Deep Hashing Learning for Visual and Semantic Retrieval of Remote Sensing Images

Weiwei Song, *Student Member, IEEE*, Shutao Li, *Fellow, IEEE*, and Jón Atli Benediktsson, *Fellow, IEEE*

***Abstract*— Driven by the urgent demand for managing remote sensing big data, large-scale remote sensing image retrieval (RSIR) attracts increasing attention in the remote sensing field. In general, existing retrieval methods can be regarded as visual-based retrieval approaches that search and return a set of similar images to a given query image from a database. Although these retrieval methods have delivered good results, there is still a question that needs to be addressed: can we obtain the accurate semantic labels of the returned similar images to further help analyzing and processing imagery? To this end, in this article, we redefine the image retrieval problem as visual and semantic retrieval of images. Especially, we propose a novel deep hashing convolutional neural network (DHCNN) to retrieve similar images and classify their semantic labels simultaneously in a unified framework. In more detail, a convolutional neural network (CNN) is used to extract high-dimensional deep features. Then, a hash layer is perfectly inserted into the network to transfer the deep features into compact hash codes. In addition, a fully connected layer with a softmax function is performed on the hash layer to generate the probability distribution of each class. Finally, a loss function is elaborately designed to consider the label loss of each image and similarity loss of pairs of images simultaneously. Experimental results on three remote sensing data sets demonstrate that the proposed method can achieve state-of-art retrieval and classification performance.**

***Index Terms*— Classification, deep learning, hashing learning, remote sensing, retrieval.**

I. Introduction

With the rapid development of remote sensing observation technology, the acquisition of remote sensing images has been largely enhanced not only in volume but also in resolution. However, these large-scale and high-resolution remote sensing images have also resulted in the significant challenge of how to efficiently manage and analyze the remote sensing big data. Over the past several decades, remote sensing image retrieval (RSIR), which aims to search and return a set of similar images from a database to a given query image, has received increased interest in the remote sensing community.

For RSIR, one of the challenges is how to design a retrieval system to return similar images in an accurate and efficient manner. Early retrieval methods for remote sensing images mainly exploited manually annotated tags, e.g., geographical location, acquisition time, or sensor type, to search for similar images. This kind of approach, called text-based image retrieval, usually obtains imprecise retrieval results since the visual information of images cannot be fully represented via annotated tags. By contrast, content-based image retrieval (CBIR) that employs the features extracted directly from the images for retrieval tasks has achieved great success in recent years [1], [2]. A CBIR system mainly has two stages: feature extraction and similarity computation. In the first stage, the query image and images from the database are represented by feature descriptors, respectively. Then, a set of relevant images is searched by ranking the similarity between the query image and each item from the database in the second stage. The extracted features can be divided into three types: low-, mid-, and high-level features. Designing a low-level feature descriptor requires engineering skills and domain expertise. Various low-level features have been exploited in RSIR, such as spectral features [3], texture features [4], and shape features [5]. Mid-level features exhibit superiority over low-level features in representing remote sensing images by exploiting powerful encoding techniques, e.g., bag-of-visual words (BoVW) [6], Fisher vector (FV) [7], and vector of locally aggregated descriptors (VLAD) [8]. However, the above features belong to handcrafted features that are limited to accurately describe the semantic information of remote sensing images.

Recently, deep learning has made a great breakthrough in the computer vision field due to its powerful ability for feature extraction [9], [10]. Motivated by those successful applications, deep learning has been introduced into the remote sensing field, including the hyperspectral image classification [11]–[14] and remote sensing scene recognition [15]–[17]. In addition, researchers have also attempted to take advantage of high-level features extracted from deep neural networks for RSIR [18]. Especially, the deep features derived from convolutional neural networks (CNNs) are used to represent remote sensing images and further retrieve relevant images [19], [20].

Manuscript received July 19, 2020; revised September 24, 2020; accepted October 27, 2020. This work was supported in part by the National Natural Science Fund of China under Grant 61890962 and Grant 61520106001; in part by the Science and Technology Plan Project Fund of Hunan Province under Grant CX2018B171, Grant 2017RS3024, and Grant 2018TP1013; and in part by the Science and Technology Talents Program of the Hunan Association for Science and Technology under Grant 2017TJ-Q09. *(Corresponding author: Shutao Li.)*

Weiwei Song and Shutao Li are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China, and also with the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Changsha 410082, China (e-mail: weiwei_song@hnu.edu.cn; shutao_li@hnu.edu.cn).

Jón Atli Benediktsson is with the Faculty of Electrical and Computer Engineering, University of Iceland, 101 Reykjavk, Iceland (e-mail: benedikt@hi.is).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TGRS.2020.3035676.

Digital Object Identifier 10.1109/TGRS.2020.3035676

Nevertheless, most existing retrieval methods, including handcrafted features-based methods and deep features-based methods, adopt the Euclidean distance as similarity criteria, which is no longer suitable for real-time retrieval goals due to the time-consuming computation. In order to overcome the above problem, hashing methods have been largely developed for RSIR [21]–[23]. These hashing-based approaches aim to learn a set of hash functions to encode the high-dimensional image features into low-dimensional Hamming space, where each image is represented by a binary hash code. By generating a hash-code table for all images, the retrieval can be easily completed via hash lookup or Hamming ranking. Recently, deep hashing-based methods that take full advantage of deep networks and hashing learning deliver better performance for RSIR. For example, Li *et al.* [22] introduced deep hashing neural networks (DHNNs) for large-scale RSIR. In such work, a deep feature learning neural network and a hashing learning neural network were exploited to learn high-level semantic features and compact hash codes, respectively. In [23], cross-source RSIR was investigated via source-invariant deep hashing CNNs (DHCNNs). In [24], a metric and hash-code learning network was proposed to learn a semantic-based metric space while simultaneously producing binary hash codes for fast and accurate retrieval of remote sensing images in large archives.

However, existing image retrieval approaches that only return similar images to a given query image from a database may no longer meet the need for further image analysis and processing. Let us consider this question: given a query image, can we return similar images and, at the same time, obtain their semantic labels? In fact, solving this problem can bring in many advantages. First, we can better evaluate a retrieval algorithm with semantic labels of returned relevant images, especially for multilabel image retrieval tasks. In addition, we can also roughly predict the class distribution of an unknown database by analyzing the semantic labels of similar images.

In this article, we redefine the traditional image retrieval problem as visual and semantic retrieval of images, which aims to retrieve a set of visually similar images to a given query image and simultaneously classify their semantic labels. To this end, we propose a novel DHCNN to learn compact hash codes for efficient retrieval and discriminative features for accurate classification. In more detail, a pretrained CNN is adopted to extract high-dimensional deep features from raw remote sensing images. Then, a hash layer is built to encode the high-dimensional deep features into low-dimensional hash codes. In addition, a fully connected layer with a softmax function is used to generate class distribution. Finally, we elaborately design a loss function to train DHCNN in an end-to-end way. Once DHCNN is trained enough, for a query image, we can generate its hash code by binarizing the output of the hash layer, and then, the retrieval can be easily completed via Hamming distance ranking. In addition, the semantic labels of images, including the query image and its similar images, can be assigned by feeding their semantic features into the softmax classifier.

The main contributions of this article can be summarized as follows: a novel DHCNN is proposed to achieve fast retrieval and simultaneously accurate classification for the large-scale remote sensing images in a unified framework. Different from the existing deep hashing network-based methods, the proposed DHCNN can generate more discriminative feature representation by fusing similarity information between image pairs and semantic information of each image, which obtains better retrieval and classification results.

The rest of this article is organized as follows. Section II briefly introduces preliminary knowledge, including CNNs and hashing learning. Section III describes the proposed method in detail. The comprehensive experimental results and the corresponding analyses are presented in Section IV. Finally, Section V makes some concluding remarks.

II. Preliminary Knowledge

In this article, we aim to take advantage of CNNs and hashing learning to enhance the representation of features. In the feature space, images from the same classes are mapped closely to each other, and images from the different classes are mapped far apart. In this section, we will briefly introduce some preliminary knowledge, including CNNs and hashing learning.

