A Deep Cross-Modality Hashing Network for SAR and Optical Remote Sensing Images Retrieval

Wei Xiong, Zhenyu Xiong, Yang Zhang, Yaqi Cui, and Xiangqi Gu

***Abstract*—The content-based remote sensing image retrieval (CBRSIR) has recently become a hot topic due to its wide applications in analysis of remote sensing data. However, since conventional CBRSIR is unsuitable in harsh environments, this article focuses on the cross-modality CBRSIR (CM-CBRSIR) between synthetic aperture radar (SAR) and optical images. Besides the large interclass and small intraclass in CBRSIR, CM-CBRSIR is limited by prominent modality discrepancy caused by different imaging mechanisms. To address this limitation, this study proposes a deep cross-modality hashing network. First, we transform optical images with three channels into four different types of single-channel images to increase diversity of the training modalities. This helps the network to mainly focus on extracting the contour and texture shared features and makes it less sensitive to color information for images across modalities. Second, we combine any type of randomly selected transformed images and its corresponding SAR or optical images to form image pairs that are fed into the networks. The training strategy, with paired image data, eliminates the large cross-modality variations caused by different modalities. Finally, the triplet loss, in combination with the hash function, helps the modal to extract the discriminative features of images and upgrade the retrieval efficiency. To further evaluate the proposed modality, we construct a SAR-optical dual-modality remote sensing image dataset containing 12 categories. Experimental results demonstrate the superiority of the proposed method with regards to efficiency and generality.**

***Index Terms*—Cross-modality content-based remote sensing image retrieval (CM-CBRSIR), deep cross-modality hashing network (DCMHN), modality discrepancy, synthetic aperture radar (SAR)-optical dual-modality remote sensing image dataset (SODMRSID).**

I. Introduction

UNPRECEDENTED advances in earth observation technologies, over the past few decades, have caused a significant increase in both quality and quantity of remote sensing image archives [1], [2]. Generally, content-based remote sensing image retrieval (CBRSIR), which is simply defined as the search for remote sensing images of similar information content within a large archive with a given query image serving as a reference, has attracted numerous research interest due to its broad applications in management of large volumes of remote sensing data. In the field of natural images, previous studies have proposed concerted efforts for improving image retrieval tasks [3], [4]. However, unlike natural images, remote sensing images contain very small and intricate targets, making retrieval of discriminative features difficult. Advancement of the convolutional neural network (CNN) has tremendously improved both accuracy and efficiency of CBRSIR retrieval [5]–[7]. Images from optical sensors considerably limit the application of CBRSIR, since the optical sensors only function well during day time and fine weather, but not at night or under bad weather scenarios.

Synthetic aperture radar (SAR) images have several advantages, including excellent functioning at all times, and under all weather conditions. However, they present numerous limitations including low resolution, side-looking imaging, blurred target details, need for visual interpretation, and lack of wide range of target detection. On the other hand, optical remote sensing images have many advantages over SAR images. For instance, they are intuitionistic and easy to understand, have rich color and texture information, present obvious target structure characteristics, high resolution, and a large field angle. However, optical images are also greatly affected by light, cloud cover, seasons, shadows, and other conditions, hence there is a need to complement them with SAR images to ensure adequate exploitation of the aforementioned strengths. Therefore, a retrieval system that can retrieve an image across optical and SAR sensors, would operate well in almost any real-world condition. However, the modality disparity caused by the different imaging mechanisms between the two sensors complicates the retrieval task.

Rapid development of feature learning has accelerated exploration of cross-modality retrieval tasks in the field of natural image analysis. These include retrieval between image and text [8], [9], image and audio [10], [11], as well as RGB and infrared images [12], [13]. However, these methods reportedly yield unsatisfactory results when applied to remote sensing images, owing to huge differences between natural and remote sensing images. In addition, only a handful of works [14]–[16] have reported use of CM-CBRSIR, which allows sensing between panchromatic and multispectral sources. Besides, SAR images lack the specific imaging principle and presence of speckle noise, as well as the rich color information contained in optical images (see Fig. 1). Based on these factors, the existing works cannot

Manuscript received June 13, 2020; revised August 14, 2020; accepted August 31, 2020. Date of publication September 3, 2020; date of current version September 17, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61790550, Grant 61790554, and Grant 91538201 *(Corresponding author: Zhenyu Xiong.)*

The authors are with the Research Institute of information Fusion, Naval Aviation University, Yantai 264001, China (e-mail: xiongwei@csif.org.cn; x_zhen_yu@163.com; 337393724@qq.com; cui_yaqi@126.com; guxiangi 1314@163.com).

