Large-Scale Remote Sensing Image Retrieval by Deep Hashing Neural Networks

Yansheng Li, Yongjun Zhang, Xin Huang, *Senior Member, IEEE*, Hu Zhu, and Jiayi Ma

***Abstract*— As one of the most challenging tasks of remote sensing big data mining, large-scale remote sensing image retrieval has attracted increasing attention from researchers. Existing large-scale remote sensing image retrieval approaches are generally implemented by using hashing learning methods, which take handcrafted features as inputs and map the high-dimensional feature vector to the low-dimensional binary feature vector to reduce feature-searching complexity levels. As a means of applying the merits of deep learning, this paper proposes a novel large-scale remote sensing image retrieval approach based on deep hashing neural networks (DHNNs). More specifically, DHNNs are composed of deep feature learning neural networks and hashing learning neural networks and can be optimized in an end-to-end manner. Rather than requiring to dedicate expertise and effort to the design of feature descriptors, we can automatically learn good feature extraction operations and feature hashing mapping under the supervision of labeled samples. To broaden the application field, DHNNs are evaluated under two representative remote sensing cases: scarce and sufficient labeled samples. To make up for a lack of labeled samples, DHNNs can be trained via transfer learning for the former case. For the latter case, DHNNs can be trained via supervised learning from scratch with the aid of a vast number of labeled samples. Extensive experiments on one public remote sensing image data set with a limited number of labeled samples and on another public data set with plenty of labeled samples show that the proposed remote sensing image retrieval approach based on DHNNs can remarkably outperform state-of-the-art methods under both of the examined conditions.**

***Index Terms*— Deep hashing neural networks (DHNNs), large-scale remote sensing image retrieval, remote sensing big data (RSBD) mining, supervised learning from scratch, transfer learning.**

Manuscript received June 2, 2017; revised August 25, 2017; accepted September 19, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 41601352, Grant 41322010, Grant 41522110, and Grant 61503288, in part by the China Postdoctoral Science Foundation under Grant 2016M590716 and Grant 2017T100581, and in part by the Fundamental Research Funds for the Central Universities under Grant 2042016KF0054. *(Corresponding author: Yongjun Zhang).*

Y. Li and Y. Zhang are with the Department of Photogrammetry, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: yansheng.li@whu.edu.cn; zhangyj@whu.edu.cn).

X. Huang is with the Department of Remote Sensing, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: xhuang@whu.edu.cn).

H. Zhu is with the Department of Radio and Television Engineering, College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: peter.hu.zhu@gmail.com).

J. Ma is with the Department of Communication Engineering, Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jiayima@whu.edu.cn).

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

Digital Object Identifier 10.1109/TGRS.2017.2756911

I. Introduction

WITH the rapid development of remote sensing observation technologies, we have entered an era of remote sensing big data (RSBD) [1]–[3]. There is no doubt that RSBD contain invaluable information. Due to the large volume of RSBD, manual information extraction from RSBD is time consuming and prohibitive. Hence, useful information must be automatically drawn from RSBD. Driven by the demand from multiple fields (e.g., disaster rescue), automatic knowledge discovery from RSBD has become increasingly urgent. Among emerging RSBD mining efforts [1], content-based large-scale remote sensing image retrieval [4]–[8] has attracted an increasing amount of research interest due to its broad applications.

In earlier remote sensing image retrieval systems, remote sensing image retrieval mainly relied on manual tags in terms of sensor types, waveband information, and geographical locations of remote sensing images. As a consequence, the retrieval performance of these systems was highly dependent on the availability and quality of manual tags. However, the manual generation of tags is often time consuming and becomes especially prohibitive when the volume of remote sensing images increases considerably. In fact, recent efforts show that the visual contents of remote sensing images themselves are more relevant than manual tags [9]. Hence, researchers have begun to exploit ways to search through similar remote sensing images in terms of visual content. Specifically, Wang and Song [10] used the spatial relationships of classification results to measure similarities between two remote sensing images. With this approach, however, image retrieval performance is highly dependent on classification accuracy levels. To avoid this dependence, numerous feature descriptors have been specifically designed for indexing remote sensing images. More specifically, local invariant [11], morphological [12], textural [13]–[16], and data-driven features [17]–[19] have been evaluated in terms of content-based remote sensing image retrieval tasks. To further improve image retrieval performance levels, we have proposed a multiple feature-based remote sensing image retrieval approach [20] that not only considers handcrafted features but also utilizes data-driven features via unsupervised feature learning [21]. In addition, Wang *et al.* [22] proposed a multilayered graph model for hierarchically refining retrieval results from coarse to fine. For the aforementioned methods, the visual contents of remote sensing images are often represented by thousands of dimensional feature descriptors. Exhaustively comparing the high-dimensional feature descriptor of an inquiry remote sensing image with each image in a data set is computationally expensive and impossible to achieve when the volume of a data set is oversized.

To address the aforementioned problems with exhaustive high-dimensional feature searching, two strategies may be employed: improving search methods and reducing the dimensions of feature descriptors. The former strategy is implemented by using data partition algorithms that recursively split data spaces into subspaces and record these divisions via a tree structure. In benefiting from this data partitioning strategy, the search speed of tree-based methods [4]–[6] is significantly improved, but retrieval performance levels decrease dramatically, especially when the dimension of the original feature descriptor is very high [23]. Unfortunately, the dimensions of feature descriptors of remote sensing images are often very high. To avoid this issue, several researchers have exploited feature reduction methods for large-scale remote sensing image retrieval. Recently, hashing learning methods [7], [8] have been introduced into large-scale remote sensing image retrieval tasks. These hashing learning methods take handcrafted feature descriptors with dimensions that are often very high as an input and map high-dimensional feature vectors (HDFVs) to low-dimensional binary feature vectors (LDBFVs). Accordingly, the complexity of exhaustive searches using LDBFV is dramatically reduced relative to that of HDFV. Although existing hashing learning methods can significantly increase search speeds, retrieval accuracy levels still fail to meet the demands of practical applications. In view of the great successes of deep learning methods [24]–[26] in recently developed applications, replacing low-level handcrafted features of hashing learning methods [7], [8] with high-level semantic features of deep learning can further improve retrieval performance levels. To fully employ the respective merits of deep and hashing learning, deep hashing neural networks (DHNNs) [27]–[29] have been proposed by pioneers of the computer vision community, and exciting results of large-scale natural image retrieval tasks have been retrieved. Generally, remote sensing images differ considerably from natural images in both spectral and spatial domains. Due to this substantial gap, DHNNs trained in a natural image data set cannot be applied directly to large-scale remote sensing image retrieval tasks. Hence, the modeling and learning of DHNNs based on specific remote sensing image retrieval tasks deserve more exploration.

Based on the aforementioned considerations, this paper proposes a novel large-scale remote sensing image retrieval approach based on DHNNs. More specifically, this paper presents a comprehensive study of DHNNs and introduces DHNNs into large-scale remote sensing image retrieval tasks. To clarify fundamental theories of DHNNs, this paper provides a systematic review of existing DHNNs. Different from existing DHNNs studies [27]–[29], this paper for the first time illustrates the importance of the similarity weight and quantization loss function of DHNNs. To cover as many cases as possible, DHNNs are utilized in two remote sensing situations: remote sensing data sets with limited and sufficient quantities of labeled samples. For the former case, the deep feature learning module of DHNNs can be derived from suitable pretrained neural networks, and the hashing learning module of DHNNs is randomly initialized; then, DHNNs can be incrementally trained using the limited number of labeled samples available. For the latter case, DHNNs can be randomly constructed based on the specific data characteristics of remote sensing images and then trained from scratch using a sufficient number of labeled samples. Compared to existing hashing learning methods [7], [8] that have been applied to large-scale remote sensing image retrieval, some recently presented hashing learning methods [30], [31], and three existing DHNN methods [27]–[29], the DHNNs proposed in this paper can achieve significant performance improvements when applied to two public remote sensing image data sets, where one includes a limited number of labeled samples and the other contains a sufficient number of labeled samples. As a whole, the main contributions of this paper are twofold.

1) From a methodological perspective, this paper provides a systematic review of DHNNs and illustrates the importance of critical components of DHNNs that are disregarded in existing DHNNs.
2) In terms of applications, for the first time, DHNNs are employed for large-scale remote sensing image retrieval. To cover as many remote sensing applications as possible, this paper illustrates ways to design and train DHNNs for large-scale remote sensing image retrieval when labeled samples are scarce and sufficient.

This paper is organized as follows. A comprehensive review of DHNNs is given in Section II, where we also list key parameters of DHNNs that can significantly affect performance outcomes. In Section III, we introduce solutions for designing and training DHNNs for large-scale remote sensing image retrieval in cases involving scarce and sufficient numbers of labeled samples. Using two public remote sensing image data sets, the overall performance of the proposed approach based on DHNNs and comparisons with state-of-the-art approaches are reported in Section IV. Finally, Section V presents the conclusion.

II. Deep Hashing Neural Networks

In the last decade, deep learning [24]–[26] has achieved considerable success when applied to nearly all computer vision tasks due to its superiority in terms of feature representation. In the remote sensing community, deep learning methods have been successively utilized for remote sensing image scene classification [32]–[35], hyper-spectral image classification [36]–[38], SAR image classification [39], [40], remote sensing image object recognition [41], [42], and so forth. Generally, the dimension of the feature vector output generated by these deep learning methods [32]–[42] is often very high and may be acceptable for these processing tasks. However, large-scale image retrieval based on HDFVs is impossible, as noted above.