A. Convolutional Neural Networks

Recently, CNNs have made a great breakthrough in many fields, e.g., image classification [9], object detection [25], and semantic segmentation [10]. A typical CNN mainly consists of a stack of alternating convolutional layers and pooling layers with a number of fully connected layers. If $\mathbf{X}^{l-1}$ is the input of a convolutional layer, the output of this layer can be computed by

$$\mathbf{X}^l = \sigma\left(\mathbf{X}_i^{l-1} * \mathbf{W}^l + \mathbf{b}^l\right) \tag{1}$$

where $\mathbf{W}^l$ and $\mathbf{b}^l$ are the weight and bias, respectively. The operator $*$ represents discrete convolution operation, and σ refers to the activation function that is utilized to improve the nonlinearity of the network. Subsequently, a pooling layer may be inserted after convolutional layers to reduce the spatial size of the feature maps, which can improve the robustness of features. Finally, all features of the previous layer are combined in fully connected layers to extract abstract semantic features. In addition, a softmax function is applied to the last fully connected layer to generate the probability distribution of classes.

In the past several years, there are many powerful CNNs, e.g., AlexNet [9], CaffeNet [26], GoogLeNet [27], VGG [28], and ResNet [29], that have been developed. Although these networks were trained on natural image data sets (e.g., ImageNet [30]), the extracted features still exhibit powerful generalization ability on remote sensing data sets [31]. In addition, considering that the available training samples are relatively small and labeling unknown samples is very difficult in the remote sensing field, we transfer existing deep CNNs to our DHCNN to reduce the need for training samples.

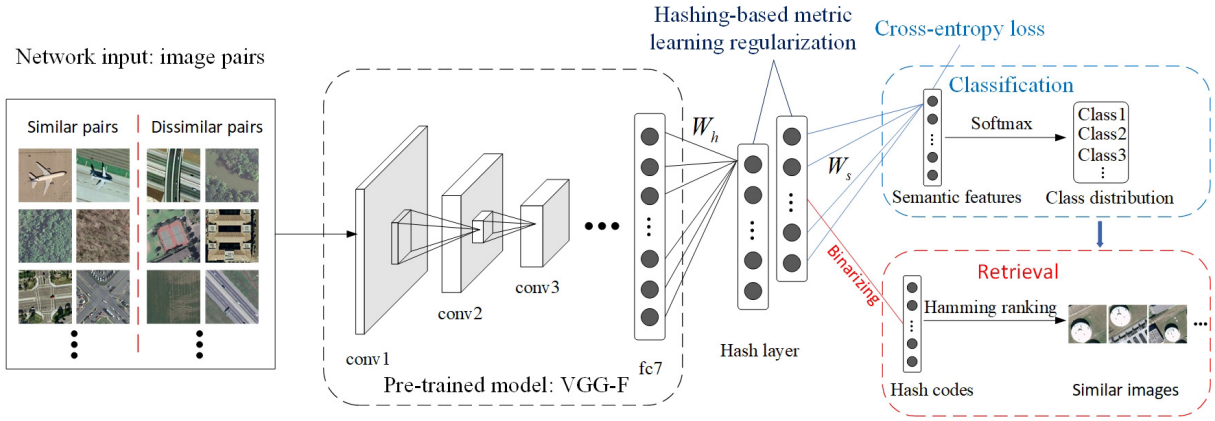

Fig. 1. Proposed DHCNN for visual and semantic retrieval of remote sensing images. First, a pretrained CNN is introduced to extract deep features. Then, a hash layer with metric learning regularization is used to transfer high-dimensional deep features into low-dimensional hash codes. Furthermore, a fully connected layer with a softmax classifier is applied to generate class distribution. With the hash codes and class distribution, a set of similar images to the given query image and their semantic labels are easily obtained via hashing ranking.

B. Hashing Learning

Due to its encouraging efficiency in both speed and storage, the hashing technique has been widely used in large-scale image retrieval [32]–[36]. Given a training set of N points $\{\mathbf{x}_i\}_{i=1}^N$, each point is represented as a D-dimensional feature vector. The goal of hashing is to learn a nonlinear function $f : \mathbf{x} \longmapsto h \in \{-1, 1\}^K$ to encode each point $\mathbf{x}$ into compact K-bit hash code $h = f(\mathbf{x})$. Existing learning-based hashing methods can be roughly divided into two categories: unsupervised hashing and supervised hashing.

Unsupervised hashing methods use the unlabeled training data to learn a set of hash functions that can encode input data points into binary codes. The representative methods include spectral hashing (SH) [37], iterative quantization (ITQ) [38], and density sensitive hashing [39]. By contrast, supervised hashing aims to generate similarity-preserving representations with shorter hash codes by utilizing supervised information, e.g., pointwise labels, pairwise labels, or ranking labels. In the past several years, there are many successful supervised hashing methods that have been developed for fast image retrieval, including binary reconstruction embedding (BRE) [40], minimal loss hashing (MLH) [41], sparse embedding and least variance encoding (SELVE) [42], and supervised hashing with kernels (KSH) [43]. By utilizing the supervised information, images from the same classes have small feature distance, while the images from the different classes have large features distance in the Hamming space.

Most of the existing hashing methods use handcrafted visual features to encode each input image, which may degrade their hashing performance because handcrafted features do not necessarily capture accurate similarity of images. Recently, many research studies have been focused on integrating the hashing technique into CNNs, which delivers satisfying performance for image retrieval. For example, Xia *et al.* [44] proposed a two-stage method to train CNN to fit binary codes computed from the pairwise similarity matrix. Li *et al.* [45] performed simultaneous feature learning and hash code learning for image retrieval with pairwise labels. In [46], a deep hashing with a regularized similarity learning framework was proposed to generate compact and bit-scalable hashing codes for image retrieval and person reidentification. In addition, Zhang *et al.* [47] focused on the problem of unsupervised deep hashing and discovered pseudolabels to train a deep network for scalable image retrieval.

TABLE I

CONFIGURATION OF DEEP NETWORK USED IN DEEP FEATURE EXTRACTION, WHICH IS TRANSFERRED FROM VGG-F [48]

Layer	Configuration
conv1	filter $96 \times 7 \times 7$, stride 2×2, pad 0, LRN, pool 3×3
conv2	filter $256 \times 5 \times 5$, stride 1×1, pad 1, pool 2×2
conv3	filter $512 \times 3 \times 3$, stride 1×1, pad 1
conv4	filter $512 \times 3 \times 3$, stride 1×1, pad 1
conv5	filter $512 \times 3 \times 3$, stride 1×1, pad 1, pool 3×3
full6	4096, dropout
full7	4096, dropout

III. PROPOSED FRAMEWORK

In order to cope with the challenge of high intraclass and low interclass variabilities that exist in remote sensing images, we combine deep learning and hashing learning to minimize the feature distance between similar image pairs and maximize the feature distance between dissimilar image pairs. To this end, we design an object function to simultaneously consider the semantic information of each image and similarity feature between image pairs. Through the proposed DHCNN, we can extract discriminative semantic features for accurate classification and learn compact hash codes for efficient retrieval. Fig. 1 illustrates the proposed DHCNN which consists of a pretrained CNN, a hash layer, and a fully connected layer with a softmax classifier. In the following part, we will introduce the proposed method in detail.

A. Deep Feature Extraction

In general, training a CNN from scratch requires a large number of training samples to learn the model parameters.