Digital Object Identifier 10.1109/JSTARS.2020.3021390

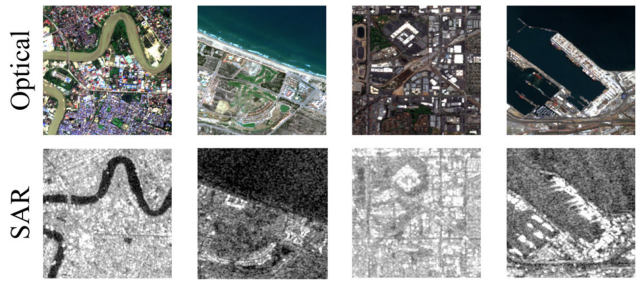


Fig. 1. Sample images from our proposed SODMRSID datasets.

effectively be extended to image retrieval between SAR and optical images.

To solve the aforementioned CM-CBRSIR challenges, we propose a deep cross-modality hashing network (DCMHN). First, to extract rich image features from different spectrum channels, we transform optical images with RGB three channels into four types of single-channel images via different spectrum channels including red, blue, green, and gray. Second, to solve the cross-modality discrepancy caused by imaging mechanisms, we randomly select four types of the transformed single-channel images to form image pairs with each corresponding SAR or optical images. Finally, the triplet loss combined with the hash function helps the modal to extract the discriminative features of images and upgrade the retrieval efficiency. We propose a new SAR-optical dual-modality remote sensing image dataset (SODMRSID), comprising 12 categories, owing to absence of any open source cross-modality dataset between SAR and optical image for retrieval tasks. To validate the effectiveness of this method, we perform extensive experiments on the proposed SODMRSID.

The main contributions of this article can be summarized as follows.

1) To the best of our knowledge, this is the first study conducting CM-CBRSIR between SAR and optical sensors, and proposes the possibility and potential values of CM-CBRSIR.
2) We provide an end-to-end framework for CM-CBRSIR, coupled with good flexibility and controllability. Besides, the model, proposed herein, is applicable to other cross-modality tasks in the field of remote sensing images.
3) We provide a large-scale benchmark dataset, named SODMRSID, that can be used to evaluate the proposed method and largely advance the task of cross-modality image processing technology between SAR and optical sensors.

The rest of this article is organized as follows: Section II reviews existing literature related to CBRSIR, supervised cross-modality hash methods, and retrieval of cross-modality in remote sensing; Section III describes the SAR-optical dual-modality remote sensing image dataset (SODMRSID); Section IV presents our proposed DCMHN, Section V outlines the experimental results and analyses; and Section VI concludes this article.

II. Related Work

A. Content-Based Remote Sensing Image Retrieval (CBRSIR)

The most important part of CBRSIR entails extracting the effective features of images. However, a key challenge to this involves designing a robust feature extractor that can accommodate the diversity of remote sensing image types as well as the complexity of remote sensing image content. Most existing feature extractors are based on low-level visual features, including global features related to spectral (color) [17], texture [19], shape [22], as well as local features based on scale invariant feature transform (SIFT) [23], difference of Gaussian (DoG) [24], and speeded up robust features (SURF) [25]. To represent the highly complex remote sensing images, most approaches produce more discriminative features by aggregating local features, such as bag-of-words (BoW) [26], a vector of locally aggregated descriptors (VLAD) [27], and Fisher vector (FV) [28] or their variants. However, the features obtained by these methods are handcrafted, requiring sufficient domain expertise and engineering skills. Generally, handcrafted features cannot accurately describe the rich information of the images due to the complex background of the remote sensing image. Specifically, the same class of remote sensing images might have a diverse appearance [29]. Numerous studies have shown superior performances to the traditional handcrafted features in CBRSIR, based on the great success of the CNN in representing high-level visual features of images [31]–[38]. Similarly, numerous studies are underway to increase retrieval accuracy by extracting more discriminative features of the images, considering the small and intricate targets contained in the remote sensing images. For instance, some researchers have combined the attention mechanism with multitask learning to extract discriminative features of the remote sensing image [39]. On the other hand, a deep hashing neural network was successfully used to solve CBRSIR in a large-scale dataset and transform the image feature to binary codes [40]. Moreover, both deep semantic features and weighted distance were reportedly used to successfully construct a retrieval framework and improve performance [41]. To further cope with large-scale complex retrieval problems in remote sensing, a two-steps strategy was reportedly used to obtain multihash codes, achieving a high retrieval accuracy over a short period of time [42]. Besides, a novel multilabel method based on fully convolutional network [67] is used for CBRSIR task, which shows great advantages over some single-label methods for interpreting complex remote sensing images.

