Multiscale Context Deep Hashing for Remote Sensing Image Retrieval

Dongjie Zhao, Yaxiong Chen, and Shengwu Xiong

***Abstract*—With the advancement of remote sensing satellites and sensor technology, the quantity and diversity of remote sensing imagery have exhibited a sustained trend of growth. Remote sensing image retrieval has gained significant attention in the realm of remote sensing. Hashing methods have been widely applied in remote sensing image retrieval due to their high computational efficiency, low storage cost, and effective performance. However, existing remote sensing image retrieval methods often struggle to accurately capture the intricate information of remote sensing images. They often lack high attention to key features. The neglect of multiscale and saliency information in remote sensing images can result in feature loss and difficulties in maintaining the balance of hash codes. In response to the issues, we introduce a multiscale context deep hashing network (MSCDH). First, we can obtain finer-grained multi-scale features and achieve a larger receptive field by incorporating the proposed multiscale residual blocks. Then, the proposed multicontext attention modules can increase the perceptual field and suppress the interference from irrelevant information by aggregating contextual information along channels and spatial dimensions. The experimental results on the UCMerced dataset and WHU-RS dataset demonstrate that the proposed method achieves state-of-the-art retrieval performance.**

***Index Terms*—Attention mechanism, deep hash, multiscale context information.**

Manuscript received 24 April 2023; revised 15 June 2023; accepted 14 July 2023. Date of publication 26 July 2023; date of current version 7 August 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160604, in part by the Project of Sanya Yazhou Bay Science and Technology City under Grant SCKJ-JYRC-2022-76 and Grant SCKJ-JYRC-2022-17, in part by the NSFC under Grant 62101393 and Grant 62176194, in part by the Major Project of IoV under Grant 2020AAA001, in part by the Youth Fund Project of Hainan Natural Science Foundation under Grant 6220N344, in part by the CAAI-Huawei MindSpore Open Fund under Grant CAAIXSJLJJ-2022-001A, in part by the Knowledge Innovation Program of Wuhan-Basic Research, in part by the Sanya Science and Education Innovation Park of Wuhan University of Technology under Grant 2022KF0020, in part by the Fundamental Research Funds for the Central Universities under Grant WUT:223110001, supported by High-performance Computing Platform of YZBSTCACC, in part by the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX1148, in part by the Hainan Special PhD Scientific Research Foundation of Sanya Yazhou Bay Science and Technology City under Grant HSPHDSRF-2023-03-002, and in part by MindSpore, which is a new deep learning computing framework (https://www.mindspore.cn/). *(Corresponding author: Shengwu Xiong.)*

The authors are with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China, also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China, also with the Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China, and also with the Chongqing Research Institute, Wuhan University of Technology, Chongqing 401122, China (e-mail: zhaodj@whut.edu.cn; 593544199@qq.com; xiongsw@whut.edu.cn).

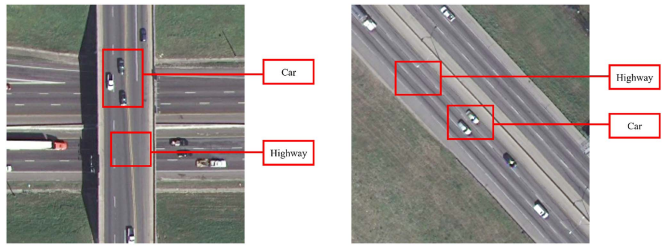
This article has supplementary downloadable material available at https://doi.org/10.1109/JSTARS.2023.3298990, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2023.3298990

I. Introduction

As SATELLITE observation technology advances rapidly, we can collect a large volume of high-resolution, accurate remote sensing (RS) data from Earth observation satellites every day. The huge RS image database contains rich geographic information, which plays an important role in many fields such as environmental monitoring, disaster rescue, meteorological analysis, economic evaluation, and ecological prediction. RS image retrieval technology refers to finding RS images in a dataset that are identical or similar to its content based on a specified query image [1], [2]. Due to the massive amount of data, researchers focus on utilizing computer technology for efficient storage, management, and analysis of these huge volumes of RS imagery. There are two categories based on how the images are processed. First, the text-based image retrieval (TBIR) [3] technology utilizes text annotations to describe the content of images. Although easy to implement, manual annotation is limited by the cognitive level, linguistic expression, and subjective judgment, and requires a significant amount of time and labor cost, making it impractical for large-scale applications. Second, the content-based image retrieval (CBIR) [4] only requires tagging image categories and includes two modules: 1) image feature extraction and 2) image retrieval. The features of feature extraction can be divided into three levels: 1) low-level; 2) middle-level; and 3) high-level semantic features. Low-level feature representation is a feature description method of RS images, which is constructed by color features [5], [6], spectral features [7], texture features [8], [9], [10], and shape features [11], [12] of RS images. Color features can provide a visual representation of the objects present in an image, with minimal dependence on image size, orientation, and viewpoint. Texture features are independent of color or brightness and exhibit properties such as hierarchical structure, scale, and translation invariance. It is more sensitive to the spatial variation of pixel intensity in RS images and pays more attention to the surface properties of image regions, and a single point cannot calculate texture features. Compared to low-level features, middle-level features carry more abundant information and exhibit greater expressive power. For example, the Bag of Visual Words (BOVW) [13] model obtains the visual dictionary by clustering based on local features. A feature encoding technique called Vector of Locally Aggregated Descriptors [14] builds a dictionary using techniques for clustering and encodes features by keeping track of the distances between local features and cluster centers. Scale-Invariant Feature Transform has the advantages of

Fig. 1. Remote sensing images have multi-scale features. (a) Larger scale airplane image. (b) Smaller scale airplane image.



(a) (b)

Fig. 2. Remote sensing images with high interclass similarity. (a) Overpass. (b) Freeway.

robustness to occlusion and invariance to viewpoint and illumination conditions.