However, in the remote sensing field, the available training samples are relatively small, and labeling unknown samples is costly and time-consuming work. To solve this problem, we adopt a pretrained CNN model to decrease the burden on training samples. Here, we take the VGG-F [48] model as an example to explicate the part of deep feature extraction, which is illustrated in the black dotted box in Fig. 1. Especially, the network parameters of VGG-F, including the first five convolutional layers and two fully connected layers, are transferred to our DHCNN. The detailed configuration is shown in Table I. For convolutional layers, "filter" specifies the number of convolution filters and kernel size; "stride" and "pad" indicate the convolutional strides and spatial padding, respectively; "LRN" refers to local response normalization [9]; and "pool" specifies the max-pooling size. For fully connected layers, "4096" indicates the feature dimension, and the dropout technique [9] is applied to full6 and full7. The activation function for all weight layers is the REctification Linear Unit (ReLU) [9].

If there are N training samples denoted as $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, the corresponding set of labels can be represented as $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$, where $\mathbf{y}_i \in \mathbb{R}^C$ is the ground-truth vector of sample $\mathbf{x}_i$ with only one element being 1 and others being 0, where C is the total number of image scene classes. For arbitrary remote sensing image $\mathbf{x}_i \in \mathbf{X}$, we can extract its deep features (i.e., the output of the fc7 layer) denoted as $\mathbf{f}_i$ by

$$\mathbf{f}_i = \Phi(\mathbf{x}_i; \theta), \quad i = 1, 2, \ldots, N \tag{2}$$

where Φ is the network function characterized by the parameter θ existed in the first seven layers of VGG-F. This propagation actually performs a series of nonlinear and linear transformations, including convolution, pooling, and nonlinear mapping.

B. Hashing-Based Metric Learning

In light of the large intraclass and low interclass variabilities that exist in remote sensing images, we adopt hashing-based metric learning to constrain images from the same classes to be encoded as closely as possible and images from the different classes to be encoded far away each other in feature space. To this end, we use pairwise input to train our network, which can explore the similarity/disimilarity information between images. Let $(\mathbf{x}_i, \mathbf{x}_j)$ be a pair of images, and we define its label s_{ij} such that $s_{ij} = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ come from same class and 0 otherwise. As mentioned earlier, we can easily obtain their deep features $(\mathbf{f}_i, \mathbf{f}_j)$ via forward propagation. Subsequently, a hash layer is inserted after the pretrained CNN to transfer the high-dimensional deep features into compact K-bit hash codes, which can be formulated as

$$\mathbf{b}_t = \operatorname{sgn}(\mathbf{u}_t), \quad t = i, j \tag{3}$$

where $\mathbf{u}_t = \mathbf{W}_h \mathbf{f}_t + \mathbf{v}_h$ is the hash-like feature, $\mathbf{W}_h \in \mathbb{R}^{K \times 4096}$ denotes a weight matrix, $\mathbf{v}_h \in \mathbb{R}^{K \times 1}$ denotes a bias vector, and $\operatorname{sgn}(\cdot)$ performs elementwise operation for a matrix or a vector, i.e., $\operatorname{sgn}(x) = 1$ if $x > 0$ and -1 otherwise.

After obtaining the hash codes $\mathbf{B} = \{\mathbf{b}_t\}_{t=1}^N$ for all the samples, the likelihood of the pairwise labels $S = \{s_{ij}\}$ can be defined as

$$p(s_{ij}|\mathbf{B}) = \begin{cases} \varphi(\omega_{ij}), & s_{ij} = 1 \\ 1 - \varphi(\omega_{ij}), & s_{ij} = 0 \end{cases} \tag{4}$$

where $\varphi(\cdot)$ is the logistic function and $\varphi(x) = (1/(1 + e^{-x}))$, $\omega_{ij} = (1/2)\mathbf{b}_i^T \mathbf{b}_j$. Based on the above definition, the loss function can be given by taking the negative log-likelihood of the observed pairwise labels in S

$$\begin{aligned} \mathcal{L}_1 &= -\log\ p(S|\mathbf{B}) = -\sum_{s_{ij} \in S} \log\ p(s_{ij}|\mathbf{B}) \\ &= -\sum_{s_{ij} \in S} (s_{ij}\omega_{ij} - \log(1 + e^{\omega_{ij}})). \end{aligned} \tag{5}$$

However, directly solving the problem (5) is very hard due to the discrete values in formulation. Motivated by Li *et al.* [45], the above loss function can be reformulated in a discrete way

$$\mathcal{L}_2 = -\sum_{s_{ij} \in S} (s_{ij}\psi_{ij} - \log(1 + e^{\psi_{ij}})) + \beta \sum_{i=1}^N \|\mathbf{u}_i - \mathbf{b}_i\|_2^2 \tag{6}$$

where $\psi_{ij} = (1/2)\mathbf{u}_i^T \mathbf{u}_j, i, j = 1, 2, \ldots, N$, β is a regularization parameter that can constrain $\mathbf{u}_i$ approach to $\mathbf{b}_i$. Through minimizing $\mathcal{L}_2$, the feature distance in the Hamming space (i.e., Hamming distance) between similar samples can be optimized to be as small as possible, and the Hamming distance between dissimilar samples becomes as large as possible.

C. Object Function and Solving

Different from the existing deep hashing methods for image retrieval that only utilize similarity information between images to learn hash codes [23], [45], we also consider semantic information of each image to further improve the ability of feature representation. To this end, a fully connected layer with a softmax function is added after the hash layer to generate the class distribution for each image. This procedure can be represented by

$$\mathbf{t}_k = \operatorname{softmax}(\mathbf{W}_s \mathbf{u}_k + \mathbf{v}_s), \quad k = 1, 2, \ldots, N \tag{7}$$

where $\mathbf{W}_s \in \mathbb{R}^{C \times K}$ and $\mathbf{v}_s \in \mathbb{R}^{C \times 1}$ denote the weight matrix and bias vector, respectively. Then, we adopt cross-entropy loss to minimize the error between the predicted label and ground-truth label

$$\mathcal{L}_3 = -\frac{1}{N} \sum_{i=1}^N \langle \mathbf{y}_i, \log \mathbf{t}_i \rangle \tag{8}$$

where $\langle\rangle$ represents inner production operation. By minimizing the loss function $\mathcal{L}_3$, the CNN can learn discriminative semantic features of each images.

As mentioned earlier, loss function $\mathcal{L}_2$ aims to learn the similarity information between images, and $\mathcal{L}_3$ aims to learn the label information of each image. Here, we design a new loss function to simultaneously consider the similarity

information and label information to improve the network performance. This new loss function is defined as

$$\mathcal{L}_4 = \eta\mathcal{L}_2 + (1-\eta)\mathcal{L}_3 \tag{9}$$

where $\eta \in [0, 1]$ is the regularization parameter to balance the label information and similarity information. Especially, when $\eta = 0$, the object function only utilizes label information of each image. On the other hand, only the similarity information between images is considered when $\eta = 1$. Finally, our object function is to minimize the loss function $\mathcal{L}_4$, that is

$$\begin{aligned}\mathcal{J} = \min\mathcal{L}_4 = \min\Bigg\{\eta\Bigg(-\sum_{s_{ij}\in S}(s_{ij}\psi_{ij} - \log(1+e^{\psi_{ij}})) \\ + \beta\sum_{i=1}^{N}\|\mathbf{u}_i - \mathbf{b}_i\|_2^2\Bigg) \\ + (1-\eta)\left(-\frac{1}{N}\sum_{i=1}^{N}\langle\mathbf{y}_i, \log\mathbf{t}_i\rangle\right)\Bigg\}.\end{aligned} \tag{10}$$

In order to solve the problem in (10), we adopt the stochastic gradient descent (SGD) algorithm to learn the parameters, including $\mathbf{W}_h$, $\mathbf{W}_s$, $\mathbf{v}_h$, $\mathbf{v}_s$, and θ. First, we compute the gradients of the objective function $\mathcal{J}$ with respect to $\mathbf{t}_i$ and $\mathbf{u}_i$, which can be represented by

$$\frac{\partial\mathcal{J}}{\partial\mathbf{t}_i} = (1-\eta)\frac{\partial\mathcal{L}_3}{\partial\mathbf{t}_i} = -(1-\eta)\frac{1}{N}\frac{\mathbf{y}_i}{\mathbf{t}_i} \tag{11}$$

