Deep Hashing Based on Class-Discriminated Neighborhood Embedding

Jian Kang, *Member, IEEE*, Ruben Fernandez-Beltran, *Senior Member, IEEE*, Zhen Ye, Xiaohua Tong, *Senior Member, IEEE*, and Antonio Plaza, *Fellow, IEEE*

Abstract—**Deep-hashing methods have drawn significant attention during the past years in the field of remote sensing (RS) owing to their prominent capabilities for capturing the semantics from complex RS scenes and generating the associated hash codes in an end-to-end manner. Most existing deep-hashing methods exploit pairwise and triplet losses to learn the hash codes with the preservation of semantic-similarities which require the construction of image pairs and triplets based on supervised information (e.g., class labels). However, the learned Hamming spaces based on these losses may not be optimal due to an insufficient sampling of image pairs and triplets for scalable RS archives. To solve this limitation, we propose a new deep-hashing technique based on the class-discriminated neighborhood embedding, which can properly capture the locality structures among the RS scenes and distinguish images class-wisely in the Hamming space. An extensive experimentation has been conducted in order to validate the effectiveness of the proposed method by comparing it with several state-of-the-art conventional and deep-hashing methods. The related codes of this article will be made publicly available for reproducible research by the community.**

Index Terms—**Content-based image retrieval, deep hashing, deep learning, fast similarity search, hashing, remote sensing.**

I. Introduction

SPACEBORNE and airborne remotely sensed images offer an important tool to deal with current societal needs as well as future challenges [1]. From the study of the spectral properties of the Earth surface [2]–[4], through the visual detection of specific targets [5]–[7], to planning and monitoring of land-cover [8]–[10], there are multiple domains where remote sensing (RS) images become particularly useful, and the growing development of different Earth observation (EO) missions exemplifies this fact [11]. As a result, recent years have witnessed an explosive growth in RS image collections, aimed at implementing big-scale operational services which demand new efficient methodologies to manage and retrieve relevant information from the massive resulting RS archives [10], [12]–[14]. Logically, content-based image retrieval (CBIR) technology plays a fundamental role in this regard.

In general, CBIR is concerned about providing users with those images which satisfy their queries according to their visual content [15]. Therefore, three main components are typically involved in the retrieval process: a query, characterized by one or more visual examples of the concept of interest; an image archive, which is used to extract images related to the query concept; and a ranking function, which aims at sorting the archive according to the relevance to the query. From a computational perspective, the ranking function is one of the most important parts of the retrieval system, since it needs to process the whole image archive to discover the most relevant samples. Whereas the traditional k-nearest neighbors approach has been shown to be among the most popular and effective methods in standard content-based retrieval [16], this solution often may not apply in actual operational RS environments due to the high computational burden of the exhaustive search process over big-scale image archives [17], [18].

Recently, hashing techniques have proven their potential in alleviating these limitations within the RS field [19]–[21]. Hashing-based retrieval methods [22] aim at projecting the original data into a set of compact binary codes (hash codes) in order to conduct the CBIR process in a very efficient way by taking advantage of the fast computation of the Hamming distance. That is, the input query (together with the image archive) are mapped into a low-dimensional binary space and then the ranking process can be efficiently computed using simple bit-wise exclusive-OR operations. Note that this approach avoids exhaustively comparing the query image with each sample in the archive by making use of an approximate nearest neighbors ranking, which is generally able to obtain good performance in practical RS applications [23]. Besides, it also provides important memory savings, due to the inherent compression of the binary representation. This represents a huge advantage in large-scale RS scenarios [24].

In the literature, it is possible to find several types of hashing methods which exploit different machine learning

Manuscript received May 11, 2020; revised July 12, 2020 and August 20, 2020; accepted September 27, 2020. Date of publication September 30, 2020; date of current version October 14, 2020. This work was supported in part by the National Key Research, and Development Project of China under Grant 2018YFB0505400 and Grant 2017YFB0502700, in part by the National Natural Science Foundation of China under Grant 41631178, in part by the Ministry of Science, Innovation, and Universities of Spain (RTI2018-098651-B-C54), in part by the Valencian Government of Spain (GV/2020/167), FEDER-Junta de Extremadura (Ref. GR18060), and in part by the European Union under the H2020 EOXPOSURE Project (734541). *(Corresponding author: Zhen Ye.)*

Jian Kang is with the School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China (e-mail: kangjian_1991@outlook.com).

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).

Zhen Ye and Xiaohua Tong are with the College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China (e-mail: yezhen0402@126.com; xhtong@tongji.edu.cn).

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).

Digital Object Identifier 10.1109/JSTARS.2020.3027954

paradigms [22]. Whereas some conventional algorithms, such as the locality-sensitive hashing (LSH) [25], the iterative quantization (ITQ), or the density sensitive hashing (DSH) [26], are able to produce positive results with standard imagery [27], the special complexity of the RS image domain often makes that specialized hashing techniques are required to effectively retrieve RS optical data [20], [21]. Specifically, deep-hashing methods have recently shown prominent capabilities in RS due to the great potential of convolutional neural networks (CNN) to extract highly relevant features from aerial scenes [28]. Several research works in the RS literature exemplify this trend [29]–[32].

Despite the success achieved by these and other relevant methods, most of the existing deep-hashing approaches rely on pairwise [32] or triplet loss functions [31], [33] to learn the corresponding hash codes, while preserving the semantic relationships in the resulting Hamming space. However, the process of effectively sampling image pairs and triplets when training these methods may become a critical aspect in RS. Note that the constant development of the acquisition technology (together with the unprecedented availability of airborne and spaceborne optical data) are substantially increasing the semantic complexity and volume of RS archives. As a result, RS CBIR systems are expected to deal with a growing land-cover within-class diversity and between-class similarity, which may be particularly important in large-scale datasets [17], [34]. Nonetheless, current deep-hashing models are typically optimized by stochastically sampling image pairs or triplets within each mini-batch, which may eventually constrain the number of positive and negative land-cover sample concepts that can be considered in each training iteration. Consequently, this limited semantic scope may lead to an insufficient (or unaffordable) learning process, thus motivating the development of new deep-hashing models to effectively learn binary codes from unconstrained RS archives [12].

In order to relieve these limitations, this article proposes a new deep-hashing method for RS scene retrieval, which offers a novel formulation with important advantages over existing deep-learning hashing schemes [31], [32]. Specifically, we define the class-discriminated neighborhood embedding (CDNE), which pursues to enhance the land-cover semantic information of the binary representations by sufficiently capturing the locality structures among RS scenes and class wisely distinguishing images in the Hamming space. To achieve this goal, three main components take part in the presented design: first, the scalable neighborhood component analysis (SNCA), focused on discovering the neighborhood structure in the metric space; second, the cross entropy (CE) loss, aimed at preserving the land-cover class discrimination capability; and 3) the quantization loss, directed to generate the final binary codes. Additionally, we also define two optimization procedures (based on the memory bank and momentum update) in order to train the proposed deep-hashing approach. A comprehensive experimental comparison, including two benchmark RS image archives and multiple state-of-the-art hashing methods, is conducted to illustrate the advantages of our contribution when generating hash codes for efficient retrieval of RS scenes. The main contributions of this article can be summarized as follows.