Manual feature extraction is too expensive and impractical for large-scale datasets. As deep learning techniques evolve, convolutional neural networks [15] use a multilayer convolutional architecture to perform a high-dimensional nonlinear mapping of the image and obtain high-level semantic features of the image. The image features generated using the CNNs method encompass richer and more abstract semantic information. In comparison to manual feature extraction methods, CNNs often achieve superior performance in retrieval and classification [16] tasks with higher accuracy. When dealing with large-scale RS images, the massive amount of data requires increased storage space, and the computational complexity of the retrieval process grows with the increase in data. Approximate nearest neighbor (ANN) [17] methods have been shown to effectively address the challenges of dimensionality explosion and high storage space, as demonstrated in previous studies. In many ANN search methods, hash algorithms [18], [19], [20] have demonstrated notable advantages in terms of feature storage and retrieval speed by mapping high-dimensional image features into shorter binary codes. The hashing algorithm learns to compare the hash codes of images in the Hamming space, and then sorts similar images in ascending order of Hamming distance based on the hash codes to output the results. The combination of deep learning and hash learning in deep hashing methods has achieved encouraging retrieval results in image retrieval tasks [21], [22], [23], [24]. The RS images in Fig. 1 exhibit diverse scale variations; the performance of convolutional neural networks may be suboptimal. Owing to the difficulty in accurately capturing the characteristics of important objects that occupy a small area in complex scenes, it is easy to confuse the background with the main body of the image. In addition, RS images still exhibit high interclass similarity. For instance, as depicted in Fig. 2, the target features of highways and overpasses in RS images are similar. The prominent features are vehicles and roads, with low discriminability. During retrieval, it is prone to put the two into different categories. Moreover, there is significant intraclass variation also, as evident in Fig. 4 where all images belong to the commercial category but exhibit substantial visual differences. Therefore, we designed an effective feature extraction method that tackles a crucial issue in remote RS retrieval. We propose a multiscale context deep hashing network (MSCDH) for RS image retrieval.

1) We propose a new hashing network that leverages the multiscale features of RS images, obtaining multiscale contextual information. This enhances the intragroup correlation of hash codes by incorporating multiscale contextual information and category semantics.
2) In order to obtain more discriminative RS image features and mitigate the degradation of the network due to deepening, we propose multiscale residual blocks (MSRBs). By utilizing multiple channel groups, the MSRBs aim to extract multiscale image features in a more fine-grained manner.
3) The proposed multicontext attention (MCA) module aggregates contextual information to enhance perception and suppress irrelevant interference. Meanwhile the attention map can highlight the distinguishability of the acquired features.

II. Related Work

A. RS Image Retrieval Methods

Traditional RS image retrieval methods rely on handcrafted features to express the content of an image. Deep learning methods have the capability to establish low-level feature information and deep semantic information in remote sensing images, enabling abstract representation of RS images. As a result, these methods are more effective in RS image retrieval tasks. Liu et al. [25] proposed an unsupervised RS image retrieval method, which transforms similarity learning into deep ordinal classification results. The problem of relying on a vast number of tagged samples in traditional RS image retrieval is addressed. Xiao et al. [26] proposed a deep compression coding method to learn low-dimensional features in CNN, which better preserves the global information and spatial structural information of images.

Although deep learning based methods can yield favorable outcomes in the context of dealing with large-scale RS images, optimizing retrieval efficiency and storage space utilization are basic factors to consider. Hashing methods are widely used in remote sensing image retrieval because of their compact storage space and efficient processing speed. The supervised hashing algorithm learns the hash function by projecting the samples into Hamming space using label information. The unsupervised deep hashing approach is to generate compact and efficient hash codes without label or category information. Li et al. [27]

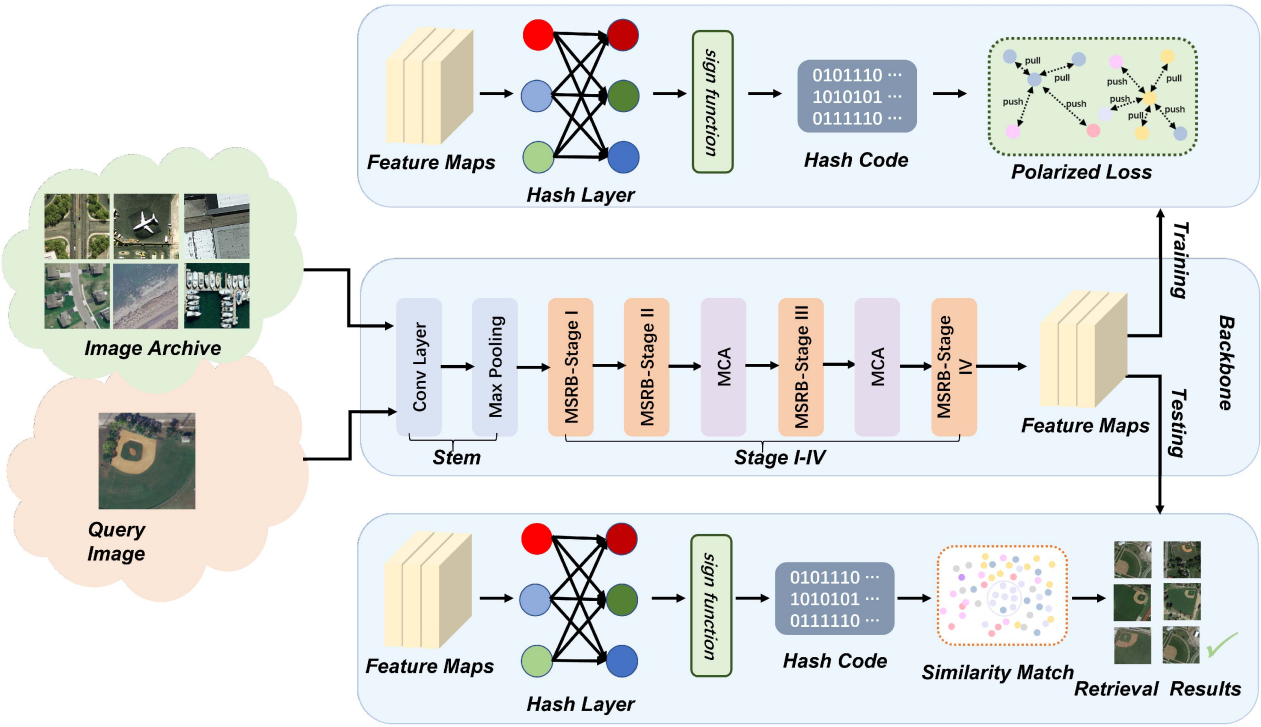


Fig. 3. Framework of MSCDH.