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial\mathbf{u}_i} &= \eta\frac{\partial\mathcal{L}_2}{\partial\mathbf{u}_i} + (1-\eta)\frac{\partial\mathcal{L}_3}{\partial\mathbf{u}_i} \\ &= \eta\left(\frac{1}{2}\sum_{j:s_{ij}\in S}(a_{ij}-s_{ij})\mathbf{u}_j \right. \\ &\quad \left. + \frac{1}{2}\sum_{j:s_{ji}\in S}(a_{ji}-s_{ji})\mathbf{u}_j + 2\beta(\mathbf{u}_i - \mathbf{b}_i)\right) \\ &\quad + (1-\eta)\left(-\frac{1}{N}\mathbf{W}_s^T(\mathbf{y}_i - \mathbf{t}_i)\right)\end{aligned} \tag{12}$$

where $a_{ij} = \varphi((1/2)\mathbf{u}_i^T\mathbf{u}_j)$. Then, we can further compute gradients of the objective function $\mathcal{J}$ with respect to $\mathbf{W}_s$, $\mathbf{v}_s$, $\mathbf{W}_h$, $\mathbf{v}_h$, and θ by utilizing chain rule

$$\frac{\partial\mathcal{J}}{\partial\mathbf{W}_s} = \frac{\partial\mathcal{J}}{\partial\mathbf{t}_i}\frac{\partial\mathbf{t}_i}{\partial\mathbf{o}_i}\frac{\partial\mathbf{o}_i}{\partial\mathbf{W}_s} = \frac{\partial\mathcal{J}}{\partial\mathbf{t}_i}\odot\mathbf{t}_i\odot(\mathbf{y}_i - \mathbf{t}_i)\mathbf{u}_i^T \tag{13}$$

$$\frac{\partial\mathcal{J}}{\partial\mathbf{v}_s} = \frac{\partial\mathcal{J}}{\partial\mathbf{t}_i}\frac{\partial\mathbf{t}_i}{\partial\mathbf{o}_i}\frac{\partial\mathbf{o}_i}{\partial\mathbf{v}_s} = \frac{\partial\mathcal{J}}{\partial\mathbf{t}_i}\odot\mathbf{t}_i\odot(\mathbf{y}_i - \mathbf{t}_i) \tag{14}$$

$$\frac{\partial\mathcal{J}}{\partial\mathbf{W}_h} = \frac{\partial\mathcal{J}}{\partial\mathbf{u}_i}\frac{\partial\mathbf{u}_i}{\partial\mathbf{W}_h} = \frac{\partial\mathcal{J}}{\partial\mathbf{u}_i}\mathbf{f}_i^T \tag{15}$$

$$\frac{\partial\mathcal{J}}{\partial\mathbf{v}_h} = \frac{\partial\mathcal{J}}{\partial\mathbf{u}_i}\frac{\partial\mathbf{u}_i}{\partial\mathbf{v}_h} = \frac{\partial\mathcal{J}}{\partial\mathbf{u}_i} \tag{16}$$

$$\frac{\partial\mathcal{J}}{\partial\Phi(\mathbf{x}_i;\theta)} = \frac{\partial\mathcal{J}}{\partial\mathbf{u}_i}\frac{\partial\mathbf{u}_i}{\partial\Phi(\mathbf{x}_i;\theta)} = \mathbf{W}_h^T\frac{\partial\mathcal{J}}{\partial\mathbf{u}_i} \tag{17}$$

where $\mathbf{o}_i = \mathbf{W}_s\mathbf{u}_i + \mathbf{v}_s$, and the operation $\odot$ denotes element-wise multiplication. Finally, we can update all parameters by using the gradient descent method as follows:

$$\xi = \xi - \mu\frac{\partial\mathcal{J}}{\partial\xi}, \quad \xi = \mathbf{W}_s, \mathbf{W}_h, \mathbf{v}_s, \mathbf{v}_h, \Phi(\mathbf{x}_i;\theta) \tag{18}$$

where μ is the learning rate.

D. Retrieval and Classification

Once the DHCNN is trained enough, we can obtain the hash codes and class labels for all samples from the database. Especially, for an arbitrary image $\mathbf{x}_q$, its hash code $\mathbf{b}_q$ and class label c_q can be determined by

$$\mathbf{b}_q = \text{sgn}(\mathbf{u}_q) = \text{sgn}(\mathbf{W}_h\mathbf{f}_q + \mathbf{v}_h) \tag{19}$$

$$c_q = \arg\max_{k=1,2,\ldots,C}\mathbf{t}_i^k \tag{20}$$

where $\mathbf{t}_i^k$ is the kth component of vector $\mathbf{t}_i$. Finally, given a query image, a set of similar images and their class labels can be easily returned via ranking Hamming distance between the query image and images from the database.

IV. Experimental Results and Analysis

To validate the effectiveness of the proposed method for RSIR and classification, comprehensive experiments are conducted on three widely used remote sensing image data sets.

A. Data Sets and Experimental Settings

1) The University of California, Merced Data (UCMD) [49] was manually extracted from large images downloaded from the United States Geological Survey (USGS). The UCMD is widely used for evaluating the performance of retrieval and classification for remote sensing images. It contains 21 land cover categories, each category includes 100 images of 256 × 256 pixels, and the spatial resolution of each pixel is 0.3 m. Some classes in UCMD are highly overlapping, e.g., medium residential and dense residential, which makes it a challenging data set. The detailed class information is shown in the left column of Table II.
2) WHU-RS data set [50] is a remote sensing scene data set that was collected from Google Earth. The images are divided into 19 classes; each class has approximately 50 images with 600 × 600 pixels. In this data set, the differences in the resolution, scale, and orientation in some scene images make it more complicated than those in UCMD. The middle column of Table II shows the detailed class information of this data set.
3) The Aerial Image Data (AID) [51] is a large-scale remote sensing image data set that was collected with the goal of advancing the state of the art in scene classification of remote sensing images. The data set has a number of 10 000 images within 30 classes. Each class consists of 220–420 images of size of 600 × 600 pixels. The images in AID are from different remote imaging sensors and the spatial resolution varies greatly between around 0.5–8 m, which brings more challenges for scene classification and image retrieval than the single-source

TABLE II
CLASS INFORMATION OF UCMD, WHU-RS, AND AID

UCMD				WHU-RS				AID			
No.	Name	No.	Name	No.	Name	No.	Name	No.	Name	No.	Name
1	agricultural	16	parking lot	1	airport	16	railway station	1	airplane	16	mountain
2	airplane	17	river	2	beach	17	residential	2	bare land	17	park
3	baseball field	18	runway	3	bridge	18	river	3	baseball field	18	parking
4	beach	19	spa-residential	4	commercial	19	viaduct	4	beach	19	playground
5	buildings	20	storage tanks	5	desert			5	bridge	20	pond
6	chaparral	21	tennis court	6	farmland			6	center	21	port
7	den-residential			7	football field			7	church	22	railway station
8	forest			8	forest			8	commercial	23	resort
9	freeway			9	industrial			9	den-residential	24	river
10	golf course			10	meadow			10	desert	25	school
11	harbor			11	mountain			11	farmland	26	spa-residential
12	intersection			12	park			12	forest	27	square
13	med-residential			13	parking lot			13	industrial	28	stadium
14	home park			14	pond			14	meadow	29	storage tanks
15	overpass			15	port			15	med-residential	30	viaduct

images, such as UCMD. The class information of this data set is given in the right column of Table II.