In tailoring deep learning techniques to large-scale image retrieval, DHNNs have been proposed in [27]–[29]. More specifically, DHNNs are composed of deep feature learning neural networks (DFLNNs) for high-level semantic feature representation and of hashing learning neural networks (HLNNs) for compact feature representation, and can

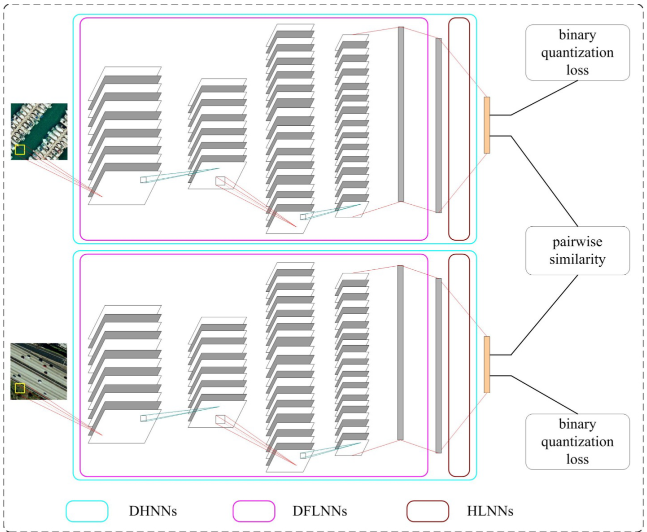


Fig. 1. Visualization of DHNNs and corresponding learning constraints. Subcomponents of DHNNs, including DFLNNs and HLNNs, are also shown.

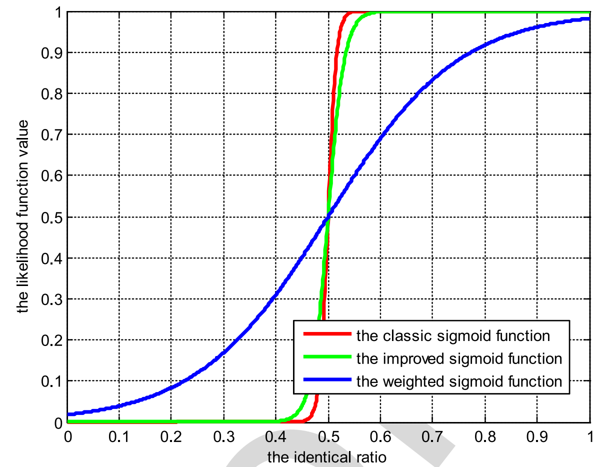


Fig. 2. Visual comparison of different sigmoid functions. In the visual comparison, the length of the binary feature is set to 64, and the similarity factor is set to 0.25. In addition, the identical ratio is calculated by dividing the number of identical bits between two binary features by the length of the binary feature.

be jointly optimized in an end-to-end manner. We note that joint optimization benefits render the feature representation and hashing mapping modules simultaneously optimal for a specific task.

To clearly describe the features of DHNNs, model formulations and learning paradigms for DHNNs are introduced in Sections II-A and II-B.

A. Modeling of DHNNs

Based on existing approaches [27]–[29], DHNNs can be represented by the integration of DFLNNs and HLNNs. More specifically, DFLNNs are composed of multiple convolutional and fully connected layers and pursue the high-level semantic feature representation of an input image scene. In addition, HLNNs can be constructed from one fully connected layer and aim at mapping the high-dimensional feature representation of DFLNNs for compact feature representation (i.e., the LDBFV). Unlike the high-dimensional feature representation of DFLNNs, the feature representation of DHNNs is extremely compact and can be applied to large-scale image retrieval tasks.

As depicted in Fig. 1, each image shares the same neural networks (i.e., DHNNs) throughout the compact feature representation process, and DHNNs can be optimized under constraints such as binary quantization loss and pairwise similarity constraints. More specifically, the binary quantization loss can render each element of the final feature representation of the DHNNs approach as -1 or 1, and the pairwise similarity constraint can cause similarities between feature representations of DHNNs to agree with real similarities based on manual labels of image scenes.

For an image data set $\{(I_i, y_i)|i = 1, 2, \ldots, N\}$, where I_i denotes the image and y_i denotes its label, the similarity matrix $\Theta \in R^{2\times N\times N}$ for the given image data set is specifically defined as $\Theta^1_{i,j} + \Theta^2_{i,j} = 1$, where $\Theta^1_{i,j} = 1$, if $y_i = y_j$ and $\Theta^1_{i,j} = 0$, if $y_i \neq y_j$.

Assuming that low-dimensional binary vectors of the image data set $\mathbf{I} = \{I_i\}_{i=1}^N$ can be represented by $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^N$, where $\mathbf{b}_i = \{-1, 1\}^l$ and l denotes the length of the binary feature vector, the likelihood function of the pairwise similarity Θ can be defined as

$$\begin{cases} P\left(\Theta^1_{i,j} = 1|\mathbf{B}\right) = \sigma\left(\Omega_{i,j}\right) \\ P\left(\Theta^2_{i,j} = 1|\mathbf{B}\right) = 1 - \sigma\left(\Omega_{i,j}\right) \end{cases} \tag{1}$$

where $\Omega_{i,j} = \mathbf{b}_i^T \mathbf{b}_j$ and $\sigma(\Omega_{i,j}) = 1/(1+e^{-\Omega_{i,j}})$ is the classic sigmoid function that easily leads to a large saturation zone where its gradient is close to 0.

In the literature, the classic sigmoid function $\sigma(\Omega_{i,j}) = 1/(1 + e^{-\Omega_{i,j}})$ is adopted in [27], and the improved sigmoid function $\sigma(\Omega_{i,j}) = 1/(1 + e^{-\Omega_{i,j}/2})$ is utilized in [29]. However, both sigmoid functions adopted in [27] and [29] would result in the generation of large saturation zone, which hinders the updating of network parameters through backpropagation. To avoid this result, this paper proposes the use of a weighted sigmoid function $\sigma(\Omega_{i,j}) = 1/(1 + e^{-\Omega_{i,j}/w})$, where $w = s \cdot l$ is the similarity weight, s is the similarity factor, and l is the length of the binary feature $\mathbf{b}$. Fig. 2 intuitively shows why the proposed weighted sigmoid function can effectively decrease the saturation zone relative to the classic sigmoid function used in [27] and the improved sigmoid function used in [29]. For the case illustrated in Fig. 2, the classic and improved sigmoid functions should cause the objective optimization function used in (2) to enter the saturation zone when the identical ratio exceeds 0.6 or falls below 0.4. In contrast, the weighted sigmoid function can cause the objective optimization function to pursue a higher identical ratio when two remote sensing images share the same visual content and vice versa.

The ideal binary feature representations $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^N$ are unknown in advance. Under the similarity matrix constraint Θ, we can determine binary representations by minimizing the following cross-entropy function:

$$\begin{aligned}\min_{\mathbf{B}} E &= \sum_{\Theta_{i,j}\in\Theta}\sum_{k=1}^{2}\left(-\Theta_{i,j}^{k}\log P\left(\Theta_{i,j}^{k}=1\middle|\mathbf{B}\right)\right)\\ &= \sum_{\Theta_{i,j}\in\Theta}\left(\Theta_{i,j}^{1}\Omega_{i,j}+\log(1+e^{\Omega_{i,j}})\right). \qquad (2)\end{aligned}$$

To draw a link between deep feature learning and hashing learning, we give the parameter formulation of DFLNNs and HLNNs in the following. Let $\mathbf{\Lambda}$ denote all parameters of multilayers of DFLNNs, and let $\{\mathbf{W},\mathbf{v}\}$ denote the weights of HLNNs. For a given input image I_i, the high-dimensional semantic feature representation of DFLNNs can be represented by $\mathbf{d}_i = \varphi(I_i;\mathbf{\Lambda})$, where $\mathbf{d}_i \in R^d$, and the continuous low-dimensional feature representation of HLNNs can be represented by $\mathbf{f}_i = \mathbf{W}^T\mathbf{d}_i + \mathbf{v} = \mathbf{W}^T\varphi(I_i;\mathbf{\Lambda}) + \mathbf{v}$, where $\mathbf{f}_i \in R^l$, $\mathbf{W}\in R^{d\times l}$, and $\mathbf{v}\in R^l$.

To simultaneously optimize the DFLNNs and HLNNs, the optimization function shown in (2) can be converted into

$$\begin{aligned}\min_{B,\Lambda,W,\upsilon} E^1 = &\sum_{\Theta_{i,j}\in\Theta}\left(\Theta_{i,j}^{1}\Upsilon_{i,j}+\log(1+e^{\Upsilon_{i,j}})\right)\\ &+\eta\sum_{i=1}^{N}\|\mathbf{f}_i-\mathbf{b}_i\|_1 \qquad (3)\end{aligned}$$

where $\Upsilon_{i,j} = \mathbf{f}_i^T\mathbf{f}_j/P$, P is the similarity penalty, and η is the regularization coefficient. Using formula derivation, it is not difficult to see that P varies with the selection of sigmoid functions. The similarity penalty P is equal to 1, 2, and $w = s\cdot l$ when the classic sigmoid function in [27], the improved sigmoid function in [29], and the weighted sigmoid function are, respectively, adopted.