1) We propose a new deep-hashing metric learning model specifically designed to deal with the high data volume and semantic complexity of RS images in retrieval tasks. The presented approach is able to learn a metric space based on CNN models that preserve the discrimination capability of land-cover concepts in the resulting Hamming space.
2) We define two optimization mechanisms (memory bank and momentum update) for training the proposed approach while preserving the consistency of the feature embeddings generated on the whole RS archive.
3) We demonstrate the superiority of our model (in the task of retrieving RS scenes) with respect to multiple state-of-the-art hashing methods, over different benchmark datasets. The related codes will be released for reproducible research inside the RS community.[1]

The rest of this article is organized as follows. Section III presents the proposed deep-hashing metric learning model for RS CBIR. Section II reviews some related works while highlighting their main limitations. Section IV contains the experimental part of the work, conducted on different publicly available benchmark datasets. Section V concludes this article and provides some hints at plausible future research lines.

II. Related Work

A. *Conventional Hashing*

Among conventional hashing methods, one of the most popular techniques for CBIR is the LSH [25]. In particular, this approach is based on using several hashing functions (computed by random projections) in order to guarantee a high probability of collision in the Hamming space for similar input samples. Similarly, the kernelized LSH [35] takes advantage of the so-called kernel embeddings to compute such random projections over the embedding space, using a reduced number of input samples. Despite the effectiveness of LSH-based techniques, alternative hashing methods have been also developed in the literature. For instance, it is the case of the anchor graphs hashing (AGH) [36]. Specifically, AGH builds an approximate neighborhood graph (anchor graph) where the underlying manifold structure of the input data is preserved. Then, it hierarchically thresholds the lower eigenfunctions of the anchor graph Laplacian to generate the hashing functions. In [37], Kong *et al.* define the isotropic hashing, which simultaneously combines different projection schemes to produce the final binary characterizations. Another relevant approach is the compressed hashing [38] that makes use of the sparse coding formulation as hashing framework. Other works also propose alternative hashing frameworks that provide even superior results. For example, Gong *et al.* present [39] the ITQ, which is one of the most successful conventional hashing methods. In more detail, ITQ formulates the hashing problem in terms of minimizing the quantization error after mapping the input data to the vertices of a zero-centered binary hypercube. Another successful method is the DSH [26], which follows the rationale of LSH but using improved binary projections that better encapsulate the original data distribution. Additionally,

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

Xia *et al.* propose [40] another significant hashing method which introduces a sparsity-inducing regularizer to reduce the number of parameters when learning the corresponding projection operator.

B. Deep Hashing

Despite their effectiveness, conventional hashing methods tend to require rather lengthy binary codes to achieve accurate CBIR results when dealing with complex multi-dimensional optical data, which generally makes other alternatives more convenient to effectively retrieve RS data [20], [21], [41], [42]. Among all the conducted research, deep-hashing models have recently exhibited great potential in RS due to the prominent success of CNNs to uncover highly discriminating features from aerial scenes [28], [43]. As a result, several deep-hashing methods have been developed in the most recent RS literature. For instance, this is the case of the work by Li *et al.*, who define [29] the deep hashing neural network (DHNN). In details, DHNN uses a deep feature learning network to first extract features from RS scenes and then applies a hashing learning network to compact such feature representations as binary codes. To achieve this goal, the authors employ a binary quantization loss to encode each output feature as a binary value, and a pairwise similarity constraint to enforce similarities between image pairs according to the corresponding semantic annotations. In [30], Song *et al.* improve this framework by presenting the deep hashing CNN (DHCNN) which is able to achieve even better results. Specifically, the DHCNN makes use of a pretrained CNN to initially extract deep features from RS images. Then, a hash layer (with metric learning regularization) and a fully connected layer (with a softmax classifier) are used to produce the hash codes together with the class distribution for the retrieval process. Additionally, Roy *et al.* develop [31] the metric-learning hashing network (MiLaN), which takes advantage of pretrained CNN features to build a metric space using a joint loss function based on the triplet loss formulation [44]. Different from previous pairwise-based deep-hashing methods, this approach considers both positive and negative RS scenes along the optimization process for learning the output binary codes. In the case of [32], Li *et al.* propose an alternative deep-hashing framework for RS that uses quantized convolutional layers. More specifically, the quantized deep learning to hash is based on a deep feature extraction network with binary filter weights and 2-bit activations, together with a hash code learning network that generates the binary hash codes using a weighted pairwise loss function.

C. Current Limitations in RS Applications

From low-level image features to high-level land-cover semantic concepts, hashing techniques are required to face a particularly important semantic gap in RS CBIR to satisfy user queries. Whereas many of the existing deep-hashing methods try to relieve this gap by using pairwise [32] or triplet loss functions [31], the process of sampling image pairs and triplets when training these models is still a critical point to effectively preserve the semantic relationships among land-cover concepts in the resulting Hamming space. Note that the increasing availability of RS data, together with the constant development of the acquisition technology, are generating an unprecedented complexity in the task of encapsulating the Earth surface visual semantics into only a few bits of information [45]. That is, ongoing RS CBIR systems need to cope with an increasing within-class diversity and between-class similarity of land-cover semantic concepts, that may compromise the efficacy of current deep-hashing techniques under large-scale RS scenarios [17]. State-of-the-art deep-hashing models, such as [31] and [32], are typically optimized by stochastically sampling image pairs (or triplets) within each mini-batch, which may limit the number of positive and negative land-cover concepts that can be considered each training iteration. Precisely, this constraint may still leads to insufficient discriminability in the resulting Hamming space, due to the increasing the semantic complexity and volume of RS archives [12].

D. Novelty of the Proposed Approach

In order to address these challenges, this article proposes a new deep-hashing method for RS scene retrieval that jointly exploits three different components: the SNCA, CA, and quantization terms. Unlike other works available in the literature [31], [32], [46], the proposed approach integrates these three components for simultaneously covering three key factors of the RS-based hashing problem: data volume, semantic complexity, and binarization loss. With the increasing expansion of remotely sensed big data [47], hashing systems are expected to deal with vast RS archives. In this situation, the SNCA component is focused on efficiently uncovering the neighborhood structure in the metric space from scalable RS data. Moreover, the ongoing evolution of the acquisition technology, together with other inherent factors (e.g., instrument positions, lighting conditions, sensor types, image corrections, etc.) also make the RS data more complex and difficult to understand [48]. Precisely, the proposed approach employs the CA term to further alleviate the large-scale variance problem in RS. Finally, the quantization term pursues to adequately binarize the corresponding embedding into a Hamming space to allow efficient retrieval of RS images by content. In addition to these improvements, the proposed approach has been defined using two different optimization algorithms (based on the memory bank and momentum contrast [49]) to further improve the intraclass discrimination capability when generating the hash codes from RS data. Compared with different state-of-the-art methods, the proposed approach is able to achieve a better performance than the methods in [29]–[31], and [50], which also indicates the novelty and advantages provided by this article within the RS community.