We systematically compare our method with some state-of-the-art hashing methods, including traditional and deep hashing methods. For traditional approaches, we compare with SH [37], ITQ [38], DSH [39], and SELVE [42] from unsupervised hashing and KSH [43] from supervised hashing. The deep hashing methods include DPSH [45], DHNNs with L2 regularization (DHNNs-L2) [22], and asymmetric deep supervised hashing (ADSH) [52]. For traditional methods, we utilize the 4096-D CNN feature extracted from the fc7 layer of VGG-F to represent each remote sensing image. For the deep hashing methods, we first resize all images tobe 224×224 pixels and then directly feed the raw image pixels into deep networks. It is worth mentioning that we adopt the same network architecture (i.e., VGG-F) for DPSH, DHNNs-L2, and our proposed DHCNN to achieve fair comparison. For SELVE, the dimension of the sparse embedding vector, the number of nearest anchors, and the balanced parameter λ are set to be 300, 10, and 0.1, respectively. For DSH, three parameters α, p, and, r are set to be 3, 1.5, and 3, respectively, where α controls the group number, p is the number of iterations in the k-means, and r is parameter for r-adjacent group. For KSH, the Gaussian RBF kernel $k(x, y) = \exp(-\|\text{x-y}\|/2\sigma^2)$ is exploited. For ADSH, the iteration numbers T_{out} and T_{in} are set to be 50 and 3, respectively, and the hyperparameter γ is set to be 200. For DPSH and DHNNs-L2, the regularization coefficient η is set to be 10. In addition, similarity weight ω is set to be 32 for DHNNs-L2. For the proposed DHCNN, the hyperparameters η and β are set to be 0.2 and 25, respectively. All experiments are performed on a computer equipped with an Intel Core i7-8700K with 3.7 GHz, 64-GB memory, and an Nvidia GeForce GTX 1080Ti. In experiments, we adopt four widely used metrics to evaluate the retrieval performance, i.e., mean average precision (MAP), precision@k, recall@k, and precision–recall. MAP is used to evaluate the overall retrieval performance. The rest of the three metrics, including precision@k, recall@k, and precision–recall, are used to compare the retrieval results of all methods in terms of precision-k, recall-k, and precision–recall curves, respectively, where k is a predefined number of the returned similar images.

B. Retrieval Results

Considering that DPSH, DHNNs-L2, and our proposed DHCNN have utilized the same framework to extract deep hash features, we first analyze the qualitative results for three methods on three data sets. Fig. 2 shows query examples with the top ten retrieved images on the UCMD, WHU-RS, and AID, respectively. For each query example, the top, middle, and bottom rows represent retrieval results obtained by DPSH, DHNNs-L2, and our proposed DHCNN, respectively. The green rectangle marks true positives, while the red rectangle marks false positives. The "TL" and "PL" represent the true label and predicted label of images, respectively. From this figure, we can see that the retrieval results obtained by DPSH and DHNNs-L2 are not good for some challenging classes that exhibit very high interclass similarity (e.g., medium residential and dense residential classes of UCMD, school, and dense residential classes of AID). In addition, for the query example on WHU-RS, the object (i.e., airplane) is very small compared with the background, which results in the unsatisfactory retrieval performance for DPSH and DHNNs-L2. By contrast, our method still returns all true positives in the top ten retrieved images, which exhibits a great advantage in coping with the problem of high interclass similarity and challenging resolution of the object in remote sensing images. More importantly, different from existing retrieval methods, our method can retrieve similar images and simultaneously classify their semantic labels in a unified framework. As shown in Fig. 2, our method can precisely predict the true label of retrieved images and achieve satisfactory classification performance.

We also report the quantitative comparison among all methods. Table III shows the image retrieval results in terms of MAP with different hash bits on the three data sets. From Table III, we can observe that, in most cases, the supervised methods outperform the unsupervised methods and the

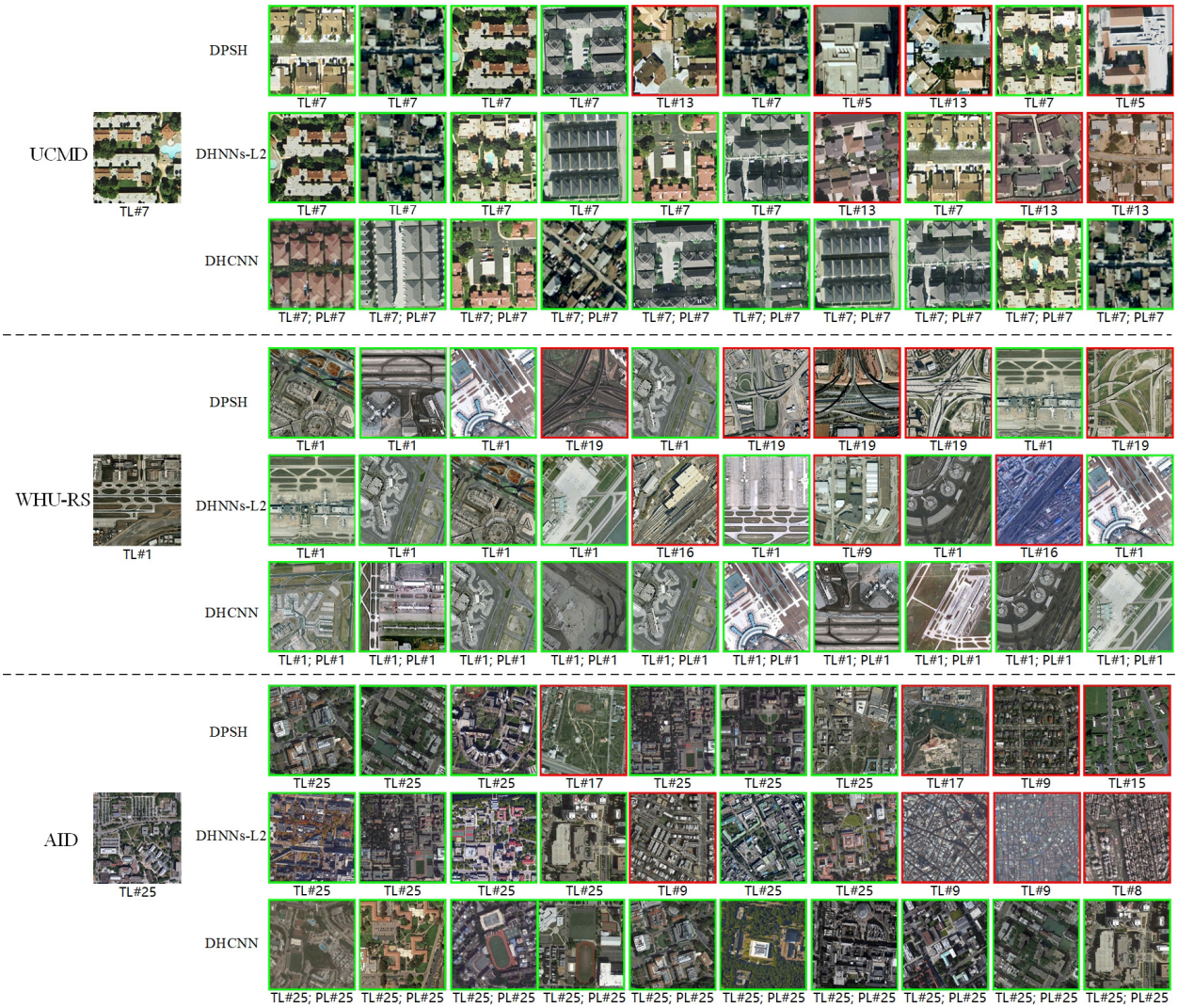

Fig. 2. Query examples with top ten retrieved images on three data sets. The first column is the query images from UCMD, WHU-RS, and AID, respectively. The rest of the ten columns refer to the searched images. For each query example, the top, middle, and bottom rows represent retrieval results obtained by DPSH [45], DHNNs-L2 [22], and our proposed method DHCNN. The green rectangle marks true positives, while the red rectangle marks false positives. The "TL" and "PL" represent the true label and predicted label of images, respectively.