The aforementioned methods were all aimed at solving the optical image retrieval tasks. Based on the specific image content of SAR images, several methods have been conducted to improve retrieval performance of SAR images. For example, a compression-based image retrieval technique was previously designed for measuring similarities between SAR on the original and despeckled TerraSAR-X images [43]. In addition, a general SAR image retrieval approach was developed, according to the region-based similarity measure and semantic categorization [44], whereas an image reranking method was used to improve the retrieval accuracy of SAR images [45]. Moreover, multiscale

property and speckle noise were successfully applied to help in designing a content-based SAR image retrieval method [46], whereas the fly algorithm, based on hash codes, effectively improved the retrieval speed and reduce the storage cost for the SAR image retrieval task [47]. In addition, an unsupervised domain adaption model for SAR image retrieval reportedly coped with the shortage of labeled SAR images [48]. Despite these methods adequately addressing the single-modality CBRSIR for SAR sensors, they cannot be effectively extended to CM-CBRSIR.

B. Supervised Cross-Modality Hash Methods

Hashing methods have attracted considerable attention due to their low storage costs and fast retrieval speed. However, constructing semantic correlations among heterogeneous features from different modalities with the binary hash codes remains the method's main challenge. Semantic correlation maximization (SCM), which integrates semantic label information into hashing codes, was previously used to reduce the storage cost and improve the query speed [49]. Similarly, studies have successfully applied semantic information and manifold structure of data to reveal the association among heterogeneous modalities [50]. Apart from these, a discrete method was used to improve accuracy and training speed by directly learning the binary hash codes [51], with the discriminative hash codes produced by learning the modality-specific hash functions [52]. Since these traditional methods mainly use hand-crafted features to learn binary vectors, their performance in real-world applications is limited by the independent feature extraction process.

Generally, deep cross-modality hashing methods are superior to traditional cross-modality ones, owing to their powerful feature representation capability. To learn the modality-specific information, an end-to-end deep learning architecture, which generates compact hash codes, has been previously used [53]. However, this architecture cannot be extended to other cross-modal cases. Deep cross-modal hashing (DCMH), which utilizes both hash codes and feature learning strategies and can be optimized from scratch in the same deep learning framework, has been proposed [54]. An adversarial cross-modal retrieval (ACMR) method based on the adversarial learning approach is also proposed, and shown to successfully generate discriminative and modality-invariant binary hash codes for the data across different modalities [55]. Typically, deep learning-based cross-modal hashing methods can outperform traditional ones, both on retrieval efficiency and accuracy. However, all these methods work for cross-modality retrieval tasks in natural images or documents, which are extremely different from remote sensing ones, both in spatial and spectral resolution. Therefore, the complexity of remote sensing images limits performance of these methods on remote sensing area.

C. Cross-Modality Retrieval in Remote Sensing

Rapid development of remote sensing technology has gradually increased the types of remote sensing data that can be acquired by different sensors. Consequently, cross-model retrieval has received widespread attention in recent years. Cross-modality retrieval techniques can be divided into three categories: the retrieval tasks allowing the model between remote sensing images and spoken audio [56], [57], remote sensing images and sentences [58]–[60], and panchromatic and multispectral images [14]–[16]. However, the retrieval between audio and image differs from the cross-modality retrieval between images from different modalities. There is not much connection between the two retrieval tasks, because the semantic information contained in remote sensing image exceeds that in an audio signal. Thus, the model cannot be directly applied to CM-CBRSIR between SAR and optical images. Additionally, the network structure of text feature extraction is not applicable to the images, making it unsuitable for our task. Furthermore, methods for solving the retrieval tasks between panchromatic and multispectral images pay little attention to images' texture information, which is extremely important during presentation of features from SAR images. Inspired by these studies, we adopt a DCMHN approach to solve the CS-CBRSIR task between SAR and optical images.