III. CDNE for Deep Hashing

The proposed end-to-end deep-hashing method for RS image retrieval (CDNE) is made up of three main parts. First, we consider a backbone CNN architecture to generate the corresponding feature embedding space for the input RS images. Second, we define a new loss function which consists of three joint components: a class discrimination, a neighborhood embedding, and a quantization term. Third, we define an optimization mechanism based on the momentum update to train the CDNE model. Fig. 1 summarizes the proposed framework in a graphical way.

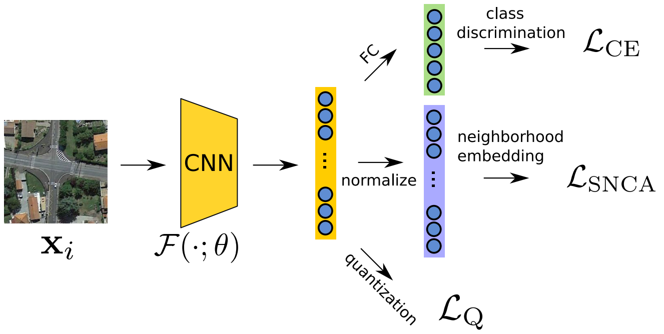

$\mathbf{x}_i$
CNN
$\mathcal{F}(\cdot;\theta)$
FC
class discrimination
$\mathcal{L}_{\mathrm{CE}}$
normalize
neighborhood embedding
$\mathcal{L}_{\mathrm{SNCA}}$
quantization
$\mathcal{L}_{\mathrm{Q}}$

Fig. 1. Proposed end-to-end deep-hashing framework for RS image retrieval (CDNE).

The details of our approach will be provided in the following sections. Nonetheless, let us start by defining the notation used in this article.

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ be an RS image dataset consisting of M images, and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ is the associated set of label vectors, where each label vector $\mathbf{y}_i$ is represented by the one-hot vector, i.e., $\mathbf{y}_i \in \{0,1\}^C$, where C is the total number of classes. If the image is annotated by class c, the cth element of $\mathbf{y}_i$ is 1. We aim to learn a hashing function $\mathcal{F}(\cdot;\theta)$, represented by a CNN model, which can map the image $\mathbf{x}_i$ into a sequence of binary values, denoted as $\mathbf{b}_i$, with the length of L (i.e., $\mathbf{b}_i \in \{0,1\}^L$). The set θ represents the learnable parameters of the CNN model. Specifically, we note $\mathbf{h}_i$ as the output feature via $\mathcal{F}(\cdot;\theta)$ associated to $\mathbf{x}_i$, (i.e., $\mathbf{h}_i = \mathcal{F}(\mathbf{x}_i;\theta)$), and $\mathbf{f}_i$ is the normalized feature on the unit sphere (i.e., $\mathbf{f}_i = \mathbf{h}_i/\|\mathbf{h}_i\|_2$). The main aim of a deep hashing system is that the distances of the generated hash codes measured in the Hamming space can reveal the semantic-similarities among the associated images. To preserve such characteristics, we utilize supervised information (i.e., class labels) to guide the training of the system. The images associated with the same semantic category are located closer in the Hamming space than the others (belonging to a different category). To reach this goal, we propose a new deep hashing method based on CDNE. In the following sections, we will describe the proposed joint loss function (see Section III-A) and the associated optimization algorithm (see Section III-B).

A. Loss Function

In order to encode the RS images sharing the same class label located nearby in the Hamming space, we utilize neighborhood component analysis [51] for the neighborhood embedding. Given a pair of images $(\mathbf{x}_i, \mathbf{x}_j)$, their cosine similarity s_{ij} can be measured by the inner product of the normalized features obtained from a CNN model, i.e., $s_{ij} = \mathbf{f}_i^T \mathbf{f}_j$. Then, the probability (p_{ij}) of $\mathbf{x}_i$ selecting $\mathbf{x}_j$ as its neighbor in the feature space, is defined as

$$p_{ij} = \frac{\exp(s_{ij}/\lambda)}{\sum_{k\neq i}\exp(s_{ik}/\lambda)} \tag{1}$$

where λ denotes the temperature parameter controlling the concentration level of the sample distribution [52]. The higher the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, the higher the chance that $\mathbf{x}_j$ can be selected as a neighbor of $\mathbf{x}_i$ in feature space. Such probability is often termed as *leave-one-out* distribution on $\mathcal{X}$. Thus, $\mathbf{x}_i$ can be correctly classified with the probability p_i as

$$p_i = \sum_{j\in\Omega_i} p_{ij} \tag{2}$$

where $\Omega_i = \{\mathbf{y}_i = \mathbf{y}_j\}$ denotes the indices of the images sharing the same class with respect to $\mathbf{x}_i$. For scalable datasets, such neighborhood embedding strategy can be modeled by exploiting the SNCA [53] with the following loss function:

$$\mathcal{L}_{\mathrm{SNCA}} = -\frac{1}{|\mathcal{X}|}\sum_i \log(p_i). \tag{3}$$

As investigated in our previous work [46], the stochastic optimization of 3 can lead to the discovery of the inherent locality structure among images in the feature space, especially when there exists high intraclass variations within the dataset. However, the class discrimination capability may not be well characterized by just utilizing SNCA. To overcome this limitation, we introduce CE loss, where class-wise prototypes can be learned and the associated image features are optimized to be aligned with respect to them. The CE loss is described by

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\mathcal{X}|}\sum_i\sum_c y_i^c \log(p_i^c) \tag{4}$$

where p_i^c denotes the probability that $\mathbf{x}_i$ is classified into the class c, formulated as

$$p_i^c = \frac{\exp(\mathbf{w}_c^T\mathbf{h}_i)}{\sum_j \exp(\mathbf{w}_j^T\mathbf{h}_i)} \tag{5}$$

where $\mathbf{w}_c$ are the learned parameters from class c. In order to make $\mathbf{h}_i$ approximate the binary values, the quantization loss is also involved to train the CNN model with the following form:

$$\mathcal{L}_Q = \sum_i \|\mathbf{h}_i - \mathbf{b}_i\|_2^2, \quad \mathbf{b}_i = \mathrm{sign}(\mathbf{h}_i) \tag{6}$$

where $\mathrm{sign}(\cdot)$ represents the signum function. Finally, the proposed joint loss function for training the hashing system is described by

$$\mathcal{L}_{\mathrm{CDNE}} = \mathcal{L}_{\mathrm{SNCA}} + \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_Q. \tag{7}$$

Note that, unlike other deep metric learning applications, the loss formulation considered in this article does not include any penalty hyperparameter for the sake of simplicity and the binary nature of final Hamming space. The general framework of the proposed hashing system, including the defined loss composition, is illustrated in Fig. 1.