TABLE III

IMAGE RETRIEVAL RESULTS IN TERMS OF MAP WITH 16, 32, AND 64 HASH BITS ON THE THREE DATA SETS. THE SCALES OF TEST QUERY SETS ARE 420 (20% SAMPLES PER CLASS), 510 (50% SAMPLES PER CLASS), AND 5000 (50% SAMPLES PER CLASS) FOR UCMD, WHU-RS, AND AID, RESPECTIVELY. THE MAPS ARE COMPUTED USING ALL TRAINING SETS, AND THE BEST VALUES ARE SHOWN IN BOLDFACE

Method	UCMD			WHU-RS			AID		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
DHCNN (Our method)	**0.9652**	**0.9698**	**0.9802**	**0.9210**	**0.9510**	0.9612	**0.8905**	**0.9297**	**0.9427**
ADSH [52]	0.9051	0.9559	0.9672	0.9134	0.9414	**0.9649**	0.7993	0.9172	0.9373
DHNNs-L2 [22]	0.8279	0.8746	0.9240	0.8372	0.8758	0.9123	0.7993	0.8238	0.8912
DPSH [45]	0.7660	0.8420	0.8975	0.6632	0.7347	0.8146	0.6832	0.7389	0.7768
KSH-CNN [43]	0.7550	0.8362	0.8722	0.6749	0.7316	0.7952	0.4826	0.5815	0.6326
ITQ-CNN [38]	0.4265	0.4563	0.4764	0.4733	0.5385	0.5491	0.2335	0.2731	0.2999
SELVE-CNN [42]	0.3612	0.4036	0.3858	0.3988	0.4709	0.4730	0.3458	0.3787	0.3681
DSH-CNN [39]	0.2882	0.3307	0.3459	0.3092	0.3987	0.4020	0.1605	0.1808	0.1972
SH-CNN [37]	0.2952	0.3008	0.2931	0.3477	0.3497	0.3486	0.1269	0.1699	0.1621

deep hashing methods further exceed the traditional hashing methods. Comparing with DPSH and DHNNs-L2, our method achieves great improvements by combining similarity information between image pairs with semantic information of each image to guide network training. Although our method does not obtain the best result obtained by ADSH on the

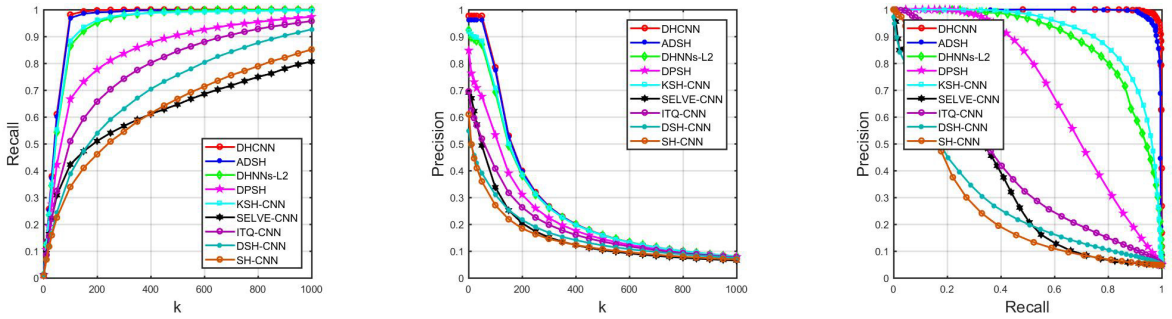

Fig. 3. Retrieval results on UCMD with 64-bit hash code. (a) Recall@k. (b) Precision@k. (c) Precision–recall.

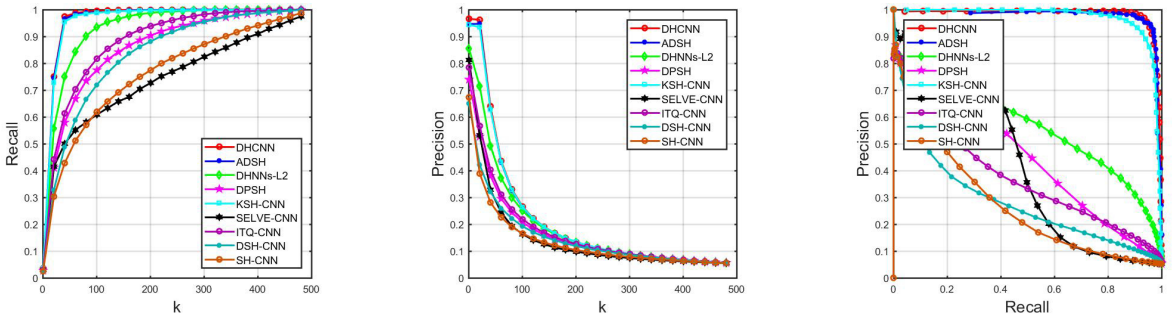

Fig. 4. Retrieval results on WHU-RS with 64-bit hash code. (a) Recall@k. (b) Precision@k. (c) Precision–recall.

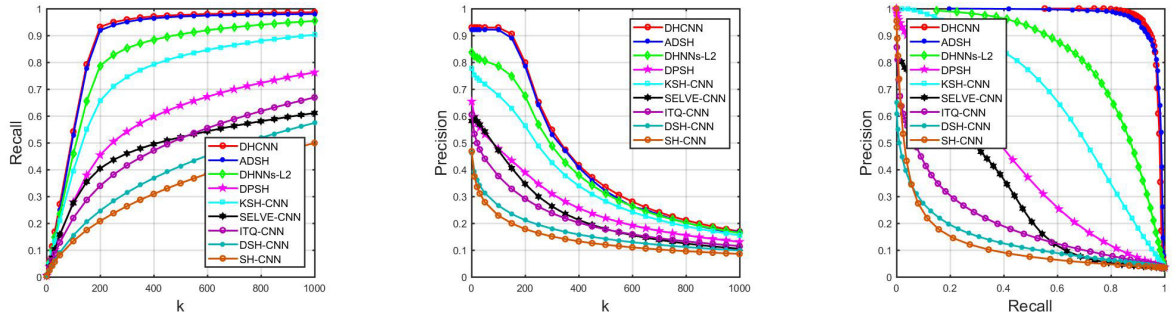

Fig. 5. Retrieval results on AID with 64-bit hash code. (a) Recall@k. (b) Precision@k. (c) Precision–recall.

WHU-RS data set when the hash bit is set to 64, we can still see that our result is very close to the best result, i.e., the MAP value of our method is only 0.0037 lower than the value of ADSH. In addition, our method is less sensitive than other compared approaches to the hash bits. For example, when the hash bits change from 64 to 16, the MAP values on AID decrease 0.0522 and 0.1380 for our method and ADSH, respectively. The abovementioned experimental results validate the effectiveness of our method for RSIR.

Finally, we further compare the retrieval results of different approaches in terms of recall@k, precision@k, and precision–recall metrics. Figs. 3–5 show the corresponding retrieval results on three data sets, where the hash bit is set to be 64 for all methods. From Figs. 3–5, we can see that our method outperforms other compared methods in most cases for recall@k and precision@k metrics. At the same time, the precision–recall curve also shows the great advantages of our method over other approaches on the three data sets.

C. Effects of Different Training Samples on MAP

In this section, we conduct experiments to analyze the effect of a different number of training samples on retrieval results under 64-bit hash bits. Here, we only compare the proposed method with DHNNs-L2 and ADSH, and the rest of the compared approaches are excluded due to the poor retrieval results obtained by these methods. The ratio between training and all samples are set to be 0.2, 0.4, 0.6, and 0.8 for three data sets, respectively. The rest of the samples are regarded as the test set to evaluate the retrieval performance.