We note that the optimization function used in (3) takes the pairwise similarity constraint and the binary quantization loss function into consideration. Intuitively, the optimization function shown in (3) is equivalent to that used in (4). As the optimization function used in (3) and (4) uses the L1 norm to define the quantization loss, the corresponding DHNNs optimized by (3) or (4) are referred to as DHNNs-L1 in the following. In the proposed DHNNs-L1, the weighted sigmoid function is adopted and $\Upsilon_{i,j}$ in (4) is equal to $\mathbf{f}_i^T\mathbf{f}_j/w$, where $w = s\cdot l$ is the similarity weight. In contrast, the existing deep hashing method used in [27] employs the classic sigmoid function, which renders $\Upsilon_{i,j}$ used in (4) equal to $\mathbf{f}_i^T\mathbf{f}_j$. The binary quantization loss from the L1 norm is also adopted in [28]

$$\begin{aligned}\min_{\Lambda,W,\upsilon} E^1 = &\sum_{\Theta_{i,j}\in\Theta}\left(\Theta_{i,j}^{1}+\Upsilon_{i,j}+\log(1+e^{\Upsilon_{i,j}})\right)\\ &+\eta\sum_{i=1}^{N}\||\mathbf{f}_i|-\mathbf{1}\|_1. \qquad (4)\end{aligned}$$

Unlike the function used in (3) and (4), the optimization function used in (5) employs the square of the L2 norm to define the quantization loss. In the following, the DHNNs optimized by (5) are referred to as DHNNs-L2. Unlike the proposed DHNNs-L2, the existing deep hashing approach used in [29] adopts the improved sigmoid function, rendering $\Upsilon_{i,j}$ in (5) equal to $\mathbf{f}_i^T\mathbf{f}_j/2$

$$\begin{aligned}\min_{B,\Lambda,W,\upsilon} E^2 = &\sum_{\Theta_{i,j}\in\Theta}\left(\Theta_{i,j}^{1}\Upsilon_{i,j}+\log(1+e^{\Upsilon_{i,j}})\right)\\ &+\eta\sum_{i=1}^{N}\|\mathbf{f}_i-\mathbf{b}_i\|_2^2. \qquad (5)\end{aligned}$$

As noted above, we comprehensively review DHNN methods [27]–[29] employed in the literature under the cross-entropy optimization framework employed in (2). In diverging from prior efforts, the importance of the similarity weight w is revealed for the first time. In addition, we evaluate the final performance of DHNNs when applied under different quantization loss functions.

In Section II-B, ways to learn DHNNs-L1 and DHNNs-L2 from (3) and (5) are demonstrated in detail.

B. DHNN Learning

Given that the volume of training samples is generally very large, we adopt a batch-based learning strategy widely adopted in deep learning to optimize DHNNs-L1 used in (3) and DHNNs-L2 used in (5). More specifically, for each iteration, we sample a batch of data to learn parameters until all data are processed. As $\mathbf{B}$ and $\{\mathbf{\Lambda},\mathbf{W},\mathbf{v}\}$ are dependent on one another in (3) or (5), we adopt an alternative way to learn them. Therefore, one parameter is updated while other parameters remain fixed.

Regardless of whether we optimize DHNNs-L1 or DHNNs-L2, binary feature vectors $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^N$ should be first estimated based on neural network parameters $\{\mathbf{\Lambda},\mathbf{W},\mathbf{v}\}$

$$\mathbf{b}_i = \mathrm{sign}(\mathbf{f}_i) = \mathrm{sign}(\mathbf{W}^T\varphi(I_i;\mathbf{\Lambda})+\mathbf{v}) \qquad (6)$$

where $\mathrm{sign}(\cdot)$ maps each element of the feature vector to -1 or 1 based on the sign of the given element.

To learn neural network parameters via the backpropagation algorithm, we must compute derivatives of the optimization function. In the following, we, respectively, give the derivatives of optimization functions used in (3) and (5).

To learn the parameters employed in DHNNs-L1, the derivative of the optimization function used in (3) with respect to $\mathbf{f}_i$ should be computed as illustrated in (7). The optimization function used in (3) with respect to $\mathbf{f}_i$ is nondifferentiable due to its use of the L1 norm. As noted in [28], (7) gives derivatives on multiple intervals that can be written as

$$\frac{\partial E^1}{\partial \mathbf{f}_i^m} = \begin{cases}\sum\limits_{j:\Theta_{i,j}\in\Theta}\left(\sigma\left(\mathbf{f}_i^T\mathbf{f}_i/(s\cdot l)\right)-\Theta_{i,j}^1\right)\mathbf{f}_j^m+\eta, & \mathbf{f}_i^m\geq 1\\ \sum\limits_{j:\Theta_{i,j}\in\Theta}\left(\sigma\left(\mathbf{f}_i^T\mathbf{f}_i/(s\cdot l)\right)-\Theta_{i,j}^1\right)\mathbf{f}_j^m+\eta, & -1\leq\mathbf{f}_i^m\leq 0\\ \sum\limits_{j:\Theta_{i,j}\in\Theta}\left(\sigma\left(\mathbf{f}_i^T\mathbf{f}_i/(s\cdot l)\right)-\Theta_{i,j}^1\right)\mathbf{f}_j^m-\eta, & \text{otherwise}\end{cases} \qquad (7)$$

where l is the length of $\mathbf{f}_i$ and $m = 1:l$.

Algorithm 1 Optimization Process for DHNNs-L1

Input: Training images $\mathbf{I} = \{I_i\}_{i=1}^{N}$ with the pairwise similarity matrix Θ;

Output: Weights for DHNNs-L1 $\{\mathbf{\Lambda}, \mathbf{W}, \mathbf{v}\}$ and by-product binary features $\mathbf{B}$;

Repeat

Randomly sample a batch of images from the training images. For each image I_i in the sampled batch, execute the following operations:

- Compute the high-dimensional feature from $\mathbf{d}_i = \varphi(I_i; \Lambda)$ by forward propagation;
- Calculate the low-dimensional binary feature from $\mathbf{b}_i = \mathrm{sign}(\mathbf{W}^T \mathbf{d}_i + \mathbf{v})$ using Eq. (6);
- Calculate derivatives of the optimization function using Eq. (7) - Eq. (10);
- Update weights $\{\mathbf{\Lambda}, \mathbf{W}, \mathbf{v}\}$ based on the derivatives via back propagation;

Continue until all images are processed over a fixed number of iterations

Algorithm 2 Optimization Process for DHNNs-L2

Input: Training images $\mathbf{I} = \{I_i\}_{i=1}^{N}$ with the pairwise similarity matrix Θ;

Output: Weights for DHNNs-L2 $\{\mathbf{\Lambda}, \mathbf{W}, \mathbf{v}\}$ and by-product binary features $\mathbf{B}$;

Repeat

Randomly sample a batch of images from the training images. For each image I_i in the sampled batch, execute the following operations:

- Compute the high-dimensional feature by $\mathbf{d}_i = \varphi(I_i; \mathbf{\Lambda})$ by forward propagation;
- Calculate the low-dimensional binary feature $\mathbf{b}_i = \mathrm{sign}(\mathbf{W}^T \mathbf{d}_i + \mathbf{v})$ from Eq. (6);
- Calculate derivatives of the optimization function from Eq. (11) - Eq. (14);
- Update weights $\{\mathbf{\Lambda}, \mathbf{W}, \mathbf{v}\}$
- based on the derivatives by back propagation;

Continue until all images are processed with a fixed number of iterations

Furthermore, we can calculate derivatives of (3) with respect to $\{\mathbf{\Lambda}, \mathbf{W}, \mathbf{v}\}$, which can refer to the following:

$$\frac{\partial E^1}{\partial \varphi(I_i; \mathbf{\Lambda})} = \mathbf{W}\frac{\partial E^1}{\partial \mathbf{f}_i} \tag{8}$$

$$\frac{\partial E^1}{\partial \mathbf{W}} = \varphi(I_i; \mathbf{\Lambda})\left(\frac{\partial E^1}{\partial \mathbf{f}_i}\right)^T \tag{9}$$

$$\frac{\partial E^1}{\partial \mathbf{v}} = \frac{\partial E^1}{\partial \mathbf{f}_i}. \tag{10}$$

To illustrate, we summarize the optimization process employed for DHNNs-L1 as Algorithm 1.

In the following, we give the optimization solution for DHNNs-L2. As for the optimization process for DHNNs-L1, we must determine the derivative of the optimization function used in (5) with respect to $\mathbf{f}_i$. In benefiting from the L2 norm, the optimization function used in (5) with respect to $\mathbf{f}_i$ is differentiable. More specifically, the closed-form gradient is as follows:

$$\frac{\partial E^2}{\partial \mathbf{f}_i} = \sum_{j:\Theta_{i,j}\in\Theta} \left(\sigma\left(\mathbf{f}_i^T \mathbf{f}_j/(s \cdot l)\right) - \Theta_{i,j}^1\right)\mathbf{f}_j^m + 2\eta(\mathbf{f}_i - \mathbf{b}_i). \tag{11}$$

Based on the gradient result shown in (11), derivatives of the optimization function shown in (5) with respect to $\{\mathbf{\Lambda}, \mathbf{W}, \mathbf{v}\}$ can be computed from

$$\frac{\partial E^2}{\partial \varphi(I_i; \Lambda)} = \mathbf{W}\frac{\partial E^2}{\partial \mathbf{f}_i} \tag{12}$$

$$\frac{\partial E^2}{\partial \mathbf{W}} = \varphi(I_i; \mathbf{\Lambda})\left(\frac{\partial E^2}{\partial \mathbf{f}_i}\right)^T \tag{13}$$

$$\frac{\partial E^2}{\partial \mathbf{v}} = \frac{\partial E^2}{\partial \mathbf{f}_i}. \tag{14}$$

To avoid confusing this process with the optimization process employed for the DHNNs-L1, we summarize the optimization process of DHNNs-L2 as Algorithm 2.

