>
</tr>
<tr>
<td>LSEP</td>
<td>2.75</td>
<td>99.79</td>
<td>0.09</td>
</tr>
<tr>
<td>GRN-SNDL-BCE</td>
<td>2.71</td>
<td>99.70</td>
<td>0.10</td>
</tr>
<tr>
<td>ResNet50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrastive</td>
<td>2.28</td>
<td>97.02</td>
<td>0.15</td>
</tr>
<tr>
<td>BCE</td>
<td>2.64</td>
<td>98.99</td>
<td>0.11</td>
</tr>
<tr>
<td>GRN-SNDL</td>
<td>2.71</td>
<td>99.64</td>
<td>0.10</td>
</tr>
<tr>
<td>LSEP</td>
<td>2.77</td>
<td>99.81</td>
<td>0.09</td>
</tr>
<tr>
<td>GRN-SNDL-BCE</td>
<td><strong>2.80</strong></td>
<td><strong>99.92</strong></td>
<td><strong>0.08</strong></td>
</tr>
<tr>
<td>WideResNet50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrastive</td>
<td>2.22</td>
<td>96.47</td>
<td>0.15</td>
</tr>
<tr>
<td>BCE</td>
<td>2.62</td>
<td>99.37</td>
<td>0.11</td>
</tr>
<tr>
<td>GRN-SNDL</td>
<td>2.73</td>
<td>99.44</td>
<td>0.10</td>
</tr>
<tr>
<td>LSEP</td>
<td>2.76</td>
<td><strong>99.87</strong></td>
<td>0.09</td>
</tr>
<tr>
<td>GRN-SNDL-BCE</td>
<td><strong>2.80</strong></td>
<td><strong>99.87</strong></td>
<td><strong>0.08</strong></td>
</tr>
</tbody>
</table>

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 $K$NN 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 $K$NN 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 $K$NN 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 embedddings 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 $\sigma$ are the two main parameters of the proposed framework. With ResNet18, in Table IV we calculate the $F_1^s$ (%) of the $K$NN 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
Fig. 9. Image retrieval examples based on ResNet50-LSEP and ResNet50-GRN-SNDL-BCE. (a), (d) and (g) are the query images from UCM, AID and DFC15 datasets, respectively. (b), (e) and (h) are the top 5 nearest neighbors retrieved from the associated training sets, based on ResNet50-LSEP. (c), (f) and (i) are retrieved based on ResNet50-GRN-SNDL-BCE.
TABLE IV
Sensitivity analysis of parameter $D$ in the proposed model (GRN-SNDL) based on the $F_1^s$ (%) of the KNN classification.

<table>
<thead>
<tr>
<th></th>
<th>UCM</th>
<th>AID</th>
<th>DFC15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D = 32$</td>
<td>87.68</td>
<td>89.33</td>
<td>92.72</td>
</tr>
<tr>
<td>$D = 64$</td>
<td>87.52</td>
<td>88.56</td>
<td>92.98</td>
</tr>
<tr>
<td>$D = 128$</td>
<td>88.47</td>
<td>89.13</td>
<td>93.08</td>
</tr>
</tbody>
</table>

TABLE V
Sensitivity analysis of parameter $\sigma$ in the proposed model (GRN-SNDL) based on the $F_1^s$ (%) of the KNN classification.

<table>
<thead>
<tr>
<th></th>
<th>UCM</th>
<th>AID</th>
<th>DFC15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 0.05$</td>
<td>87.83</td>
<td>89.60</td>
<td>93.97</td>
</tr>
<tr>
<td>$\sigma = 0.1$</td>
<td>88.47</td>
<td>89.13</td>
<td>93.08</td>
</tr>
<tr>
<td>$\sigma = 0.15$</td>
<td>87.04</td>
<td>86.90</td>
<td>91.60</td>
</tr>
<tr>
<td>$\sigma = 0.2$</td>
<td>85.71</td>
<td>85.76</td>
<td>92.34</td>
</tr>
</tbody>
</table>

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 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. 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


**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.

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