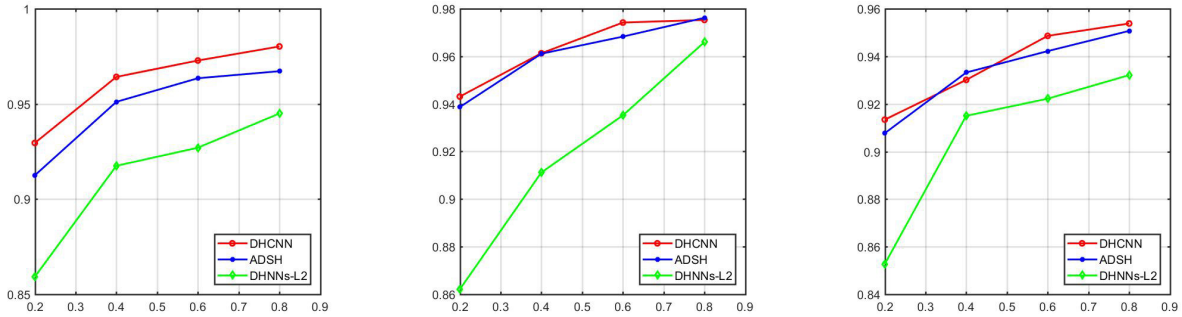

Fig. 6. Retrieval results under different train ratio on (a) UCMD, (b) WUH-RS, and (c) AID.

TABLE IV
COMPARISON OF COMPUTATIONAL TIME (IN SECONDS) FOR THE PROPOSED DHCNN, DHNNS-L2 [22], AND DPSH [45] ON UCMD AND WHU-RS

Method	UCMD		WHU-RS	
	Training	Test	Training	Test
DHCNN (our method)	663.89	2.56	203.92	1.12
DHNNs-L2 [22]	643.95	3.01	196.58	1.47
DPSH [45]	647.20	3.05	196.88	1.51

TABLE V
CLASSIFICATION RESULTS IN TERMS OF OA (SHOWN IN PERCENTAGES) ON THREE DATA SETS. THE RATIO OF TRAINING AND ALL SAMPLES IS SET TO 80%, 50%, AND 50% PER CLASS FOR UCMD, WHU-RS, AND AID, RESPECTIVELY. THE BEST VALUES ARE SHOWN IN BOLDFACE

Method	UCMD	WHU-RS	AID
DHCNN (Our method)	**97.68**	**96.22**	**93.48**
SPP-net+MKL [53]	95.72	94.84	92.91
MSP-FV-VGG-F [54]	96.90	-	93.20
DCA-Fusion [55]	95.84	95.56	91.87
GBRCN [56]	95.53	91.34	91.40
CaffeNet [51]	95.02	95.62	89.53
VGG-VD16 [51]	95.21	94.12	89.64
GoogLeNet [51]	94.32	95.74	86.39

The retrieval results in terms of MAP are shown in Fig. 6. From this figure, we can see that DHNNs-L2 is very sensitive to the number of training samples. When a small number of training samples are available, the retrieval results of DHNNs-L2 on three data sets dramatically decrease. By contrast, the MAP values of DHCNN and ADSH steadily rise with the number of training samples increasing. In addition, we also find that DHCNN exhibits the obvious advantage over ADSH on the UCMD. Especially, the MAP value of DHCNN is about 0.15 higher than that of ADSH for all separation scenarios of training and test sets. Although the above advantage is not distinct on WHU-RS and AID, we can still see that DHCNN delivers the highest MAP values for the majority of separation scenarios. Based on the above analyses, we can conclude that the proposed DHCNN can achieve satisfactory retrieval performance under different separation scenarios of training and test sets and, at the same time, exhibit advantage over other compared methods to some extent.

D. Computational Time

In this section, we investigate the efficiency of the proposed method. DPSH and DHNNs-L2 are selected as the compared approaches. The experiments are conducted on the UCMD and WHU-RS data set. We randomly select 80% and 50% of samples per class as the training set for UCMD and WHU-RS, respectively, and the rest of the samples are used as a test set. Table IV shows the computational time, including training time and test time, on the two data sets when the hash bit is set to be 64. From this table, we can easily find that the training and test time of the three methods are roughly approximate. Furthermore, we can also observe that the training time of our method is slightly higher than the ones of the other two methods. The main reason is that our method has an additional semantic layer after the hash layer to consider the semantic information of each image, which brings in more network parameters. By contrast, our method is the most efficient method in terms of test time.

E. Classification Results

As one of the advantages over existing retrieval approaches, our method can precisely classify the semantic labels of the returned similar images. In other words, the retrieval and classification tasks can be simultaneously achieved in a unified framework. In this section, we conduct experiments on the UCMD, WHU-RS, and AID to validate the effectiveness of our method for remote sensing image classification. We randomly select 80%, 50%, and 50% of samples per class as the training set, and the rest of the samples are used to evaluate the classification performance. The hash bit is set to be 64 for all data sets. We adopt overall accuracy (OA) as metrics to evaluate the classification performance.

In this experiment, we compare our method with some state-of-the-art methods for remote sensing image classification, including gradient boosting random convolutional network (GBRCN) [56], deep feature fusion based on discriminant correlation analysis (DCA-Fusion) [55], multiscale pooling with FV (MSP-FV) method [54], spatial pyramid pooling-net with multiple kernel learning (SPP-net + MKL) [53], and some deep feature-based methods that extract the activations from the first fully connected layer of CaffeNet [26],

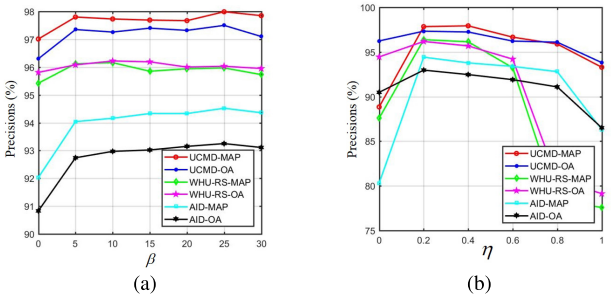

Fig. 7. Effects of different parameters on the retrieval and classification accuracies on three data sets. (a) β. (b) η.

GoogLeNet [27], and VGG-VD16 [28]. Table V shows the classification results obtained by all methods. Note that the results of all compared methods are from the corresponding references and review articles [51], [57], where the same number of training samples is used. From this table, we can see that our proposed method obtains the highest classification accuracies on UCMD and AID. Although the result of MSP-FV-VGG-F on WHU-RS is not reported due to the unavailable source code, we can still see that our method significantly outperforms other compared approaches. In addition, we also observe that the GoogLeNet performs worse than CaffeNet and VGG-VD16, which is inconsistent with the image classification of natural images.

F. Parameters Analysis

In our object function, parameter β constrains hash-like feature $\mathbf{u}_i$ approach to hash code $\mathbf{b}_i$, and parameter η balances the semantic information and similarity information. In this section, the effects of β and η on retrieval and classification performance are analyzed.

Fig. 7(a) shows the effects of β on MAP and OA on the thee remote sensing data sets, where η is set to be 0.2. From Fig. 7(a), we can see that the retrieval and classification accuracies are lowest when β equals to 0. The main reason is that the learned hash codes lack compactness since the network cannot effectively constrain the hash-like feature approach to hash code when β equals 0. The best retrieval and classification results are achieved when β equals 25 for both the UCMD and AID and 10 for WHU-RS, respectively. In addition, Fig. 7(b) shows the effects of η on MAP and OA on the three data sets, where β is set to be 25 for UCMD and AID and 10 for WHU-RS, respectively. From this figure, we can see that, when η approaches 0 (i.e., only consider label loss of each image) or 1 (i.e., only consider similarity loss of between images), the MAP and OA significantly decrease for all data sets. The MAP and OA obtain the highest values when η achieves 0.2 for all data sets. The above experimental results also validate the superiority of effectively combining the semantic information of each image and similarity information between images.

V. Conclusion

In this article, we redefine the image retrieval problem as visual and semantic retrieval of images. Especially, given a query image, a set of similar images and their semantic labels can be simultaneously obtained via a united framework. To this end, a novel DHCNN is proposed to learn compact hash codes for efficient RSIR and discriminative features for accurate semantic label classification. In more detail, we first introduce a pretrained CNN to extract high-dimensional deep features from raw remote sensing images. Then, a hash layer is perfectly inserted into CNN to learn low-dimensional hash codes. In addition, a fully connected layer with a softmax function is performed on the hash layer to generate the probability distribution of each class. Finally, a loss function that simultaneously considers the semantic information and similarity information is elaborately designed to train DHCNN. The experimental results on three remote sensing data sets demonstrate that the proposed method gives excellent results as it achieves state-of-art retrieval and classification performance.