B. Optimization Strategy

As introduced in [46], [53], we can obtain the gradient of $\mathcal{L}_{\mathrm{CDNE}}$ with respect to $\mathbf{f}_i$ based on the chain rule

$$\begin{aligned}\frac{\partial\mathcal{L}_{\mathrm{CDNE}}}{\partial\mathbf{f}_i} = &-y_i^c(1-p_i^c)\|\mathbf{h}_i\|_2\mathbf{w}_c + \frac{\lambda}{\sigma}\sum_k p_{ik}\mathbf{f}_k\\ &-\frac{\lambda}{\sigma}\sum_{k\in\Omega_i}\tilde{p}_{ik}\mathbf{f}_k.\end{aligned} \tag{8}$$

Algorithm 1: CDNE(MB).

```
Require: $\mathbf{x}_i$, and $\mathbf{y}_i$
Initialize $\theta$ and $\mathcal{B}$ (randomly), along with $\sigma$, $L$ and $m$.
for $t = 0$ to $\mathrm{maxEpoch}$ do
    Sample a mini-batch.
    Obtain $\mathbf{f}_i^{(t)}$ and $\mathbf{h}_i^{(t)}$ based on CNN with $\theta^{(t)}$.
    Calculate $s_{ij}$ with reference to $\mathcal{B}$.
    Calculate the gradients based on (8).
    Back-propagate the gradients.
    Update $\mathcal{B}$ via (9).
endfor
Ensure: $\theta$, $\mathcal{B}$
```

It can be seen that the feature embeddings of the whole dataset are required for calculating the gradients. Thus, a memory bank $\mathcal{B}$ is exploited to store the normalized features, i.e., $\mathcal{B} = \{\mathbf{f}_1, \ldots, \mathbf{f}_M\}$. After each iteration, θ can be updated by using back-propagation, and $\mathcal{B}$ can be updated by

$$\mathbf{f}_i^{(t+1)} \leftarrow m\mathbf{f}_i^{(t)} + (1-m)\mathbf{f}_i \tag{9}$$

where m is the parameter used for the proximal regularization of $\mathbf{f}_i$ based on its previous state. Since the normalized features are progressively updated in a memory bank, we term this optimization strategy as *CDNE(MB)*. The associated optimization scheme is described in Algorithm 1.

Alternatively, another optimization mechanism based on momentum update [49], [54] is proposed in our previous work [46]. In order to consistently generate the features for optimizing the CNN models, we progressively update the state of CNN models instead of updating the normalized features in the memory bank. To achieve this, an auxiliary CNN model with parameters θ_{aux} is introduced, and θ_{aux} is updated through

$$\theta_{\mathrm{aux}}^{(t+1)} \leftarrow m\theta_{\mathrm{aux}}^{(t)} + (1-m)\theta^{(t)} \tag{10}$$

where $m \in [0, 1)$ is the momentum coefficient, and the state of θ is updated using back-propagation. Based on the expected mean average of θ_{aux}, the state of the auxiliary CNN model can be more smoothly evolved than the CNN model with θ. Thus, the features in $\mathcal{B}$ are encoded by $\mathcal{F}(\cdot; \theta_{\mathrm{aux}})$, and updated through

$$\hat{\mathbf{f}}_i^{(t+1)} \leftarrow \hat{\mathbf{f}}_i^{(t)}. \tag{11}$$

We term this strategy as *CDNE(MU)*. The associated optimization scheme is described in Algorithm 2. After the training based on these two optimization strategies, the hash code for a new image $\mathbf{x}^*$ outside the archive (i.e., $\mathbf{x}^* \notin \mathcal{X}$) can be generated by

$$\mathbf{b}^* = \mathrm{sign}(\mathcal{F}(\mathbf{x}^*; \theta)). \tag{12}$$

IV. Experiments

A. Datasets

In this article, two benchmark RS image archives are utilized to evaluate the performance of the proposed method. A detailed description of the considered datasets is provided below.

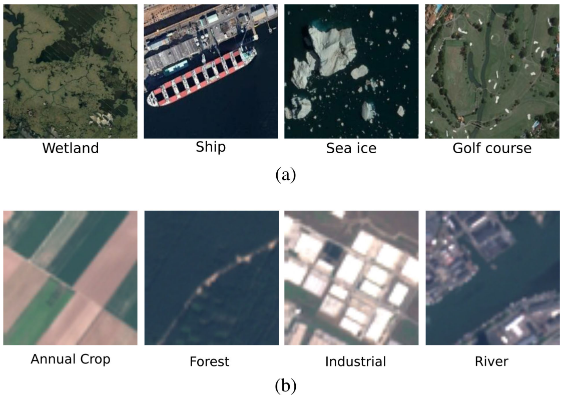


Fig. 2. Some examples of the NWPU-RESISC45 (a) and EuroSAT (b) image collections used in our experiments.

Algorithm 2: CDNE(MU).

```
Require: $\mathbf{x}_i$, and $\mathbf{y}_i$
Initialize $\theta$, $\theta_{\mathrm{aux}}$ and $\mathcal{B}$ (randomly), along with $\sigma$, $L$
  and $m$.
for $t = 0$ to $\mathrm{maxEpoch}$ do
    Sample a mini-batch.
    Obtain $\mathbf{f}_i^{(t)}$ and $\mathbf{h}_i^{(t)}$ based on CNN with $\theta^{(t)}$.
    Obtain $\hat{\mathbf{f}}_i^{(t)}$ and $\hat{\mathbf{h}}_i^{(t)}$ based on the auxiliary CNN
      with $\theta_{\mathrm{aux}}^{(t)}$.
    Calculate $s_{ij}$ based on $\mathbf{f}_i^{(t)}$ with reference to $\mathcal{B}$.
    Calculate the gradients based on (8).
    Back-propagate the gradients.
    Update the parameters $\theta_{\mathrm{aux}}$ of the auxiliary CNN
      via (10).
    Update $\mathcal{B}$ via (11).
endfor
Ensure: $\theta$, $\mathcal{B}$
```

1) *NWPU-RESISC45* [48]: This dataset is a large-scale RS archive, which contains 31500 images uniformly distributed in 45 land-cover types: airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, dense residential, desert, forest, freeway, golf course, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snow-berg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station, and wetland. Fig. 2(a) shows some examples of this dataset. All the scenes are RGB images with a total size of 256 × 256 pixels and spatial resolution ranging from 30 to 0.2 m. The NWPU-RESISC45 collection is publicly available.[2]