III. Large-Scale Remote Sensing Image Retrieval via Deep Hashing Neural Networks

In this section, we propose a novel large-scale remote sensing image retrieval approach based on the aforementioned DHNNs composed of DFLNNs and HLNNs.

As illustrated in Fig. 3, the proposed large-scale remote sensing image retrieval approach based on the DHNNs involves two stages: a training stage and a testing stage. In the training stage, the DHNNs should be trained offline using labeled remote sensing images. In the testing stage, based on the DHNNs learned from the training stage, low-dimensional binary features of the given remote sensing images can be computed based on (6). As illustrated by the testing stage presented in Fig. 3, the large-scale remote sensing image retrieval task is transformed into a feature-searching problem. As noted above, the final feature representation of the DHNNs is very compact. In benefiting from this characteristic, the large-scale remote sensing image retrieval task can be easily implemented via exhaustive feature similarity comparisons, where similarities between binary features can be efficiently computed from the hamming distance [27]–[31]. As final features of the remote sensing image generated from the DHNNs are very compact, features of remote sensing images in the large-scale remote sensing image data set can be computed in advance and then saved as the feature data set without incurring considerable storage costs. Hence, in the retrieval stage, feature extraction time dedicated to the large-scale remote sensing image data set can be saved, and it is only necessary to compute the feature representation of the inquiry image based on the DHNNs.

It is well known that deep learning-based methods are often dependent on the use of millions of labeled samples to learn complex neural network parameters [24]–[26]. The DHNNs discussed in this paper also suffer from this problem. Hence, the performance of DHNNs depends heavily on the volume of labeled samples. To broader DHNNs applications,

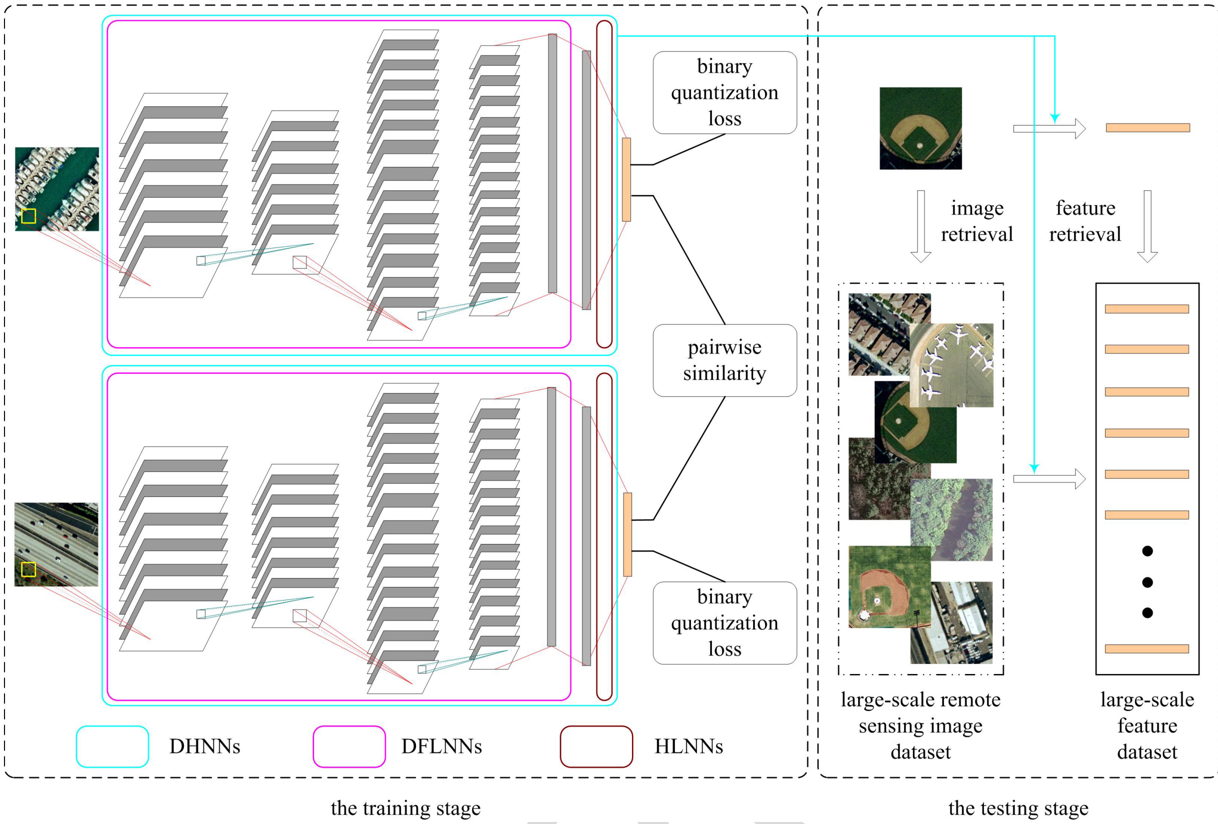


Fig. 3. Flowchart of the proposed large-scale remote sensing image retrieval approach based on DHNNs. The proposed approach involves training and testing stages. More specifically, the training stage involves learning DHNNs, and the testing stage addresses large-scale remote sensing image retrieval based on the DHNNs learned in the training stage.

Sections III-A and III-B present ways to design and train DHNNs under two typical cases for which the number of labeled remote sensing samples available is limited or sufficient.

A. Large-Scale Remote Sensing Image Retrieval by Virtue of Limited Number of Labeled Samples

In the majority of remote sensing applications, large numbers of remote sensing images are available, but labeled images are very rare. In such cases, fully learning convolutional neural networks (CNNs) from scratch is impossible. In the literature, several efforts have been made to transfer CNNs that have been pretrained in a large-scale natural image data set (e.g., ImageNet) [43] to remote sensing image tasks of scene classification [34], object recognition [39], and so on.

Inspired by such successful experiences [34], [39], we train DHNNs via transfer learning when the number of labeled remote sensing images available is very limited. More specifically, we expect to transfer CNNs pretrained on the source domain (e.g., the natural image object recognition task) to the target domain (i.e., the remote sensing image retrieval task). To this end, the DFLNNs of DHNNs can inherit from suitable pretrained CNNs (e.g., the one pretrained on ImageNet), and the HLNNs of DHNNs can be randomly initialized based on the size of the adopted DFLNNs. Furthermore, the constructed DHNNs can be incrementally trained by applying Algorithm 1 or Algorithm 2 under the supervision of a limited number of labeled remote sensing images. As the weights of DHNNs mainly concentrate on DFLNNs, a relatively strong DFLNNs initialization can decrease the optimization difficulty of DHNNs. In benefiting from the reuse of CNNs, the advocated DHNNs can be trained to achieve strong levels of generalization performance, even when the number of labeled remote sensing images available is very limited.

As a precondition to the success of this transfer learning strategy, the remote sensing image in the target domain relatively resembles the image in the source domain in terms of spectral ranges and spatial resolutions. In the training and testing stages, the remote sensing image in the target domain must be projected to the size of the image in the source domain to reuse CNNs trained in the source domain. Although the projection may lose some information on remote sensing images, this approach is still very cost effective when the remote sensing image adopted is similar to natural images. This strategy is verified for a public aerial image data set [44], and corresponding results are shown in Section IV-B.

B. Large-Scale Remote Sensing Image Retrieval With the Aid of a Sufficient Number of Labeled Samples

We note that the aforementioned transfer learning strategy for DHNNs may decline in efficacy when the remote sensing image used is significantly different from the image in the

Fig. 4. Illustration of the UCMD. The UCMD covers 21 land cover categories, and four images of each category randomly selected from the UCMD are shown.

source domain. As is well known, remote sensing images include much more spectral channels than natural images do. Hence, remote sensing images include even more cues that can be used in image analyses than natural ones do. When transferring CNNs pretrained on a natural image data set to construct the DFLNNs of DHNNs, only three RGB spectral channels of remote sensing images are used for feature representation, while the rich spectral information of remote sensing images is disregarded.

Along with the great successes of deep learning, more and more researchers have realized the importance of labeled samples. Accordingly, the remote sensing image data set with large volumes of labeled samples [45] has been released. In particular, a large-scale remote sensing image data set with manual labels is available. However, to our knowledge, no report has illustrated the feasibility of joint deep feature and hashing learning for remote sensing image data sets. To allow rich annotation information of remote sensing images to generate good yields, we attempt to specifically design and train DHNNs for remote sensing images from scratch. The solution proposed is verified based on one public satellite image data set [45], where each image contains four RGB–near infrared (NIR) spectral channels, and corresponding results are presented in Section IV.