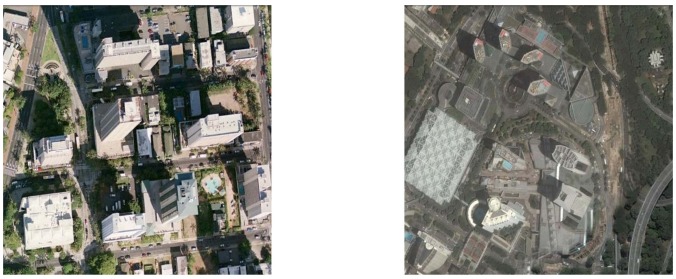


Fig. 4. RS images with large intraclass variation. (a) (b) Both belong to the commercial category but exhibit substantial visual differences.

introduced a deep hash approach for RS image retrieval. They used pretrained CNN and hash networks, and experiments on datasets of different magnitudes. Roy et al. [28] introduced a hash network based on metric learning by learning semantic-based metric spaces for generating compact hash codes to enable fast and accurate retrieval tasks. Tang et al. [29] incorporated hash learning in a generative adversarial framework to make hash codes more balanced. The performance of hash codes is optimized by optimizing the aggregation density loss function within the residual hash network. Shan et al. [30] enhanced the retrieval results by combining deep hashing methods with hard probabilistic sampling. The approach to remote sensing image retrieval called FAH [31] is proposed, which combines generative adversarial networks (GAN) with hashing methods. The entire network is divided into two modules: 1) DFLM module for extracting deep features from remote sensing images, and 2) AHLM for obtaining compact hash codes. Song et al. [32] redefined the image retrieval task as visual and semantic retrieval. They proposed a meta-hashing method [33] that utilized only a small number of labeled samples in a new category, therefore, reducing the requirement for labeled samples and improving the generalization ability of the trained hash model. Sun et al. [34] addressed the issue of traditional multiview hashing methods failing to effectively explore the latent similarity between RS images by integrating GIST features and SIFT-based BOW features of the RS images. Shen et al. [35] investigated the significance of interlabel dependencies in multilabel image retrieval tasks. They proposed a deep joint image-label hashing method called Deep Co-Image-Label Hashing (DCILH), which has outperformed existing deep learning methods in the context of multilabel tasks.

B. RS Image Retrieval Based on Attention Mechanisms

Compared with natural images, RS images are characterized by complex contents and variable scales. The attention mechanism [36], [37] can acquire the weight distribution of diverse regions through learning. Focus on the key information of interest and ignore other nonkey information. The salient information of RS images can be highlighted. To solve the problem of inaccurate extraction of target features from remote sensing images due to complex backgrounds, Wang et al. [38] refined the features from the last convolutional layer using a dual attention mechanism and input Compact Bilinear Pooling, in combination with PCA downscaling, to suppress the interference of background. Liu et al. [31] completed the feature extraction of RS images from two aspects. Attention branching is used to highlight the category features belonging to different scenarios to ensure retrieval accuracy. Xiong et al. [39] added

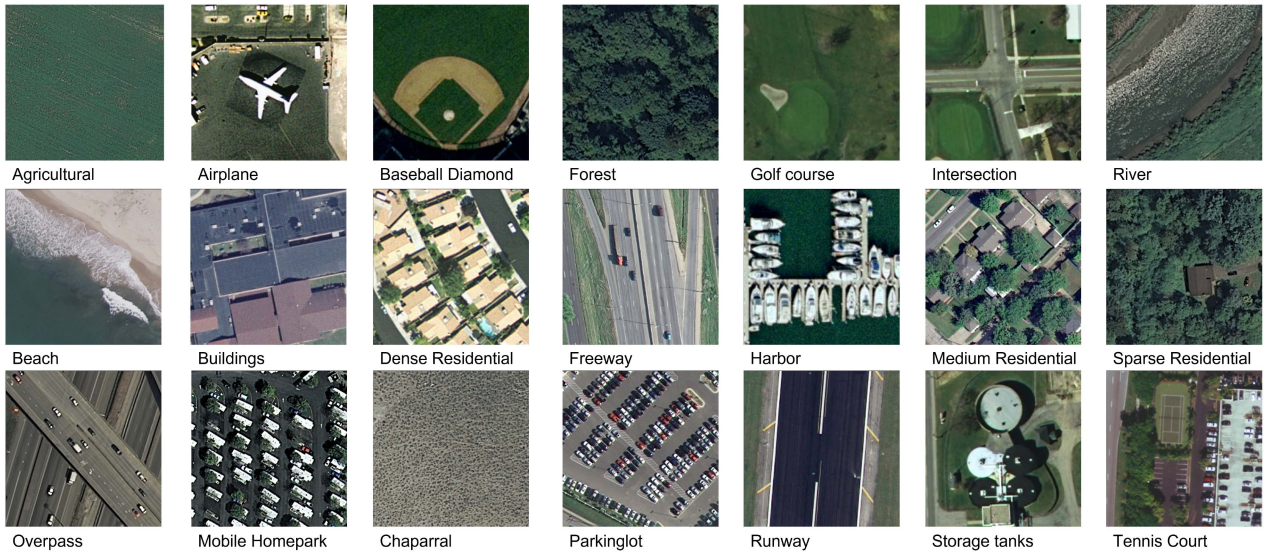


Fig. 5. Examples of the UCMerced dataset.

attention mechanisms to suppress irrelevant features. And introducing central loss in the training phase is more suitable for remote sensing image retrieval. Imbriaco et al. [40] used attention-focused local convolutional features and aggregated them using a vector of local aggregation descriptors to generate global descriptors. Therefore, our proposed MCA module combines channel attention with spatial attention to assign feature weights in a supervised manner improving the discriminative power of features on RS images.

III. Proposed Method

A. Problem We Want to Solve

With the advancement of earth observation technology and the exponential growth of RS datasets, the demands for RS image retrieval have progressively escalated. RS image content is complex and rich in detailed information. RS images are characterized by low interclass variability and low intraclass similarity. Therefore, this article proposes the MSCDH algorithm, which introduces new MSRBs. It leverages multiscale information to obtain discriminative features from RS images. We fully consider the intraclass, interclass relationships of the RS images achieving the minimization of feature distances between similar image pairs and the maximization of feature distances between dissimilar image pairs. Fig. 3 illustrates the network framework.