References

[1] X. Tang, L. Jiao, W. J. Emery, F. Liu, and D. Zhang, "Two-stage reranking for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5798–5817, Oct. 2017.

[2] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[3] T. Bretschneider, R. Cavet, and O. Kao, "Retrieval of remotely sensed imagery using spectral information content," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Dec. 2002, pp. 2253–2255.

[4] G.-S. Xia, J. Delon, and Y. Gousseau, "Shape-based invariant texture indexing," *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 382–403, Jul. 2010.

[5] G. J. Scott, M. N. Klaric, C. H. Davis, and C.-R. Shyu, "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1603–1616, May 2011.

[6] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.

[7] P. Napoletano, "Visual descriptors for content-based retrieval of remote-sensing images," *Int. J. Remote Sens.*, vol. 39, no. 5, pp. 1343–1376, Mar. 2018.

[8] S. Ozkan, T. Ates, E. Tola, M. Soysal, and E. Esen, "Performance analysis of State-of-the-art representation methods for geographical image retrieval and categorization," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 11, pp. 1996–2000, Nov. 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[11] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[12] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. Atli Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[13] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.

[14] L. Fang, Z. Liu, and W. Song, "Deep hashing neural networks for hyperspectral image feature extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1412–1416, Sep. 2019.

[15] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[16] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 1793–1802, Dec. 2016.

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

[18] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," *IEEE Trans. Big Data*, vol. 6, no. 3, pp. 507–521, Sep. 2020.

[19] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 5, p. 489, May 2017.

[20] F. Hu, X. Tong, G.-S. Xia, and L. Zhang, "Delving into deep representations for remote sensing image retrieval," in *Proc. IEEE 13th Int. Conf. Signal Process. (ICSP)*, Nov. 2016, pp. 198–203.

[21] 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.

[22] 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. 2018.

[23] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6521–6536, Nov. 2018.

[24] S. Roy, E. Sangineto, B. Demir, and N. Sebe, "Deep metric and hash-code learning for content-based retrieval of remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 4539–4542.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[26] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[27] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

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

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[31] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–51.

[32] X. Lu, Y. Chen, and X. Li, "Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 106–120, Jan. 2018.

[33] X. Lu, X. Zheng, and X. Li, "Latent semantic minimal hashing for image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 355–368, Jan. 2017.

[34] J. Gui, T. Liu, Z. Sun, D. Tao, and T. Tan, "Supervised discrete hashing with relaxation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 608–617, Mar. 2018.

[35] X. Bai, H. Yang, J. Zhou, P. Ren, and J. Cheng, "Data-dependent hashing based on p-stable distribution," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5033–5046, Dec. 2014.

[36] X. Bai, C. Yan, H. Yang, L. Bai, J. Zhou, and E. R. Hancock, "Adaptive hash retrieval with kernel based similarity," *Pattern Recognit.*, vol. 75, pp. 136–148, Mar. 2018.

[37] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1753–1760.

[38] 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. 2013.

[39] Z. Jin, C. Li, Y. Lin, and D. Cai, "Density sensitive hashing," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1362–1371, Aug. 2014.

[40] B. Kulis and T. Darrell, "Learning to hash with binary reconstructive embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1042–1050.

[41] M. Norouzi and D. M. Blei, "Minimal loss hashing for compact binary codes," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 353–360.

[42] X. Zhu, L. Zhang, and Z. Huang, "A sparse embedding and least variance encoding approach to hashing," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3737–3750, Sep. 2014.

[43] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, "Supervised hashing with kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2074–2081.

[44] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proc. AAAI Conf. Arti. Intell.*, 2014, p. 2.

[45] W. Li, S. Wang, and W. Kang, "Feature learning based deep supervised hashing with pairwise labels," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 1711–1717.

[46] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[47] H. Zhang, L. Liu, Y. Long, and L. Shao, "Unsupervised deep hashing with pseudo labels for scalable image retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1626–1638, Apr. 2018.

[48] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*. [Online]. Available: http://arxiv.org/abs/1405.3531

[49] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.

[50] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *Proc. ISPRS*, vol. 38, 2010, pp. 298–303.

[51] 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.

[52] Q. Jiang and W. Li, "Asymmetric deep supervised hashing," in *Proc. AAAI Conf. Arti. Intell.*, 2018, pp. 1–8.

[53] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 117–126, Jan. 2018.

[54] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.

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

[56] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[57] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Robust space–frequency joint representation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7492–7502, Oct. 2019.

Weiwei Song (Student Member, IEEE) received the B.S. degree from Southwest Minzu University, Chengdu, China, in 2015. He is pursuing the Ph.D. degree with the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Changsha, China.

From November 2018 to November 2019, he was a Visiting Ph.D. Student with the Department of Electrical and Computer Engineering, University of Iceland, Reykjavik, Iceland, supported by the China Scholarship Council. His research interests include hyperspectral image classification, remote sensing image retrieval, and spare representation.

Shutao Li (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees from Hunan University, Changsha, China, in 1995, 1997, and 2001, respectively.

In 2001, he joined the College of Electrical and Information Engineering, Hunan University, where he is a Full Professor. From May 2001 to October 2001, he was a Research Associate with the Department of Computer Science, The Hong Kong University of Science and Technology, Hong Kong. From November 2002 to November 2003, he was a Post-Doctoral Fellow with Royal Holloway College, University of London, London, U.K. From April 2005 to June 2005, he was a Visiting Professor with the Department of Computer Science, The Hong Kong University of Science and Technology. He has authored or coauthored over 200 refereed articles. His research interests include image processing, pattern recognition, and artificial intelligence.

Dr. Li is a member of the Editorial Board of Information Fusion and Sensing and Imaging. He received two Second-Grade State Scientific and Technological Progress Awards of China, in 2004 and 2006. He is also an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT.

Jón Atli Benediktsson (Fellow, IEEE) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

From 2009 to 2015, he was the Prorector of Science and Academic Affairs and a Professor of Electrical and Computer Engineering with the University of Iceland. In 2015, he was the Rector with the University of Iceland. He is the Co-Founder of Oxymap, Reykjavik, a biomedical start-up company. He has authored or coauthored extensively in his fields of interest. His research interests include remote sensing, image analysis, pattern recognition, biomedical analysis of signals, and signal processing.

Dr. Benediktsson is a Fellow of the International Society for Optics and Photonics (SPIE) and a member of the Association of Chartered Engineers in Iceland (VFI), the Societas Scientiarum Islandica, and the Tau Beta Pi. He was a recipient of the Stevan J. Kristof Award from Purdue University in 1991 as an Outstanding Graduate Student in Remote Sensing, the Icelandic Research Councils Outstanding Young Researcher Award in 1997, the IEEE Third Millennium Medal in 2000, the Yearly Research Award from the Engineering Research Institute, University of Iceland, in 2006, the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society (GRSS) in 2007, and the IEEE/VFI Electrical Engineer of the Year Award in 2013. He was a co-recipient of the University of Iceland's Technology Innovation Award in 2004, the 2012 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) Paper Award, the IEEE GRSS Highest Impact Paper Award in 2013, and the *International Journal of Image and Data Fusion* Best Paper Award in 2014. He was the 2011–2012 President of GRSS. He has been with the GRSS Administrative Committee since 2000. He was the Chairman of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2007 to 2010. He was the Editor-in-Chief of the IEEE TGRS from 2003 to 2008. He has been an Associate Editor of the IEEE TGRS since 1999, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003, and the IEEE ACCESS since 2013. He serves on the Editorial Board of the PROCEEDINGS OF THE IEEE and the International Editorial Board of the *International Journal of Image and Data Fusion*.