IV. Experimental Results

Section IV-A introduces widely adopted evaluation criteria used for large-scale remote sensing image retrieval. Section IV-B provides an example that shows how DHNNs are designed and trained when the number of labeled samples available is very limited. In reference to such conditions, the overall performance of DHNNs and its performance relative to other approaches are reported. With the support of plenty of labeled samples, Section IV-C illustrates the means of designing and training DHNNs and reports on the overall performance of DHNNs and compares this performance with those of state-of-the-art approaches. Finally, Section IV-D provides a brief discussion of the experimental results and describes our future work related to DHNNs.

A. Evaluation Criteria

In this paper, large-scale remote sensing image retrieval performance is quantitatively evaluated using the following two widely adopted metrics [7], [27]–[31]: the mean average precision (MAP) and the precision-recall curve. More specifically, the MAP score can be computed from

$$\mathrm{MAP} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathrm{precision}\left(R_i^j\right) \tag{15}$$

TABLE I
CONFIGURATION OF DFLNN ON UCMD

Layer	Configuration
Conv1	filter: $64\times11\times11\times3$, stride1: 4×4 , pool: 3×3 , stride2: 2×2
Conv2	filter: $256\times5\times5\times64$, stride1: 1×1 , pool: 3×3 , stride2: 2×2
Conv3	filter: $256\times3\times3\times256$, stride1: 1×1
Conv4	filter: $256\times3\times3\times256$, stride1: 1×1
Conv5	filter: $256\times3\times3\times256$, stride1: 1×1 , pool: 3×3 , stride2: 2×2
Full6	4096
Full7	4096

TABLE II
MAP VALUES OF DHNNS-L1 UNDER DIFFERENT PARAMETERS ON UCMD

	$\eta=5.0e0$	$\eta=1.0e1$	$\eta=5.0e1$	$\eta=1.0e2$	$\eta=5.0e2$
$s=0.25$	0.6009	0.9406	0.9590	0.9539	0.3109
$s=0.50$	0.7010	0.8530	0.7506	**0.9587**	0.4735
$s=0.75$	0.1650	0.2450	0.6959	0.7123	0.3627
$s=1.00$	0.7141	0.7933	0.6770	0.5898	0.1250

TABLE III
MAP VALUES OF DHNNS-L2 UNDER DIFFERENT PARAMETERS ON UCMD

	$\eta=5.0e0$	$\eta=1.0e1$	$\eta=5.0e1$	$\eta=1.0e2$	$\eta=5.0e2$
$s=0.25$	0.9433	0.9520	0.9587	0.9503	0.0486
$s=0.50$	0.8436	0.9622	**0.9718**	0.9620	0.0956
$s=0.75$	0.8989	0.9708	0.9708	0.9596	0.1449
$s=1.00$	0.8585	0.8633	0.9701	0.9654	0.4295

where $q_i \in Q$ is the inquiry image, $|Q|$ denotes the volume of the inquiry image data set, and n_i is the number of images relevant to q_i in the searching image data set. Assuming that relevant images are ordered as $\{r_1, r_2, \ldots r_{n_i}\}$ across images in the searching image data set, R_i^j is the set of ranked results from the 1-st result to the r_j-th result.

B. Experiments on the Data Set With a Limited Number of Labeled Samples

1) Evaluation Data Set: In this paper, we take the publicly available University of California, Merced remote sensing image data set (UCMD) [44] to demonstrate how to design and train DHNNs from a limited number of labeled samples. The UCMD is generated by manually labeling aerial image scenes, and it covers 21 land cover categories. More specifically, each land cover category includes 100 images of 256×256 pixels, the spatial resolution of each pixel is 30 cm, and each pixel is measured in the RGB spectral space. Four representative images of each category of the UCMD are visually shown in Fig. 4. We note that the UCMD has been widely used for the performance evaluation of remote sensing image retrieval [11], [12], [20] and remote sensing image scene classification [21], [32]–[35] efforts. Hence, the UCMD is a representative remote sensing image data set that includes a limited number of labeled samples.

2) Experimental Setup: To slightly augment the volume of the UCMD, each image from the UCMD is rotated by 90°, 180, and 270°. This strategy has been widely adopted to enlarge data sets without any manual labor [34] and can increase the size of a UCMD by a factor of 4. In the following, we describe experiments conducted on the augmented UCMD containing 8400 images. Furthermore, the inquiry image data set is composed of 1000 images randomly sampled from the augmented UCMD, and the others are taken as searching and training image data sets with a volume of 7400.

In this experiment, the DFLNNs of DHNNs are constructed by transferring the CNNs pretrained on ImageNet [46] based on the fact that the aerial image of the UCMD resembles the natural image included in ImageNet in terms of spectral ranges and spatial resolutions, and the HLNNs of DHNNs are randomly initialized based on the output size of the DFLNNs. The specific configuration of the transferred DFLNNs is shown in Table I, and the DFLNNs can process an input image of $224 \times 224 \times 3$. In Table I, "filter" specifies the number of filters, the size of a field, and the dimensions of input data, and it can be formulated as num $\times$ size $\times$ size $\times$ dim. "stride1" denotes the sliding step of the convolution operation. "pool" denotes the down sampling factor. "stride2" denotes the sliding step of the local pooling operation.

Furthermore, the constructed DHNNs are incrementally optimized by Algorithm 1 or Algorithm 2 from the training aerial image data set. To distinguish between optimization algorithms, DHNNs-L1 denotes the DHNNs optimized by Algorithm 1, and DHNNs-L2 denotes the DHNNs optimized by Algorithm 2. In the incremental optimization process, the DFLNNs and HLNNs of DHNNs can be jointly updated under the supervision of the training aerial image data set.

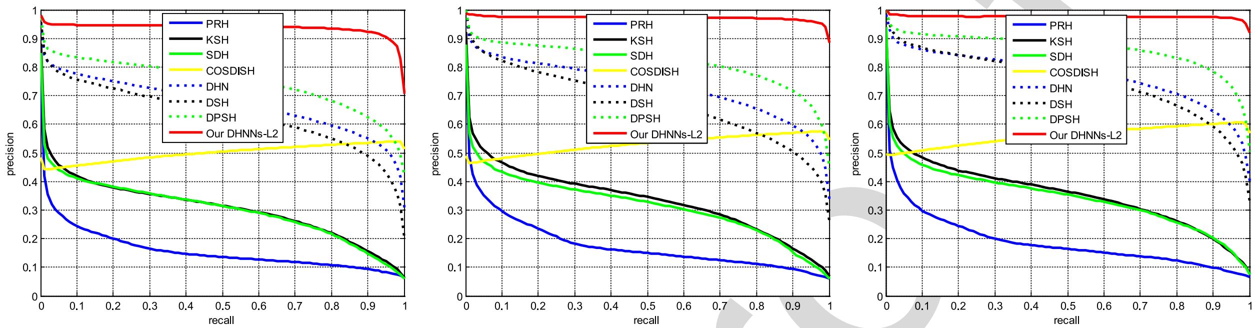
3) Overall Performance of the DHNNs: In this section, we explore the performance of DHNNs-L1 and DHNNs-L2 and the sensitivity of key parameters, including the similarity factor and regularization coefficient. In this experiment, the length of the final hashing feature is set to 64. The inquiry aerial image data set contains 1000 images, and the searching aerial image data set includes 7400 images. Based on this experimental setting, Table II reports the image retrieval performance of DHNNs-L1, and the retrieval performance is measured based on the MAP value. In addition, Table II presents sensitivity analysis results for key parameters, including the similarity factor s and the regularization coefficient η. In addition, Table III illustrates the image retrieval performance of DHNNs-L2 based on two critical parameters.

As illustrated in Tables II and III, DHNN-L2 performs better than DHNNs-L1. More specifically, DHNNs-L2 achieves the

TABLE IV

MAP VALUES OF DHNNS-L2 AND OTHER APPROACHES ON UCMD

	PRH in [7]	KSH in [8]	SDH in [30]	COSDISH in [31]	DHN in [27]	DSH in [28]	DPSH in [29]	Our DHNNs-L2
$l = 32$	0.1557	0.3039	0.2997	0.4998	0.6707	0.6317	0.7478	**0.9396**
$l = 64$	0.1744	0.3326	0.3144	0.5300	0.7313	0.6750	0.8174	**0.9718**
$l = 96$	0.1858	0.3539	0.3427	0.5594	0.7707	0.7502	0.8640	**0.9762**



(a) (b) (c)

Fig. 5. Performance of DHNNs-L2 and other methods when applied with different hashing feature lengths on UCMD. (a) Performance when $l = 32$. (b) Performance when $l = 64$. (c) Performance when $l = 96$.

best remote sensing image retrieval outcomes when the similarity factor is set to 0.50 and the regularization coefficient is equal to 5.0e1.

4) Comparisons With State-of-the-Art Approaches: With the similarity factor and regularization coefficient in DHNNs-L2 fixed, we report MAP values of our proposed DHNNs-L2 for different hashing feature lengths in Table IV. To show the superiority of the adopted DHNNs-L2, we compare it with state-of-the-art approaches, including two existing large-scale remote sensing image retrieval approaches [7], [8], two recently developed hashing learning methods [30], [31], and three existing DHNNs methods [27]–[29]. More specifically, the large-scale remote sensing image retrieval method based on partial randomness hashing (PRH) [7], the large-scale remote sensing image retrieval method based on kernel-based supervised hashing (KSH) [8], [47], the potential method based on supervised discrete hashing (SDH) [30], and the candidate method based on column sampling-based discrete supervised hashing (COSDISH) [31] are reimplemented or provided by the authors. These approaches [7], [8], [30], [31] take the 512-D GIST feature [48] as an input for hashing learning methods. To illustrate the benefits of the proposed DHNNs-L2, we also compare it with existing DHNNs models, including the deep hashing network (DHN) [27], deep supervised hashing (DSH) [28], and deep pairwise-supervised hashing (DPSH) [29]. Experimental parameters are set according to suggestions made in corresponding papers. To illustrate the superiority of the optimization function of the proposed DHNNs-L2, the DHN [27], DSH [28], and DPSH [29] are based on the same deep network architecture of the proposed DHNNs-L2. As shown in Table IV, we can easily conclude that the proposed DHNNs-L2 can clearly outperform other state-of-the-art approaches.