B. Multiscale Residual Blocks

The multiscale representation capability in convolutional neural networks is crucial. RS images manifest diverse spatial information and small-sized feature targets. Leveraging information from RS images at different scales is used to eliminate the reduction of intragroup correlation caused by insufficient local information in hash codes. We have designed a multiscale context deep hashing method addressing these issues. We adopted ResNet50 [41] as the backbone network and designed MSRBs to extract multiscale features. MSRBs can express multiscale features at a finer granularity level by grouping and merging. Fig. 7(b) shows the residual structure of ResNet, by adding a residual structure of ResNet, which incorporates a residual connection between the input and output to facilitate the propagation of information across layers and mitigate the degradation problem associated with increasing network depth. We decoupled ResNet 3×3 convolutional to make a multiscale. The structure of the MSRBs is illustrated in Fig. 7(a). The feature map of the input after 1×1 convolution is divided into s subsets, denoted by $X_i \in R^{W \times H \times C} (i = 1, 2, \cdots s)$, $W \times H$ is the spatial dimension size, and C represents the number of channels. X_i has the same spatial dimension as the input X, the channel dimension is $\frac{1}{s}$ of the original. The parameter s is used to control the scale dimension, and a larger s provides more different scales of perceptual fields. Except for X_1, each of the other feature map subsets undergoes a 3×3 convolutional layer, indicated by Q_i. The feature map subset X_i is added with Y_{i-1} and Y_{i-2} and input to Q_i to obtain the output Y. The expression defining is as follows:

$$\mathrm{Y} = \begin{cases} X_i & i = 1 \\ Q_i \left(X_i + Y_{i-1} \right) & i = 2 \\ Q_i \left(X_i + Y_{i-1} + Y_{i-2} \right) & 3 \leq i \leq s \end{cases} . \tag{1}$$

As a result of the connection operation between subsets, each 3×3 convolution layer receives information from all its previous feature map subsets. This allows the expansion of the feeling field for each subset of the previous feature maps. This structure is better suited to the characteristics of RS images and inherits the advantages of ResNet model.

C. MCA Module

Compared to ordinary images, RS images are characterized by complex content, including rich background information and

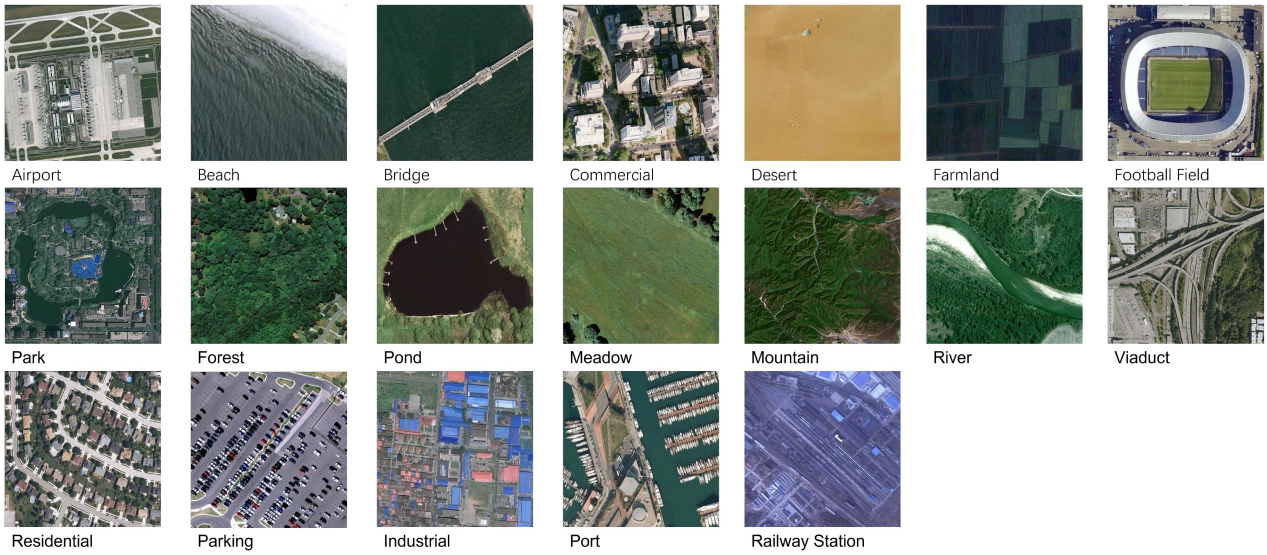


Fig. 6. Examples of the WHU-RS dataset.

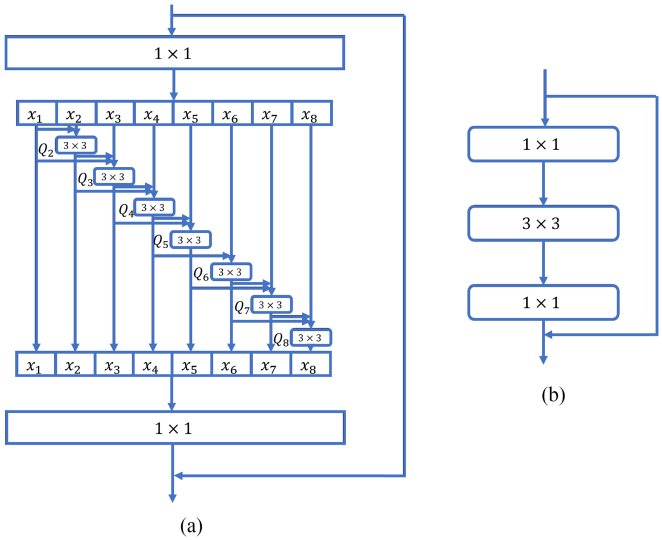


Fig. 7. The structure of MSRBs and the Residual Block in ResNet. (a) Structure of MSRBs. (b) Structure of Residual Block.