2) *EuroSAT* [55]: This dataset consists of 27000 labeled and geo-referenced Sentinel-2 images with size of 64×64

[2] [Online]. Available: http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html

pixels, spatial resolution of 10 m, and a total of 13 spectral bands covering the wavelength region from 443 to 2190 nm of the electromagnetic spectrum. Each scene belongs to one of the following 10 semantic land-cover categories: Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River, and Sea Lake. Some examples of this dataset are displayed in Fig. 2(b). The EuroSAT collection is also publicly available.[3]

B. Experimental Setup

Our experiments have been designed to test performance of the proposed deep-hashing method to retrieve RS images using a Hamming distance-based ranking. Regarding the retrieval configuration, the considered RS datasets have been randomly split into training, validation, and testing partitions, containing 70%, 10%, and 20% of the total number of samples, respectively. In more details, the training partition is used to learn the parameters of the hashing model whereas the testing partition serves as an external query set for retrieval. That is, each RS image in the testing partition is raised as an external query sample to retrieve images from the training partition. The retrieval performance is then evaluated using three different figures of merit: 1) precision; 2) recall; and 3) mean average precision (MAP). 13 defines the average precision (AP) expression

$$\mathrm{AP} = \frac{1}{Q}\sum_{r=1}^{R} P(r)\delta(r) \tag{13}$$

where Q is the number of ground-truth RS images in the dataset that are relevant with respect to the query, $P(r)$ denotes the precision for the top r retrieved images, and $\delta(r)$ is the indicator function to specify whether the rth relevant image is truly relevant to the query.

The proposed method has been implemented in PyTorch[4] and the selected backbone CNN architecture is the ResNet18 [56] model. Although other backbone networks could be used with the proposed loss and optimization mechanisms, we employ the ResNet18 in this article for the sake of simplicity. Regarding the selected parameters, σ and m are set to 0.1 and 0.5, respectively. The stochastic gradient descent optimizer is adopted for training. Besides, the initial learning rate is set to 0.01, and it is decayed by 0.5 every 30 epochs. Finally, the batch size is set to 256 and we totally train the model for 100 epochs.

To validate the effectiveness of the proposed method, we test four different code length values [16,32,64,128] to produce the corresponding binary codes for the retrieval process. In addition, we consider multiple state-of-the-art hashing methods for the experimental comparison: 1) LSH [25]; 2) PCA-ITQ [39]; 3) PCA-RR [39]; 4) DSH [26]; 5) SP [40]; 6) DHCNN [30]; 7) the triplet loss utilized in MiLaN [31] (simply termed as Triplet hereinafter); 8) DPSH [50]; and 9) DHNN [29]. In the case of conventional methods, the ResNet18 [56] trained based on the CE loss is served for extracting the input features. For training DHCNN, Triplet, DPSH, and DHNN, we fine-tune the associated learning rates (using the validation set) while utilizing the aforementioned parameter configuration to conduct a fair experimental comparison. All the experiments are conducted on an NVIDIA Tesla P100 graphics processing unit.

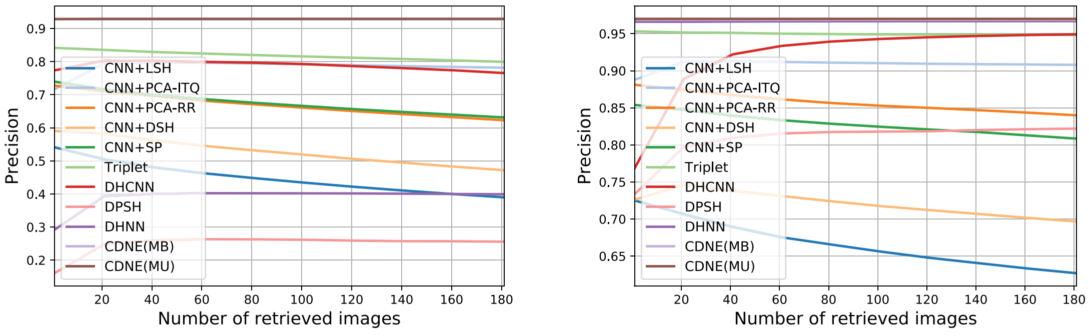
C. Experimental Results

1) Precision–Recall: Fig. 3 demonstrates the precision values with respect to the number of retrieved images when $L = 64$. When the number of retrieved images increases, the proposed method achieves, with a wide margin, the highest precision on NWPU-RESISC45 and also the best/second-best performance on EuroSAT being always among the highest results. Analyzing the other two deep hashing methods included in this experiment (i.e., DHCNN and Triplet), both CDNE(MU) and CDNE(MB) can better discover the locality structure of the images in the Hamming space based on their semantic information. In addition, Fig. 4 displays the precision–recall curves of the retrieval performances including two additional deep hashing methods in the comparison (i.e., DPSH and DHNN). It can be seen that the proposed method achieves the best performance compared to the other tested methods. For the conventional methods, the features are extracted from the CNN model trained via CE loss. For the CE loss, it has a prominent capability for class-wise discrimination. However, for the heterogeneous features extracted from complex RS scenes, the CE loss cannot well uncover the inherent locality structure for the images within individual classes. For the contrastive and triplet losses utilized in the considered deep hashing methods, the possible pairs and triplets are with the order of $\mathcal{O}(|\mathcal{X}|^2)$ and $\mathcal{O}(|\mathcal{X}|^3)$. With scalable RS datasets, they cannot be sampled sufficiently during a limited number of epochs for training. Thus, the CNN model cannot fully capture the relationships of the images in the produced Hamming space. On the contrary, the proposed approach exploits the memory bank and momentum update for updating the feature results progressively and the state of the CNN model can be optimized based on the calculated losses of the similarities between the features within a broader range of RS scenes. Therefore, the CNN model can be sufficiently trained based on the proposed method and the feature variations within individual intraclasses in the Hamming space can be captured.