To further illustrate aerial image retrieval performance outcomes, we present precision-recall curves of DHNNs-L2 and of other approaches. Fig. 5 shows the precision-recall curves of methods based on different hashing feature lengths. As illustrated in Fig. 5, DHNNs-L2 significantly outperforms the other approaches.

In addition to the above quantitative comparison with state-of-the-art approaches, we draw intuitive comparisons, as illustrated in Fig. 6. For this visual comparison, the hashing feature length of all methods is set to 96, and all methods use the same inquiry image and the same search image data set. In Fig. 6, the aerial scene containing storage tanks is taken as the inquiry image, and retrieval results of different methods are shown. As shown in Fig. 6, DHNNs-L2 clearly outperforms other methods and retrieves true aerial images, even in the midst of considerable appearance variations. Due to space limitations, we only provide one visual retrieval example, though DHNNs-L2 applies to other cases as reflected in the comprehensive results shown in Table IV and Fig. 5.

C. Experiments on the Data Set With Oversized Labeled Samples

1) Evaluation Data Set: In this section, we use a public satellite image data set based on four land cover categories (SAT4) [45] as a case to explore the feasibility of jointly learning deep feature representation and hashing mapping

Fig. 6. Visual image retrieval results of different methods examined. (a) Inquiry aerial image of the storage tanks category. (b) PRH retrieval results presented in [7]. (c) KSH retrieval results presented in [8]. (d) SDH retrieval results presented in [30]. (e) COSDISH retrieval results presented in [31]. (f) DHN retrieval results presented in [27]. (g) DSH retrieval results presented in [28]. (h) DPSH retrieval results presented in [29]. (i) Retrieval results of our DHNNs-L2. The 1st, 5th, 10th, 15th, 20th, 30th, 40th, and 50th retrieval results of each method are shown. In addition, false retrieval results are marked with red rectangles.

functions from scratch. Images in the SAT4 were drawn from the National Agriculture Imagery Program. Each image in the SAT4 includes 28×28 pixels, the spatial resolution of each pixel is 1 m, and each pixel is measured in the RGB–NIR spectral space. In addition, the SAT4 includes 500 000 images covering four land cover categories (barren land, trees,

TABLE V
CONFIGURATION OF DFLNN ON SAT4

Layer	Configuration
Conv1	filter: $32\times5\times5\times4$, stride1: 1×1 , pool: 3×3 , stride2: 2×2
Conv2	filter: $32\times3\times3\times32$, stride1: 1×1 , pool: 3×3 , stride2: 2×2
Conv3	filter: $64\times3\times3\times32$, stride1: 1×1 , pool: 2×2 , stride2: 1×1
Full4	128
Full5	128

TABLE VI
MAP VALUES OF DHNNS-L1 UNDER DIFFERENT PARAMETERS ON SAT4

	$\eta=1.0e0$	$\eta=1.0e1$	$\eta=1.0e2$	$\eta=1.0e3$	$\eta=1.0e4$
$s=0.25$	0.9694	0.9781	0.9784	0.9743	0.9459
$s=0.50$	0.9736	0.9772	0.9787	**0.9793**	0.9640
$s=0.75$	0.9765	0.9700	0.9746	0.9746	0.9613
$s=1.00$	0.9744	0.9773	0.9750	0.8450	0.9494

TABLE VII
MAP VALUES OF DHNNS-L2 UNDER DIFFERENT PARAMETERS ON SAT4

	$\eta=1.0e0$	$\eta=1.0e1$	$\eta=1.0e2$	$\eta=1.0e3$	$\eta=1.0e4$
$s=0.25$	0.9736	0.9769	0.9765	0.9736	0.6471
$s=0.50$	0.9769	0.9808	0.9811	0.9788	0.6258
$s=0.75$	0.9736	0.9785	**0.9819**	0.9756	0.6417
$s=1.00$	0.8479	0.9778	0.8615	0.9814	0.7503

grassland, and all land cover types other than the former three classes). Visual samples drawn from the SAT4 are shown in Fig. 7.

2) Experimental Setup: From this experiment, we randomly selected 1000 images from the SAT4 as an inquiry image data set, and others were used as a searching and training image data sets with a volume of 499 000. Hence, it was sufficient to learn a specific deep neural network aiming at given types of satellite images under the supervision of this training satellite image data set. In addition, the inquiry and searching image data sets were further used to evaluate image retrieval performance outcomes.

As the satellite image was measured in the RGB–NIR spectral space and the size of the image is relatively small, Table V presents the architecture of the DFLNN specifically designed for such satellite images. As shown in Table V, the architecture contains three convolutional layers and two fully connected layers and is relatively compact compared to the ImageNet network. We note that the architecture given in Table V is just one of the many candidates. This paper merely introduces a general solution for designing DFLNNs and for further constructing DHNNs. More DFLNNs architectures can be explored and evaluated in future works. Under the applied experimental setting, both the DFLNNs and HLNNs of the DHNNs were randomly initialized. Furthermore, we can use Algorithm 1 or Algorithm 2 to train it from scratch using the training satellite image data set.

3) Overall Performance of the DHNNs: In this experiment, we used a training image data set of 499 000 images to train the DHNNs from scratch using different optimization algorithms. In the following, DHNNs-L1 is the constructed DHNNs optimized by Algorithm 1, and DHNNs-L2 is the constructed DHNNs optimized by Algorithm 2. With the hashing feature length set to 64, Table VI illustrates the satellite image retrieval accuracy of DHNNs-L1 equipped with two parameters, including the similarity factor s and regularization coefficient η. Table VII reports the satellite image retrieval accuracy of DHNNs-L2 under two key parameters.

As shown in Tables VI and VII, DHNNs-L2 performs better than DHNNs-L1. DHNNs-L2 can achieve the best satellite image retrieval performance outcomes when the similarity factor s is set to 0.75 and the regularization coefficient η is equal to 1.0e2.

4) Comparisons With State-of-the-Art Approaches: According to the sensitivity analysis of the similarity factor and the regularization coefficient shown in Section IV-C-3, the similarity factor s and regularization coefficient η of the DHNNs-L2 are set as 0.75 and 1.0e2, respectively. Furthermore, Table VIII reports the accuracy of DHNNs-L2 when a different hashing feature length l is adopted. To illustrate the superiority of DHNNs-L2, we also present the accuracy of the following seven state-of-the-art approaches: PRH [7], KSH [8], SDH [30], COSDISH [31], DHN [27], DSH [28], and DPSH [29]. These shallow hashing methods [7], [8], [30], [31] used the 512-D GIST feature [48] as an input. In addition, these deep hashing methods [27]–[29] use the same deep network architecture as that employed for the proposed DHNNs-L2. For the comparisons, all methods employ the same inquiry and searching data sets. As shown in Table VIII, the proposed DHNNs-L2 achieves significant satellite image retrieval performance improvements relative to other existing methods.

To clearly show image retrieval performance variations of the different methods, we report the precision-recall curves of DHNNs-L2 and of other approaches. More specifically, Fig. 8 reports the precision-recall curves of the different methods for different hashing feature lengths. As shown in Fig. 8, the proposed DHNNs-L2 significantly outperforms the other approaches.

For the same hashing feature length $l = 96$, we report the visual retrieval results of DHNNs-L2 and other approaches in Fig. 9. As a whole, the quantitative and qualitative results illustrate the superiority of the proposed DHNNs-L2.

There is no doubt that the feature-searching module can be efficiently applied through the utilization of hashing

Fig. 7. RGB channel visualization of the adopted SAT4. More specifically, SAT4 covers four land cover categories, and 24 images of each category, randomly selected from the SAT4, are shown.