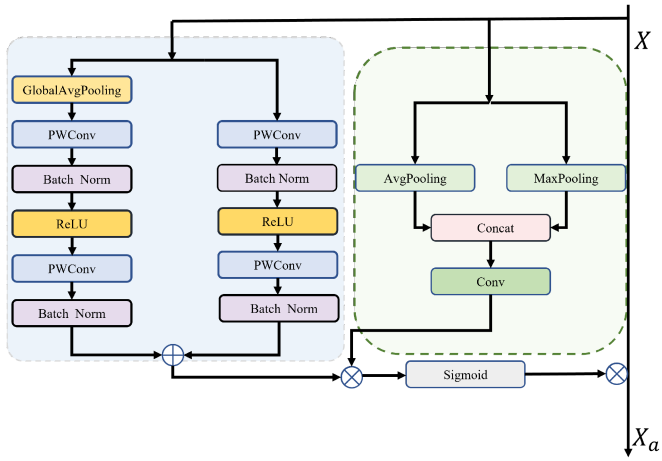


Fig. 8. Structure of the MCA module.

diverse foreground details. These factors can affect the discriminative power of convolutional neural networks in capturing high-level features. Attention mechanisms, which focus on important regions of an image, are commonly used to compensate for this limitation of CNNs. Currently, attention methods typically sample the entire region of an image. However, considering that the features extracted by CNNs from remote sensing images predominantly capture local information, we take into account the relationships among different local structures and combine the global and local features on the channel dimension to establish both global and local attention mechanisms. This approach allows for better aggregation of contextual information. Additionally, we focus on salient regions in the spatial dimension to enhance the expressive power of features. The main idea of the MCA module is to focus on the target in the region. The structure of MCA is illustrated in Fig. 8, which consists of the following two main components: 1) channel attention (left branch) and 2) spatial attention (right branch). In the channel dimension, GlobalAvgPooling and two consecutive pointwise convolutions are used to extract global feature attention. Then, the channel attention is obtained by integrating local channel attention and global attention. Finally, it can achieve aggregation of local and global contextual information. Local channel context $L_a(X)$ and the global channel context $G_a(X)$ can be expressed as

$$La(X) = BN\left(F_2\left(\delta\left(BN\left(F_1\left(GAP(X)\right)\right)\right)\right)\right), \quad (2)$$

$$Ga(X) = BN\left(F_2\left(\delta\left(BN\left(F_1(X)\right)\right)\right)\right) \quad (3)$$

where the input feature map $X \in R^{W\times H\times C}$, F_1, F_2 are pointwise convolution, GAP denotes GlobalAvgPooling, BN is Batch Norm [42], and δ denotes the sigmoid function. In the spatial dimension, the input features pass through both max pooling

and average pooling. Obtain the feature map with constant width and height and one channel number. The spatial information of the location is integrated by the information from different channels. The expression for spatial attention is

$$S(X) = \delta\left(F_3\left([AvgPool(X); MaxPool(X)]\right)\right) \tag{4}$$

where F_3 denotes 7×7 convolution. Output feature map X_a can be formulated as

$$X_a = X \otimes \sigma\left((La(X) \oplus Ga(X)) \otimes S(X)\right). \tag{5}$$

The $\otimes$ symbol indicates elementwise multiplication. The $\oplus$ denotes broadcasting addition.

D. Objective Function

Our objective in deep hashing methods is to encode images into compact hash codes while preserving the category labels and RS images consistent in Hamming space. The existing hashing methods based on pairs [43] or triples [44], [45] achieve the hash code approximation of similar images but ignore the long-tail problem of images.

To enhance the accuracy of deep hashing RS image retrieval, we use polarization loss [46] for remote sensing images with high interclass similarity and low intraclass similarity. It is a minimizable loss that can make the generated hash code have the minimum hamming distance between images of the same class while ensuring the maximum feature distance between images of different classes. Thus, the problem that the hamming distance is inconvenient to optimize is solved [47]. In MSCDH, the continuous real value $H(X)$ is obtained through the hash layer. The binary hash code $b \epsilon R^K$ is obtained by binarizing with the following sign function. It can be defined as

$$b = \text{sign}(H(x)) \tag{6}$$

where $sign()$ is the sign function. $V = H(X)$ is output vector. In order to learn the better hash function, the polarization loss is as follows:

$$\mathcal{L}_{\text{p}}(\mathbf{V}, \mathbf{T}) := \sum_{i=1}^{K} \max\left(M - V_i \cdot T_i, 0\right). \tag{7}$$

The magnitudes of each MSCDH output channel are induced over the threshold M while corresponding signs are aligned to the target vector T to minimize the polarization loss (7) during the learning phase. Since the inner product $V_i \cdot T_i$ of a hash code is inversely proportional to the hamming distance of the hash code, it maximizes the interclass hamming distance while simultaneously minimizing the intraclass hamming distance.

IV. Experiments

A. Dataset

This study employed the following two scene-specific remote sensing image datasets.

1) UCMerced dataset [48] was manually extracted from the USGS National Map Urban Area Imagery series of large images for urban areas across the country. It contains 2100 aerial images from 21 different land cover categories, in fact each category includes 100 images, each with a pixel size of 256×256 and a spatial resolution of 0.3 meters. Fig. 5 examples the UCMerced dataset.
2) WHU-RS [49] dataset was obtained from Google Earth. This data include 19 types of common RS images, such as airports, beaches, rivers, etc. Each image is 600×600 pixels in size. Some of the samples are shown in Fig. 6.

These dataset images are rich in variability in translation, spatial resolution, viewpoint than other datasets.

B. Evaluation Metrics

1) Precision@k in image retrieval tasks is defined as the proportion of retrieved results that are relevant to the query result [50], [51], [52]. The calculation formula is as follows:

$$\text{Precision@ K} = \frac{N}{K} \tag{8}$$

where N denotes the quantity of similar samples, and K denotes the count of images ranked in the top K list.
2) Mean Average Precision (MAP) [53], [54] is a comprehensive metric employed to assess the effectiveness of RS image retrieval. The calculation formula is as follows:

$$\text{MAP} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{n_i} \sum_{j=1}^{n_i} \text{precision}(R_{ij}) \tag{9}$$

where $q_i \in Q$ is a query image, and n_i is the number of image related to images q_i in the database. Suppose that the images related to samples are ranked $\{r_1, r_2, \ldots, r_{n_i}\}$ according to the correlation degree. R_{ij} is a collection of sorted search results from the top result to r_j.

These assessments, MAP and Precision@k, quantitatively evaluate the performance of RS image retrieval. It reflects the overall ranking of relevant images in the retrieval results. If the numerical values are larger, it indicates that the retrieval results of the method are better [55], [56].

C. Detailed Implementation of MSCDH

The proposed algorithm is implemented under the MindSpore and Pytorch framework. To assess the validity of the MSCDH method employed in this article, it is compared with various deep hashing methods. The learning rate is tuned to 0.001 and the batch size is specified as 32 for the model. We have experimentally validated that setting the value of s to 8 yields favorable results. We randomly selected the UCMerced and WHU-RS datasets for training and all the samples were set between 0.8 and 0.5. The value of epochs was set to 500 and we evaluated the performance of 16, 32, and 64-bit hash codes. We have set the preset target binary code T using random assignment to ensure sufficient interclass distance [57], [58].