2) MAP: Based on the image retrieval results of the compared methods, we calculate their MAP values when $R = 100$ and report them in Table I as quantitative assessment. From the obtained results, we can see that the proposed CDNE losses outperform the other methods by a large margin. For example, when $L = 16$, CDNE(MU) improves the MAP by 23% compared with the triplet loss used in MiLaN. As L increases from 16 to 128, CDNE(MU) maintains a stable image retrieval accuracy, which indicates that it is suitable for encoding scalable RS archives with low-dimensional hash codes. As an illustrative example, we present some image retrieval results (obtained with $L = 32$) from both datasets: NWPU-RESISC45 in Fig. 5(a); and EuroSAT in Fig. 5(b). Given two query images from the datasets, we retrieve their top 40 nearest neighbors based on the CDNE(MU), DHCNN, and Triplet losses, and demonstrate

[3] [Online]. Available: http://madm.dfki.de/files/sentinel/EuroSATallBands.zip

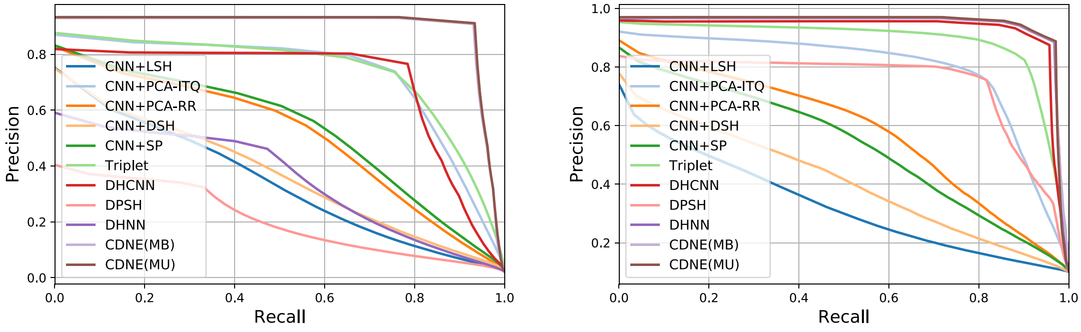
[4] [Online]. Available: https://pytorch.org/



(a)

(b)

Fig. 3. Mean precision versus the number of images retrieved based on the considered methods when $L = 64$. (a) NWPU-RESISC45. (b) EuroSAT.



(a)

(b)

Fig. 4. Precision-recall curves of the compared methods when $L = 64$. (a) NWPU-RESISC45. (b) EuroSAT.

TABLE I
MAP (%) Obtained by All the Methods on the Benchmark Datasets With Various l When $r = 100$

	NWPU-RESISC45				EuroSAT			
	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
CNN+LSH	32.98	50.38	67.82	77.35	49.44	69.37	81.24	87.89
CNN+PCA-ITQ	63.35	82.59	87.54	88.66	92.08	93.61	93.44	93.64
CNN+PCA-RR	56.70	70.79	78.07	81.36	84.99	87.58	89.57	90.93
CNN+DSH	38.84	58.49	68.03	74.36	73.08	75.01	82.53	88.90
CNN+SP	51.78	71.32	78.98	84.64	68.51	85.17	89.97	91.34
DPSH	29.94	31.09	42.51	47.98	57.10	85.76	98.81	98.40
DHNN	29.72	46.38	60.07	70.12	90.32	99.50	99.97	99.98
Triplet	76.02	84.21	87.39	89.25	94.94	96.73	97.22	98.02
DHCNN	98.07	92.12	94.24	99.95	99.51	99.29	99.95	99.38
CDNE(MB)	**99.69**	99.83	**99.91**	**99.97**	99.89	**99.99**	99.97	99.99
CDNE(MU)	99.19	**99.86**	99.87	99.96	**99.93**	99.97	**100.00**	**100.00**

some selected images in the result. In Fig. 5(a), some images belonging to *commercial area* are mistakenly considered as relevant images with respect to the query image of *church*, according to the result of DHCNN. In the Triplet result of Fig. 5(b), some *Industrial* images are wrongly retrieved with respect to the query image of *Residential*. This visual results reveal, from a qualitative perspective, that our approach can precisely retrieve the relevant images even in the presence of pattern variations within these classes.

3) Ablation Study: In order to further analyze the retrieval performances with respect to each loss item in CDNE, we conduct the corresponding ablation study on CDNE(MU). Specifically, we separately remove the SNCA and CE loss terms in CDNE(MU), which are termed as *CDNE(MU)-w/o-SNCA* and *CDNE(MU)-w/o-CE*, respectively, and calculate the MAP scores on the two benchmark datasets. As shown in Table II, when we remove the SNCA term, the image retrieval performance is significantly reduced, since the SNCA term is mainly utilized for preserving the semantic similarities among the images in the Hamming space. When we remove the CE term, the retrieval performance is slightly reduced. This indicates that the integration of the CE term can improve the quality of hash code generation by further enforcing the class discrimination capability.