TABLE VIII
MAP VALUES OF DHNNS-L2 AND OTHER APPROACHES ON SAT4

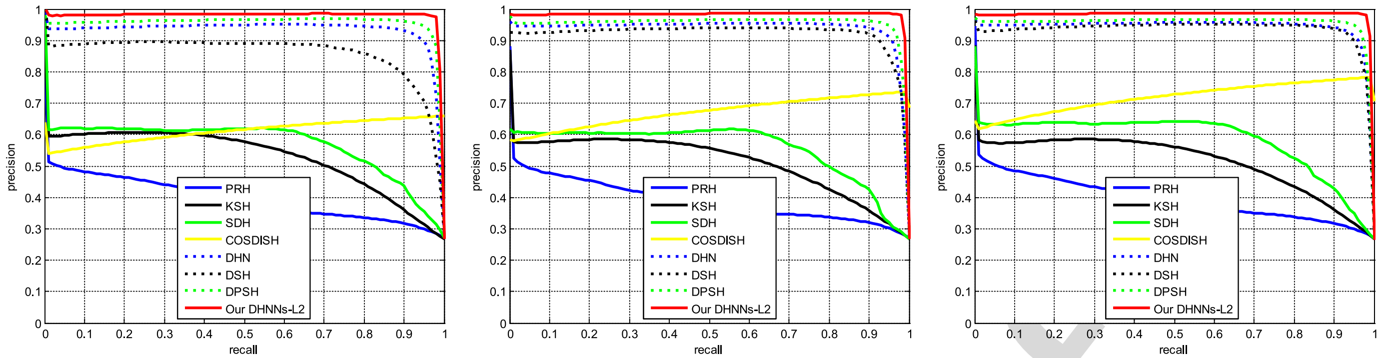
	PRH in [7]	KSH in [8]	SDH in [30]	COSDISH in [31]	DHN in [27]	DSH in [28]	DPSH in [29]	Our DHNNs-L2
$l = 32$	0.3933	0.5280	0.5681	0.6110	0.9321	0.8595	0.9554	**0.9793**
$l = 64$	0.3881	0.5103	0.5574	0.6714	0.9391	0.9212	0.9549	**0.9819**
$l = 96$	0.3946	0.5133	0.5830	0.7192	0.9431	0.9341	0.9561	**0.9830**

features [7], [8]. In practice, the efficient extraction of hashing features from images is very challenging. Fortunately, the proposed DHNNs can be easily applied with the use of parallel hardware. In this paper, the proposed DHNNs-L2 is implemented via GPU. The proposed DHNNs can extract hashing features of dozens of aerial images of the UCMD per second and can output hashing features of hundreds of satellite images of the SAT4 each second. As a whole, the proposed DHNNs-L2 is accurate and efficient.

D. Discussion and Avenues for Future Research

In the aforementioned experiments, the two remote sensing image data sets used (i.e., the UCMD and SAT4) represent two typical remote sensing image retrieval task conditions. Under these two different conditions, DHNNs can be designed and learned under a unified framework. Our two representative experiments fully show the generalization of the proposed DHNNs-L2. In addition, the experiments show that the proposed DHNNs-L2 can achieve significant performance improvements relative to the outcomes of two existing large-scale remote sensing image retrieval approaches [6], [7], two potential approaches based on recent hashing learning methods [30], [31], and three existing deep hashing methods [27]–[29].

In future work, we will explore ways to train DHNNs from scratch using large-scale labeled data with noisy, possibly



(a) (b) (c)

Fig. 8. Performance of DHNNs-L2 and other methods when applied with different hashing feature lengths on SAT4. (a) Performance when $l = 32$. (b) Performance when $l = 64$. (c) Performance when $l = 96$.

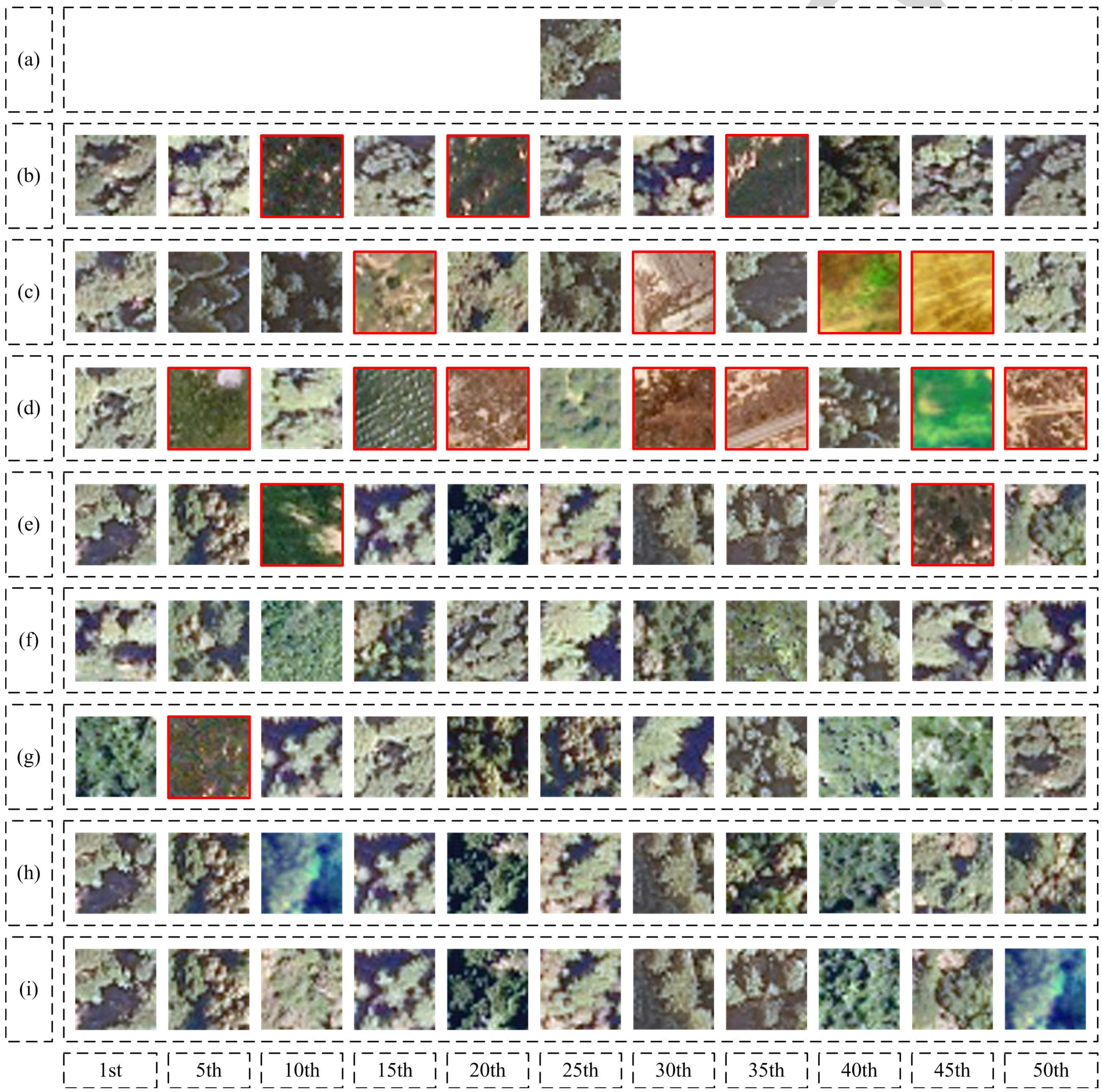


Fig. 9. Visual image retrieval results for different methods. (a) Inquiry satellite image of the tree category. (b) PRH retrieval results presented in [7]. (c) KSH retrieval results presented in [8]. (d) SDH retrieval results presented in [30]. (e) COSDISH retrieval results presented in [31]. (f) DHN retrieval results presented in [27]. (g) DSH retrieval results presented in [28]. (h) DPSH retrieval results presented in [29]. (i) Retrieval results of our DHNNs-L2. The 1st, 5th, 10th, 15th, 20th, 25th, 30th, 35th, 40th, 45th, and 50th retrieval results of each method are shown. In addition, false retrieval results are marked with red rectangles.

incorrect labels. These data are often generated at a relatively low cost. For example, remote sensing images can be efficiently labeled through crowd-sourcing [49], but labeled data can contain a certain number of incorrect labels [50]. Guided by the geography information system, remote sensing images can also be labeled automatically with the cost of a certain number of alignment errors [51]. Hence, DHNNs training from noisy labeled data should be very cost effective.

As noted above, DHNNs can output the compact semantic feature representation of an input remote sensing image in urgent need of remote sensing image interpretation. Hence, we plan to explore more applications of DHNNs such as hyper-spectral image classification [52], image matching and registration [53], [54], information fusion [55], built-up area detection [56], urban village detection [57], [58], and land cover recognition [59].

V. Conclusion

Due to an urgent need for RSBD mining, large-scale remote sensing image retrieval has attracted increasing attention. Although several efforts have been made to address issues of large-scale remote sensing image retrieval, this task remains a very challenging problem. This paper is the first to advocate the use of DHNNs to address this problem.

We conduct a comprehensive study of DHNN systems. Based on the general cross-entropy theory, we provide a systematic review of existing DHNN methods. This paper is the first to highlight the importance of the similarity weight, which is set to a constant and disregarded in existing works. To broaden the applications of DHNNs, we adapt DHNNs to two representative remote sensing cases where the remote sensing data set includes either a limited number of labeled samples or plenty of labeled samples. For these two conditions, we present the means to design and train DHNNs. Extensive experiments conducted on one public aerial image data set and one public satellite image data set demonstrate that the proposed large-scale remote image retrieval approach based on the adjusted DHNNs can remarkably outperform state-of-the-art approaches.

Large-scale remote sensing image retrieval methods and DHNNs should be increasingly adapted to address the requirements of more and more practical applications. To facilitate this, we present potential avenues for future research on DHNNs from method optimization and application perspectives. In future work, we plan to explore ways to train DHNNs using labeled data containing a certain number of errors from scratch, as such data can often be generated at a low cost. In addition, we plan to exploit the feasibility of applying DHNNs to more remote sensing image interpretation applications. Broadly speaking, DHNNs and their future extensions could realize new solutions for a broad range of remote sensing applications.

References

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

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

[3] L. Wang, H. Zhong, R. Ranjan, A. Zomaya, and P. Liu, "Estimating the statistical characteristics of remote sensing big data in the wavelet transform domain," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 324–337, Sep. 2014.