D. Evaluation of Different Parts

We evaluate MSCDH by the analysis of the following: 1) MSRBs; 2) MCA module; and 3) Polarization loss. The experiment was conducted in the following three aspects.

TABLE I
MAP OF MSCDHSC, RESNET50 ON UCMERCED DATASET WITH DIFFERENT HASH BITS

Dataset	Method	16 b	32 b	64 b
UCMerced	MSCDH_SC	0.9654	0.9744	0.9870
	ResNet50	0.9421	0.9588	0.9651
WHU-RS	MSCDH_SC	0.9521	0.9613	0.9685
	ResNet50	0.9286	0.9342	0.9356

TABLE II
TOP K RESULTS OF THE UCMERCED DATASET USING DIFFERENT ATTENTION MODULES

Methods	P10	P20	P50
MSCDH(our method)	0.9985	0.9970	0.9954
MSCDH_SC_CBAM	0.9784	0.9762	0.9753
MSCDH_SC_GCBlock	0.9893	0.9834	0.9819
MSCDH_SC_CCA	0.9653	0.9631	0.9608

TABLE III
MAP OF THE UCMERCED DATASET USING DIFFERENT ATTENTION MODULES

Methods	16 b	32 b	64 b
MSCDH(our method)	0.9764	0.9851	0.9963
MSCDH_SC_CBAM	0.9421	0.9539	0.9628
MSCDH_SC_GCBlock	0.9563	0.9614	0.9735
MSCDH_SC_CCA	0.9353	0.9415	0.9528

First, MSDCH is named MSCDHSC without using MCA. The MSCDHSC using the residual structure was compared to the original ResNet50. The hash learning processes all use polarization loss, where the hyperparameter M is set to 0.5. Table I shows the experimental results on the UCMerced and WHU-RS datasets using MSCDHSC and ResNet50 as the backbone network. Our proposed method improves MAP from 0.9651 to 0.9870 on the UCMerced dataset, an improvement of 2.19%. The WHU-RS dataset has also been enhanced in terms of MAP. Table I illustrates that the fine-grained features of remotely sensed images can be better obtained by using a network with MSRBs. It indicates that the MSRBs are more suitable for the information-rich RS images.

Furthermore, to validate the effectiveness of our multiscale attention mechanism, we compared it with several commonly used attention modules, including CBAM [59], GCBlock [60], and CCNet [61]. CBAM is a lightweight attention module that employs channel attention and spatial attention in a sequential manner. GCBlock combines NL block and SE block attention structures. CCNet utilizes Criss-Cross Attention modules to capture contextual information. The MSCDH network without the attention module is named as MSCDHSC. Table II provides the precision@K (K=10, 20, 50) retrieval results for the UCMD dataset. It can be observed that our proposed attention module is more effective in retrieving RS images. Tables III and IV present the mean average precision results of the WHU-RS dataset and UCMD dataset. Table IV presents a comparison of mean precision at different bit lengths for different attention mechanisms. Similar retrieval performance can be observed, and our proposed attention mechanism achieves better performance compared to the other attention mechanisms. The MAP of our method improved by 3.35% compared to CBAM on UCMerced dataset. MSCDH retrieval results show a 0.93% improvement compared to MSCDHSC on UCMerced dataset. Our attention mechanism weights feature from both channel and spatial perspectives, enhancing the discriminative ability of features and yielding more accurate retrieval results.

TABLE IV
MAP OF THE WHU-RS DATASET USING DIFFERENT ATTENTION MODULES

Methods	16 b	32 b	64 b
MSCDH(our method)	0.9687	0.9842	0.9876
MSCDH_SC_CBAM	0.9384	0.9415	0.9502
MSCDH_SC_GCBlock	0.9457	0.9496	0.9564
MSCDH_SC_CCA	0.9228	0.9312	0.9358

TABLE V
RETRIEVAL RESULTS OF THE UCMERCED DATASET

Methods	16 b	32 b	64 b
MSCDH(our method)	0.9764	0.9851	0.9963
AHCL	0.9709	0.9762	0.9854
DHCNN	0.9682	0.9718	0.9822
ADSH	0.9651	0.9689	0.9810
DHNNs-L2	0.9232	0.9569	0.9649
FAH	0.9010	0.9561	0.9653
DPSH	0.8382	0.9135	0.9225

TABLE VI
RETRIEVAL RESULTS OF THE WHU-RS DATASET

Methods	16 b	32 b	64 b
MSCDH(our method)	0.9687	0.9842	0.9876
AHCL	0.9661	0.9811	0.9843
DHCNN	0.9412	0.9694	0.9743
ADSH	0.9334	0.9494	0.9739
DHNNs-L2	0.8923	0.9243	0.9502
FAH	0.7776	0.9508	0.9649
DPSH	0.7245	0.7941	0.8532

According to the objective function, parameter M is utilized to restrict the distance among hash codes. We investigated the impact of parameter M on retrieval performance while setting the hash code length to 16, 32, and 64 b. Specifically, we varied the values of m from 0.1 to 3 in our experiments. The retrieval MAP for the UCMerced dataset is shown in Fig. 9 illustrates MAP for the WHU-RS dataset. It can be observed that when m is set to 0.5, the best retrieval performance is achieved on both datasets.

E. Results

To assess the validity of hash codes in our proposed MSCDH, we compared it with six nearest deep hashing methods, including AHCL [62], ADSH [63], DPSH [43], FAH [31], DHCNNs-L2 [27], and DHCNN [32]. Tables V and VI present the retrieval results of different algorithms on UCMerced, WHU-RS datasets with varying hash code lengths. On the UCMerced dataset, our method achieved a 1.09% improvement in MAP compared to the state-of-the-art AHCL method on 64-b hash codes. It can be observed that different lengths and hash codes have varying effects on the representation of features. Short hash codes have limited expressive power, while longer hash codes result in better representation performance.