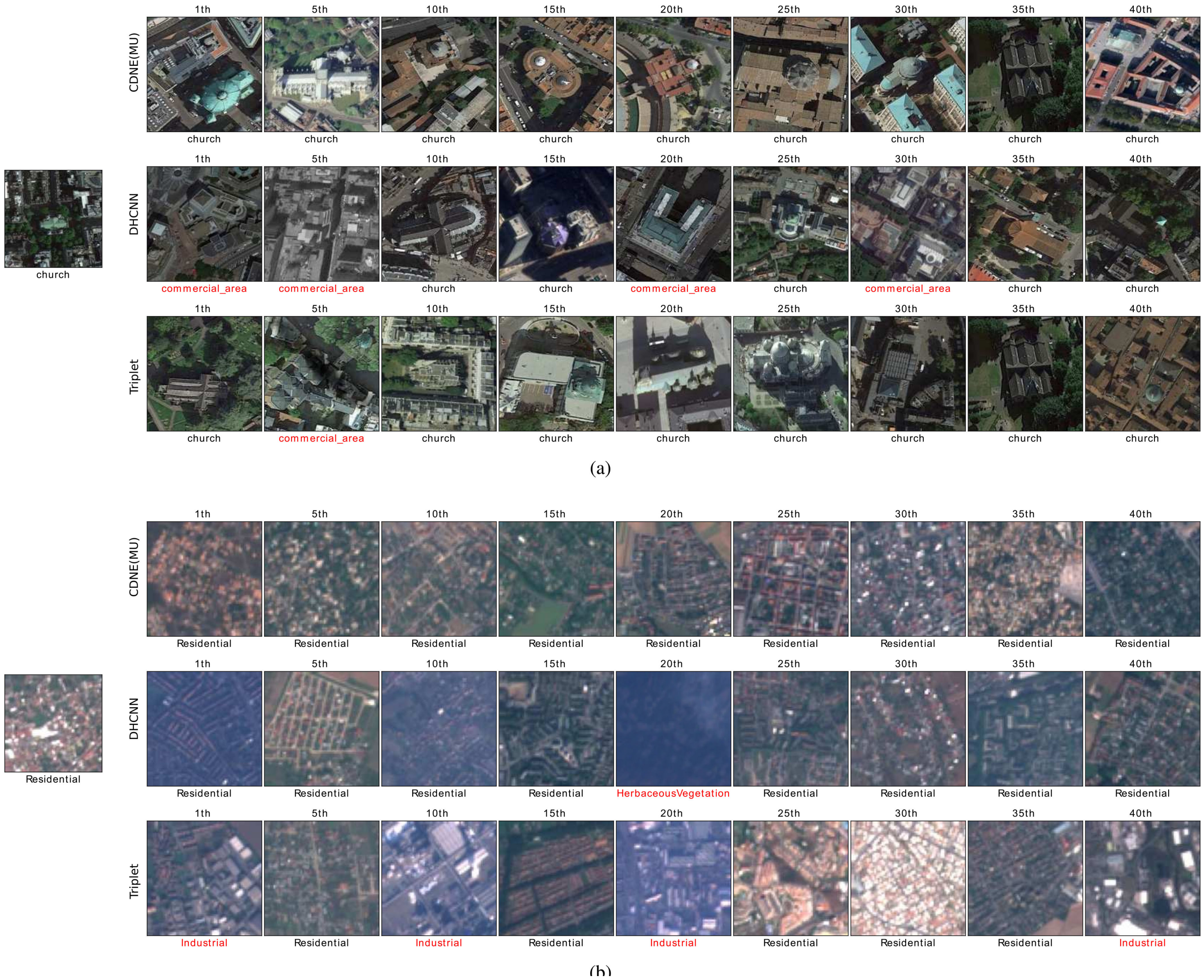

church
CDNE(MU)
1th 5th 10th 15th 20th 25th 30th 35th 40th
church church church church church church church church church
DHCNN
1th 5th 10th 15th 20th 25th 30th 35th 40th
commercial_area commercial_area church church commercial_area church commercial_area church church
Triplet
1th 5th 10th 15th 20th 25th 30th 35th 40th
church commercial_area church church church church church church church
(a)
Residential
CDNE(MU)
1th 5th 10th 15th 20th 25th 30th 35th 40th
Residential Residential Residential Residential Residential Residential Residential Residential Residential
DHCNN
1th 5th 10th 15th 20th 25th 30th 35th 40th
Residential Residential Residential Residential HerbaceousVegetation Residential Residential Residential Residential
Triplet
1th 5th 10th 15th 20th 25th 30th 35th 40th
Industrial Residential Industrial Residential Industrial Residential Residential Residential Industrial
(b)

Fig. 5. Examples of retrieved images based on CDNE(MU), DHCNN, and Triplet on the two datasets. (a) NWPU-RESISC45. (b) EuroSAT.

TABLE II

MAP (%) Obtained by CDNE(MU)-w/o-SNCA, CDNE(MU)-w/o-CE, and CDNE(MU) on the Benchmark Datasets With Various l When $r = 100$

	NWPU-RESISC45				EuroSAT			
	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
CDNE(MU)-w/o-SNCA	91.94	92.95	92.12	90.70	99.54	99.75	99.73	99.89
CDNE(MU)-w/o-CE	99.17	99.84	**99.87**	**99.96**	99.92	**99.97**	99.98	99.99
CDNE(MU)	**99.19**	**99.86**	**99.87**	**99.96**	**99.93**	**99.97**	**100.00**	**100.00**

TABLE III

MAP (%) Scores for Sensitivity Analysis on λ When $l = 64$ and $r = 100$

	NWPU-RESISC45	EuroSAT
$\lambda = 0.05$	99.87	99.97
$\lambda = 0.1$	99.87	100.00
$\lambda = 0.15$	99.93	99.98
$\lambda = 0.2$	99.90	100.00

4) Hyperparameter Analysis: Additionally, Table III shows the MAP scores (with $L = 64$ and $R = 100$) achieved by the proposed approach when varying the λ hyperparameter. As it is possible to see, the obtained retrieval results are consistent throughout the different values considered in both datasets, which denote the stability of the CDNE(MU) hashing method with respect to λ.

V. Conclusion and Future Lines

In this article, we propose a new deep-hashing method for RS CBIR, termed CDNE. Our newly proposed method learns a Hamming space where the locality structures of the RS scenes can be preserved while the land-cover semantic categories can be effectively discriminated. Specifically, we define a joint loss composed of: first, SNCA, aimed at discovering the neighborhood structures of the images with heterogeneous intraclass variations; second, CE, aimed at separating images with different semantic categories; and third, the quantization loss, aimed at pushing the feature outputs of CNN models toward the associated binary codes.

Our extensive experiments demonstrate that the proposed method outperforms other state-of-the-art conventional and deep-hashing methods. Compared with the widely used contrastive and triplet losses in deep-hashing methods, the proposed loss function can be sufficiently optimized by utilizing the memory bank, so that semantically similar images can be better grouped and semantically dissimilar images can be better separated in the resulting Hamming space.

As future work, we plan to develop further research to extend the proposed hashing system with additional hyperparameters to make it adaptive to images with multilabel information while also analyzing its implications.

References

[1] J. A. Benediktsson, J. Chanussot, and W. M. Moon, "Very high-resolution remote sensing: Challenges and opportunities," in *Proc. IEEE*, vol. 100, no. 6, pp. 1907–1910, Jun. 2012.

[2] 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.

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

[4] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-lidar and hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1470–1474, Aug. 2020.

[5] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, Mar. 2018.

[6] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 44–59, 2018.

[7] 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.

[8] N. Zerrouki, F. Harrou, and Y. Sun, "Statistical monitoring of changes to land cover," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 6, pp. 927–931, Jun. 2018.

[9] 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.

[10] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote sensing image classification and retrieval," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3016020.

[11] A. S. Belward and J. O. Skøien, "Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites," *ISPRS J. Photogrammetry Remote Sens.*, vol. 103, pp. 115–128, 2015.

[12] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," in *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.

[13] X. X. Zhu *et al.*, "So2Sat LCZ42: A benchmark dataset for global local climate zones classification," 2019, *arXiv:1912.12171*.

[14] Y. Long *et al.* "DIRS: On creating benchmark datasets for remote sensing image interpretation," 2020, *arXiv: 2006.12485*.

[15] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.*, vol. 40, no. 1, pp. 262–282, 2007.

[16] A. Alzubi, A. Amira, and N. Ramzan, "Semantic content-based image retrieval: A comprehensive study," *J. Vis. Commun. Image Representation*, vol. 32, pp. 20–54, 2015.

[17] Y. Ma *et al.*, "Remote sensing big data computing: Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 51, pp. 47–60, 2015.

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

[19] B. Demir and L. Bruzzone, "Hashing-based scalable remote sensing image search and retrieval in large archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 892–904, Feb. 2016.

[20] D. Ye, Y. Li, C. Tao, X. Xie, and X. Wang, "Multiple feature hashing learning for large-scale remote sensing image retrieval," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 11, 2017, Art. no. 364.

[21] P. Li, X. Zhang, X. Zhu, and P. Ren, "Online hashing for scalable remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 5, 2018, Art. no. 709.

[22] J. Wang, W. Liu, S. Kumar, and S.-F. Chang, "Learning to hash for indexing big data," in *Proc. IEEE*, vol. 104, no. 1, pp. 34–57, Jan. 2016.