[4] 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.

[5] K. W. Tobin *et al.*, "Automated feature generation in large-scale geospatial libraries for content-based indexing," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 5, pp. 531–540, 2006.

[6] C.-R. Shyu, M. Klaric, G. J. Scott, A. S. Barb, C. H. Davis, and K. Palaniappan, "GeoIRIS: Geospatial Information Retrieval and Indexing System—Content mining, semantics modeling, and complex queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 839–852, Apr. 2007.

[7] 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.

[8] 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.

[9] B. Demir and L. Bruzzone, "A novel active learning method in relevance feedback for content-based remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2323–2334, May 2015.

[10] M. Wang and T. Song, "Remote sensing image retrieval by scene semantic matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2874–2886, May 2013.

[11] 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.

[12] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 3023–3034, May 2014.

[13] Y. Hongyu, L. Bicheng, and C. Wen, "Remote sensing imagery retrieval based-on Gabor texture feature classification," in *Proc. 7th Int. Conf. Signal Process.*, Aug./Sep. 2004, pp. 733–736.

[14] S. Newsam, L. Wang, S. Bhagavathy, and B. S. Manjunath, "Using texture to analyze and manage large collections of remote sensed image and video data," *Appl. Opt.*, vol. 43, no. 2, pp. 210–217, 2004.

[15] B. Luo, J. F. Aujol, Y. Gousseau, and S. Ladjal, "Indexing of satellite images with different resolutions by wavelet features," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1465–1472, Aug. 2008.

[16] R. Rosu, M. Donias, L. Bombrun, S. Said, O. Regniers, and J.-P. Da Costa, "Structure tensor Riemannian statistical models for CBIR and classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 248–260, Jan. 2017.

[17] W. Zhou, Z. Shao, C. Diao, and Q. Cheng, "High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder," *Remote Sens. Lett.*, vol. 6, no. 10, pp. 775–783, 2015.

[18] 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, pp. 489–508, 2017.

[19] Z. Du, X. Li, and X. Lu, "Local structure learning in high resolution remote sensing image retrieval," *Neurocomputing*, vol. 207, pp. 813–822, Sep. 2016.

[20] Y. Li, Y. Zhang, C. Tao, and H. Zhu, "Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion," *Remote Sens.*, vol. 8, no. 9, pp. 709–723, 2016.

[21] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised multilayer feature learning for satellite image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 157–161, Feb. 2016.

[22] Y. Wang *et al.*, "A three-layered graph-based learning approach for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6020–6034, Oct. 2016.

[23] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proc. VISAPP Int. Conf. Comput. Vis. Theory Appl.*, 2009, pp. 331–340.

[24] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 26th Annu. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[27] H. Zhu, M. Long, J. Wang, and Y. Cao, "Deep hashing network for efficient similarity retrieval," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2415–2421.

[28] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2064–2072.

[29] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1711–1717.

[30] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 37–45.

[31] W.-C. Kang, W.-J. Li, and Z.-H. Zhou, "Column sampling based discrete supervised hashing," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1230–1236.

[32] 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.

[33] 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.

[34] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land–cover classification of high-resolution imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 549–553, Apr. 2017.

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

[36] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[37] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[38] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, "Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5585–5599, Oct. 2017.

[39] H. Liu, S. Yang, S. Gou, D. Zhu, R. Wang, and L. Jiao, "Polarimetric SAR feature extraction with neighborhood preservation-based deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1456–1466, Apr. 2017.

[40] J. Geng, H. Wang, J. Fan, and X. Ma, "Deep supervised and contractive neural network for SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2442–2459, Apr. 2017.

[41] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[42] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[43] 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. 2–9.

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

[45] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "DeepSat—A learning framework for satellite imagery," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, 2015, Art. no. 37.

[46] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *Proc. Brit. Mach. Vis. Conf.*, 2014, vol. 2. no. 4.

[47] 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.

[48] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[49] P. Welinder and P. Perona, "Online crowdsourcing: Rating annotators and obtaining cost-effective labels," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 25–32.

[50] H. Li, B. Yu, and D. Zhou, "Error rate analysis of labeling by crowdsourcing," in *Proc. Int. Conf. Mach. Learn. Workshop, Mach. Learn. Meets Crowdsourcing*, 2013, pp. 1–22.

[51] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 567–574.

[52] Z. Zhong, B. Fan, K. Ding, H. Li, S. Xiang, and C. Pan, "Efficient multiple feature fusion with hashing for hyperspectral imagery classification: A comparative study," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4461–4478, Aug. 2016.

[53] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.

[54] J. Ma, J. Zhao, J. Tian, X. Bai, and Z. Tu, "Regularized vector field learning with sparse approximation for mismatch removal," *Pattern Recognit.*, vol. 46, no. 12, pp. 3519–3532, 2013.

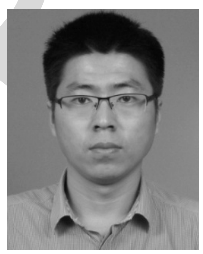
[55] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.

[56] Y. Li, Y. Tan, J. Deng, Q. Wen, and J. Tian, "Cauchy graph embedding optimization for built-up areas detection from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2078–2096, May 2015.

[57] X. Huang, H. Liu, and L. Zhang, "Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3639–3657, Jul. 2015.

[58] Y. Li, X. Huang, and H. Liu, "Unsupervised deep feature learning for urban village detection from high-resolution remote sensing images," *Photogramm. Eng. Remote Sens.*, vol. 83, no. 8, pp. 567–579, 2017.

[59] J. Fan, T. Chen, and S. Lu, "Unsupervised feature learning for land-use scene recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2250–2261, Apr. 2017.



Yansheng Li received the B.S. degree from the School of Mathematics and Statistics, Shandong University, Weihai, China, in 2010, and the Ph.D. degree from the School of Automation, Huazhong University of Science and Technology, Wuhan, China, in 2015.

Since 2015, he has been an Assistant Professor with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. Currently, he is a Visiting Assistant Professor with the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA, where he is hosted by the Distinguished Bloomberg Professor A. L. Yuille; he will hold the position till 2018. He has authored more than 20 peer-reviewed articles in international journals from multiple domains such as remote sensing and computer vision. His research interests include computer vision, machine learning, deep learning, and their applications in remote sensing.

Dr. Li has been frequently serving as a reviewer for more than six international journals including the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, *Photogrammetric Engineering and Remote Sensing, and Remote Sensing*. He is also a Communication Evaluation Expert for the National Natural Science Foundation of China.

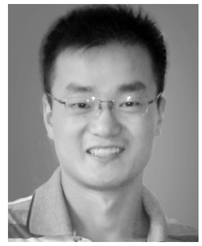
Yongjun Zhang was born in 1975. He received the B.S., M.S., and Ph.D. degrees from Wuhan University (WHU), Wuhan, China, in 1997, 2000, and 2002, respectively.

He is currently a Professor of photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, WHU. His research interests include space, aerial, and low-attitude photogrammetry, image matching, combined bundle adjustment with multisource data sets, and 3-D city reconstruction.

Xin Huang (M'13–SM'14) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009.

He is currently a Luojia Distinguished Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, where he teaches remote sensing, photogrammetry, image interpretation, etc. He is the Founder and Director of the Institute of Remote Sensing Information Processing, School of Remote Sensing and Information Engineering, Wuhan University. He has authored more than 100 peer-reviewed articles (SCI papers) in international journals. His research interests include remote sensing image processing methods and applications.

Prof. Huang is supported by The Youth Talent Support Program of China in 2017, and was supported by the China National Science Fund for Excellent Young Scholars in 2015, and the New Century Excellent Talents in University from the Ministry of Education of China in 2011. He was a recipient of the Boeing Award for the Best Paper in Image Analysis and Interpretation from the American Society for Photogrammetry and Remote Sensing in 2010, the National Excellent Doctoral Dissertation Award of China in 2012, and the winner of the IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest in 2014. In 2011, he was recognized by the IEEE GRSS as a Best Reviewer of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the lead Guest Editor of the special issue on Information Extraction From High-Spatial-Resolution Optical Remotely Sensed Imagery for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (vol. 8, no.5, May 2015) and the special issue on Sparsity-Driven High-Dimensional Remote Sensing Image Processing and Analysis for the *Journal of Applied Remote Sensing* (vol.10, no.4, Oct 2016). Since 2014, he has been an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. Since 2016, he has been an Associate Editor of *Photogrammetric Engineering and Remote Sensing*.



Hu Zhu received the B.S. degree in mathematics and applied mathematics from Huaibei Coal Industry Teachers College, Huaibei, China, in 2007, and the M.S. and Ph.D. degrees in computational mathematics and pattern recognition and intelligent systems from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2013, respectively.

In 2013, he joined the Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include pattern recognition, image processing, and computer vision.

Jiayi Ma received the B.S. degree in mathematics from the Huazhong University of Science and Technology, Wuhan, China, in 2008, and the Ph.D. degree from the School of Automation, Huazhong University of Science and Technology, in 2014.

From 2012 to 2013, he was an Exchange Student with the Department of Statistics, University of California, Los Angeles, CA, USA. From 2014 to 2015, he was a Post-Doctoral Researcher with Wuhan University, Wuhan, where he is currently an Associate Professor with the Electronic Information School. He has authored or co-authored more than 70 scientific articles. His research interests include computer vision, machine learning, and pattern recognition.