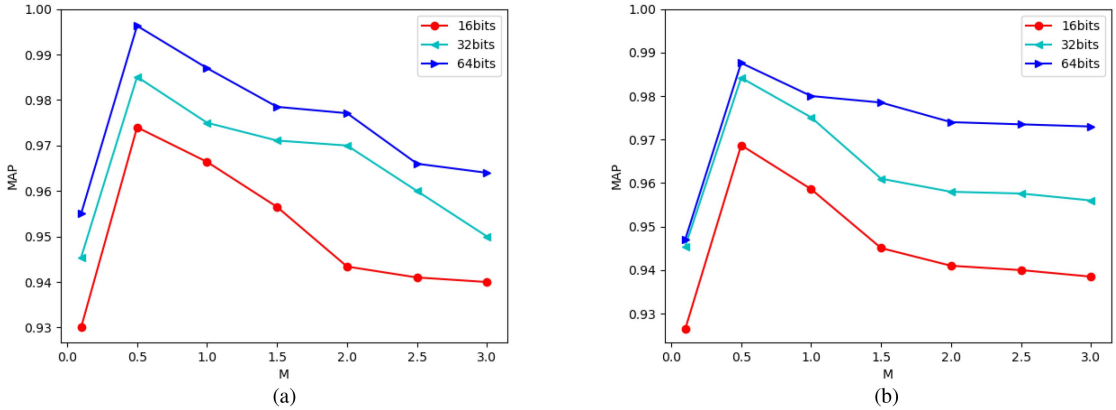


Fig. 9. Comparing hashing accuracy with different margin M settings in terms of MAP on Dataset. (a) UCMerced Dataset. (b) WHU-RS Dataset.

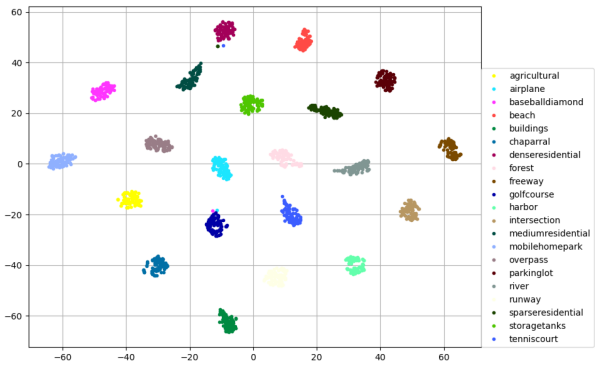


Fig. 10. 2-D scatter plot obtained by t-sne on the UCMerced dataset.

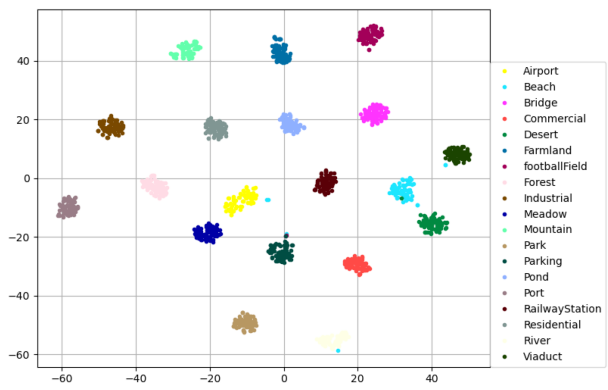


Fig. 11. 2-D scatter plot obtained by t-sne on the WHU-RS dataset.

In addition, we also analyze the clustering properties of hash codes using the t-SNE dimensionality reduction algorithm. It projects the original high-level data into a 2-D space for visualization, while the low-dimensional features processed by it can better reflect the distance distribution of the original data in high-dimensional space. Fig. 10 shows the t-SNE visualization of the 64-bit hash codes of the UCMerced dataset using the MSCDH method. Fig. 11 shows the t-SNE visualization of the 64-bit hash codes generated from the WHU-RS dataset. It can be observed that the hash codes learned by MSCDH have an easily discriminable distribution, and the hash codes of similar images are clustered together. Each class can be well separated to achieve the effect of minimizing the intraclass distance.

V. Conclusion

RS image retrieval technology is the foundation and prerequisite for many applications in the field of remote sensing. Therefore, this article addresses the problems of RS image retrieval field such as large-scale variation and large differences of similar targets. We propose the multiscale context deep hashing RS retrieval method. First, we used multiscale residual blocks to extract the multiscale features of RS images. Then, MCA mechanism is proposed to capture contextual information to weight salience features in channel and spatial dimensions. Further, a microscopic polarizable loss is used to maintain the balance and differentiation of hash codes during hash learning improving the robustness of the model in the case of complex background of remote sensing images. Experiments were conducted on the UCMerced and WHU-RS datasets to verify the validity of MSCDH.

In practical applications, supervised hashing methods may be constrained by large data volumes or limited label quantities. Moreover, they demand high accuracy and consistency in labeling. The presence of noise or errors in the labels can directly impact the learning and retrieval performance of the hashing function. Our next research goal is to develop a deep hashing model to address the supervised RS image retrieval problem.

References

[1] L. Liu, Y. Wang, J. Peng, and A. Plaza, "DFLR: Deep feature learning with latent relationship embedding for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, Jul. 2021.

[2] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 8, p. 1243, 2018.

[3] L. Wu, R. Jin, and A. K. Jain, "Tag completion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 716–727, Mar. 2013.

[4] 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.

[5] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.

[6] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proc. ACM multimedia*., 1997, pp. 65–73.

[7] T. Bretschneider, R. Cavet, and O. Kao, "Retrieval of remotely sensed imagery using spectral information content," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2002, vol. 4, pp. 2253–2255.

[8] Y. Hongyu, L. Bicheng, and C. Wen, "Remote sensing imagery retrieval based-on Gabor texture feature classification," in *Proc. Int. Conf. Signal Process.*, 2004, vol. 1, pp. 733–736.

[9] G.-S. Xia, J. Delon, and Y. Gousseau, "Shape-based invariant texture indexing," *Int. J. Comput. Vis.*, vol. 88, pp. 382–403, 2010.

[10] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[11] 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.

[12] P. Agouris, J. Carswell, and A. Stefanidis, "An environment for content-based image retrieval from large spatial databases," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 263–272, 1999.

[13] J. Yang, J. Liu, and Q. Dai, "An improved bag-of-words framework for remote sensing image retrieval in large-scale image databases," *Int. J. Digit. Earth*, vol. 8, no. 4, pp. 273–292, 2015.

[14] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating Local Descriptors Into a Compact Image Representation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit*., 2010, pp. 3304–3311.

[15] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4355–4369, May 2021.

[16] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2810–2825, Feb. 2021.

[17] S. Har-Peled and N. Kumar, "Approximate nearest neighbor search for low-dimensional queries," *SIAM J. Comput.*, vol. 42, no. 1, pp. 138–159, 2013.

[18] J. Wang et al., "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, Apr. 2018.