[23] P. Li and P. Ren, "Partial randomness hashing for large-scale remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 464–468, Mar. 2017.

[24] S. Li *et al.*, "Geospatial big data handling theory and methods: A review and research challenges," *ISPRS J. Photogrammetry Remote Sens.*, vol. 115, pp. 119–133, 2016.

[25] M. Slaney and M. Casey, "Locality-sensitive hashing for finding nearest neighbors [lecture notes]," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 128–131, Apr. 2008.

[26] Z. Jin, C. Li, Y. Lin, and D. Cai, "Density sensitive hashing," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1362–1371, Aug. 2013.

[27] W. Cao, W. Feng, Q. Lin, G. Cao, and Z. He, "A review of hashing methods for multimodal retrieval," *IEEE Access*, vol. 8, pp. 15 377–15 391, 2020.

[28] Y. Gu, Y. Wang, and Y. Li, "A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection," *Appl. Sci.*, vol. 9, no. 10, 2019, Art. no. 2110.

[29] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 950–965, Feb. 2017.

[30] W. Song, S. Li, and J. A. Benediktsson, "Deep hashing learning for visual and semantic retrieval of remote sensing images," 2019, *arXiv:1909.04614*.

[31] S. Roy, E. Sangineto, B. Demir, and N. Sebe, "Metric-learning-based deep hashing network for content-based retrieval of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: 10.1109/LGRS.2020.2974629.

[32] P. Li *et al.*, "Hashing nets for hashing: A quantized deep learning to hash framework for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7331–7345, Oct. 2020.

[33] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2018.

[34] 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.

[35] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in *Proc. Int. Conf. Comput. Vision*, vol. 9, 2009, pp. 2130–2137.

[36] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1–8.

[37] W. Kong and W.-J. Li, "Isotropic hashing," in *Proc. Adv. Neural. Inf. Process. Syst.*, 2012, pp. 1646–1654.

[38] Y. Lin, R. Jin, D. Cai, S. Yan, and X. Li, "Compressed hashing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 446–451.

[39] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2012.

[40] Y. Xia, K. He, P. Kohli, and J. Sun, "Sparse projections for high-dimensional binary codes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3332–3339.

[41] T. Reato, B. Demir, and L. Bruzzone, "An unsupervised multicode hashing method for accurate and scalable remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 276–280, Feb. 2019.

[42] 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.*, to be published.

[43] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state-of-the-art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[44] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 815–823.

[45] K. Nogueira, O. A. Penatti, and J. A. Dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, 2017.

[46] 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.*, to be published, doi: 10.1109/TGRS.2020.2991657.

[47] 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.

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

[49] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2019, *arXiv:1911.05722*.

[50] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," 2015, *arXiv:1511.03855*.

[51] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 513–520.

[52] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[53] Z. Wu, A. A. Efros, and S. X. Yu, "Improving generalization via scalable neighborhood component analysis," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 685–701.

[54] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3007029.

[55] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

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

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

In August of 2018, he was a Guest Researcher with the Institute of Computer Graphics and Vision, TU Graz, Graz, Austria. From 2019 to 2020, he was with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany. He is currently with the School of Electronic and Information Engineering, Soochow University, Suzhou, China. His research focuses on signal processing and machine learning techniques, and their applications in remote sensing. In particular, he is interested in multi-dimensional data analysis, geophysical parameter estimation based on InSAR data, SAR denoising, and deep learning based techniques for remote sensing image analysis.

Dr. Kang was the recipient of first place of the Best Student Paper Award in EUSAR 2018, Aachen, Germany. His joint work was selected as one of the 10 Student Paper Competition Finalists, in IGARSS 2020.

Ruben Fernandez-Beltran (Member, IEEE) received a B.Sc. degree in computer science, an M.Sc. degree in intelligent systems, and a Ph.D. degree in computer science from the 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, University Jaume I, as a member of the Institute of New Imaging Technologies. He has been a Visiting Researcher with the University of Bristol (U.K.), University of Cáceres (Spain), and Technische Universität Berlin (Germany). His research interests lie in multimedia retrieval, spatio-spectral image analysis, pattern recognition techniques applied to image processing and remote sensing.

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

Zhen Ye received the B.E. degree in geographic information system and Ph.D. degree in cartography and geographic information engineering from Tongji University, Shanghai, China, in 2011 and 2018, respectively.

He is currently a Postdoctoral Researcher with the Chair of Photogrammetry and Remote Sensing, Technical University of Munich, Munich, Germany. His research interests include photogrammetry and remote sensing, high-accuracy image registration, and high-resolution satellite image processing.

Xiaohua Tong (Senior Member, IEEE) received the Ph.D. degree in geographic information system from Tongji University, Shanghai, China, in 1999.

Between 2001 and 2003, he was a Postdoctoral Researcher with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China. He was a Research Fellow with Hong Kong Polytechnic University, Hong Kong, in 2006, and a Visiting Scholar with the University of California, Santa Barbara, CA, USA, between 2008 and 2009. He is currently a Professor with the College of Surveying and Geoinformatics, Tongji University. His research interests include remote sensing, geographic information system, uncertainty and spatial data quality, and image processing for high-resolution and hyper-spectral images.

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

He is currently the Head of the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura. He has authored more than 600 publications, including around 300 JCR journal articles (more than 170 in IEEE journals), 23 book chapters, and around 300 peer-reviewed conference proceeding papers. His research interests include hyperspectral data processing and parallel computing of remote sensing data.

Dr. Plaza was a member of the Editorial Board of the IEEE Geoscience and Remote Sensing Newsletter, from 2011 to 2012 and the IEEE Geoscience and Remote Sensing Magazine, in 2013. He was also a member of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS). He is also a fellow of IEEE for contributions to hyperspectral data processing and parallel computing of earth observation data. He received the recognition as a Best Reviewer of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, in 2009, and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, in 2010, for which he has served as an Associate Editor from 2007 to 2012. He was also a recipient of the Most Highly Cited Paper (2005–2010) in the *Journal of Parallel and Distributed Computing*, the 2013 Best Paper Award of the IEEE JSTARS, and the Best Column Award of the *IEEE Signal Processing Magazine*, in 2015. He was the recipient of the 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 more than 50 different journals. He has served as the Editor-in-Chief of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING from 2013 to 2017. He has guest-edited ten special issues on hyperspectral remote sensing for different journals. He is also an Associate Editor of IEEE ACCESS (received the recognition as an Outstanding Associate Editor of the journal, in 2017). He is currently serving as the Editor-in-Chief of the IEEE JOURNAL ON MINIATURIZATION FOR AIR AND SPACE SYSTEMS. For more information please visit: http://www.umbc.edu/rssipl/people/aplaza.