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

[20] 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.

[21] 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.

[22] 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, 2017.

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

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

[25] Y. Liu, L. Ding, C. Chen, and Y. Liu, "Similarity-based unsupervised deep transfer learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7872–7889, Nov. 2020.

[26] Z. Xiao, Y. Long, D. Li, C. Wei, G. Tang, and J. Liu, "High-resolution remote sensing image retrieval based on CNNS from a dimensional perspective," *Remote Sens.*, vol. 9, no. 7, p. 725, 2017.

[27] 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.

[28] 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. Dig. Int. Geosci. Remote Sens. Symp*., 2018, pp. 4539–4542.

[29] X. Tang, C. Liu, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Large-scale remote sensing image retrieval based on semi-supervised adversarial hashing," *Remote Sens.*, vol. 11, no. 17, p. 2055, 2019.

[30] X. Shan, P. Liu, G. Gou, Q. Zhou, and Z. Wang, "Deep hash remote sensing image retrieval with hard probability sampling," *Remote Sens.*, vol. 12, no. 17, p. 2789, 2020.

[31] C. Liu, J. Ma, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Deep hash learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3420–3443, Apr. 2021.

[32] W. Song, S. Li, and J. A. Benediktsson, "Deep hashing learning for visual and semantic retrieval of remote sensing images," *IEEE Trans. Geosci. Remote Sens*, vol. 59, no. 11, pp. 9661–9672, Nov. 2021.

[33] X. Tang et al., "Meta-hashing for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, Dec. 2021.

[34] Y. Sun, W. Wu, X. Shen, and Z. Cui, "Multiview inherent graph hashing for large-scale remote sensing image retrieval," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 14, pp. 10705–10715, Oct. 2021.

[35] X. Shen, G. Dong, Y. Zheng, L. Lan, I. W. Tsang, and Q. Sun, "Deep co-image-label hashing for multi-label image retrieval," *IEEE Trans. Multim.*, vol. 24, pp. 1116–1126, Oct. 2022.

[36] X. Zheng, H. Sun, X. Lu, and W. Xie, "Rotation-invariant attention network for hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 31, pp. 4251–4265, May 2022.

[37] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 58, no. 5, pp. 3232–3245, Nov. 2019.

[38] Y. Wang, S. Ji, M. Lu, and Y. Zhang, "Attention boosted bilinear pooling for remote sensing image retrieval," *Int. J. Remote Sens.*, vol. 41, no. 7, pp. 2704–2724, 2020.

[39] W. Xiong, Y. Lv, Y. Cui, X. Zhang, and X. Gu, "A discriminative feature learning approach for remote sensing image retrieval," *Remote Sens.*, vol. 11, no. 3, p. 281, 2019.

[40] R. Imbriaco, C. Sebastian, E. Bondarev, and P. H. de With, "Aggregated deep local features for remote sensing image retrieval," *Remote Sens.*, vol. 11, no. 5, p. 493, 2019.

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

[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

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

[44] X. Wang, Y. Shi, and K. M. Kitani, "Deep supervised hashing with triplet labels," in *Proc. Asian Conf. Comp. Vision*., 2017, pp. 70–84.

[45] B. Zhuang, G. Lin, C. Shen, and I. Reid, "Fast training of triplet-based deep binary embedding networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2016, pp. 5955–5964.

[46] L. Fan, K. W. Ng, C. Ju, T. Zhang, and C. S. Chan, "Deep polarized network for supervised learning of accurate binary hashing codes," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 825–831.

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

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

[49] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," 2009.

[50] X. Lu, Y. Chen, and X. Li, "Discrete deep hashing with ranking optimization for image retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 2052–2063, Jun. 2020.

[51] Y. Chen and X. Lu, "Deep discrete hashing with pairwise correlation learning," *Neurocomputing*, vol. 385, pp. 111–121, 2020.

[52] Y. Chen, X. Lu, and X. Li, "Supervised deep hashing with a joint deep network," *Pattern Recognit.*, vol. 105, 2020, Art. no. 107368.

[53] Y. Chen and X. Lu, "Deep category-level and regularized hashing with global semantic similarity learning," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 6240–6252, Dec. 2021.

[54] H. Ning, B. Zhao, and Y. Yuan, "Semantics-consistent representation learning for remote sensing image–voice retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, Mar. 2021.

[55] Y. Chen, X. Lu, and Y. Feng, "Deep voice-visual cross-modal retrieval with deep feature similarity learning," in *Proc. IEEE Pattern Recognition. Comput. Vis.*, 2019, pp. 454–465.

[56] M. Guo, C. Zhou, and J. Liu, "Jointly learning of visual and auditory: A new approach for RS image and audio cross-modal retrieval," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 11, pp. 4644–4654, Nov. 2019.

[57] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. 30th Annu. ACM Symp. Theory Comput.*, 1998, pp. 604–613.
[58] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Commun. ACM*, vol. 51, no. 1, pp. 117–122, 2008.
[59] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional Block Attention Module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[60] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1971–1980.
[61] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proc IEEE Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
[62] W. Song, Z. Gao, R. Dian, P. Ghamisi, Y. Zhang, and J. A. Benediktsson, "Asymmetric hash code learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, Jan. 2022.
[63] Q.-Y. Jiang and W.-J. Li, "Asymmetric deep supervised hashing," in *Proc. AAAI Conf. Artif. Intell*, vol. 32, no. 1, pp. 3342–3349, 2018.

Dongjie Zhao received the bachelor of engineering degree in software engineering from Northwest Normal University, Lanzhou, China, in 2021. She is currently working toward the master of engineering degree in computer science and technology with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China.

Her main research interests include image retrieval and style transfer.

Yaxiong Chen received the B.Sc. degree in mathematics from Hubei University, China, in 2014. He received the Master of Science in mathematics from Wuhan University of Technology, China, in 2017. He received the Ph.D. degrees in signal and information processing from University of Chinese Academy of Sciences, China, in 2020. he is an Associate Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China. His current research interests include pattern recognition, machine learning, hyperspectral image analysis, and medical imaging.

Shengwu Xiong received the B.Sc. degree in computational mathematics from Wuhan University, Wuhan, China, in 1987, and the M.Sc. and Ph.D. degrees in computer software and theory from Wuhan University, in 1997 and 2003, respectively.

He is currently a Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, China. His research interests include intelligent computing, machine learning, and pattern recognition.