III. SAR-Optical Dual-Modality Remote Sensing Image Dataset

The increasing ability to acquire remote sensing data has generate numerous remote sensing image scene datasets [61]–[63], [68], [69]. These existing datasets were constructed by only one kind of remote sensing data modality. Intuitively, these single-modality datasets would not cope with the increasingly complex environmental and diverse data in real-world conditions. Consequently, cross-modality datasets, which allow modality between panchromatic and multispectral images, remote sensing images and spoken audio, as well as remote sensing images and sentences, have been proposed in [14], [57], and [58]. To promote the all-time and all-weather image retrieval system, constructing a cross-modality remote sensing image dataset between SAR and optical sensors is a priority. Therefore, we collected a new SAR-optical dual-modality remote sensing image dataset. (SODMRSID).[1]

Specifically, the SODMRSID was collected from remote sensing images captured by SAR and optical sensors. The SODMRSID comprises of a great number of patch pairs, with each patch pair representing a combination of a SAR and an optical image, covering the same area across the globe and throughout all four seasons. Notably, although they show different aspects of the captured ground region because of the different geometric and radiometric appearance, the SAR and optical images in one patch-pair represent the same type of scene. SODMRSID is constructed based on SEN1-2 [64], which comprises 2 82 384 remote sensing images acquired by Sentinel-1 [65] and Sentinel-2 [66]. Dual sample description is outlined in Table I.

The SODMRSID contains a total number of 24 000 images, covering 12 typical scene classes that include agriculture, beach, forest, harbor, industrial, lake, meadow, mountain, pond, residential area, river, and water. Each class of the image consists

[1][Online]. Available: https://pan.baidu.com/s/1xR7h-NP143Ju9chGuDBDMw with password "p1b2".

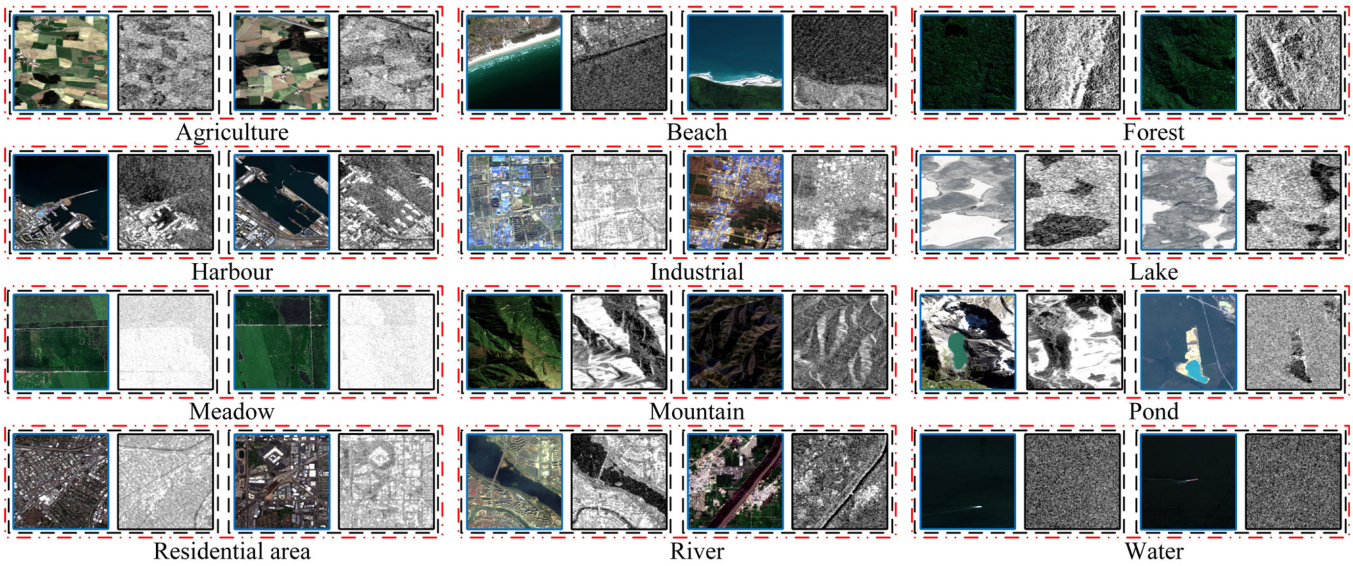


Fig. 2. Examples from the proposed SODMRSID.

TABLE I
DESCRIPTION OF THE DATASET

Data modality	Satellite sensor	Spatial resolution	Spectral channel	Image size
SAR image	Sentinel-1	10m	1	256×256
Optical image	Sentinel-2	10m	3	256×256

of 1000 SAR-optical patch-pairs, some of which are shown in Fig. 2.

The SAR-optical patch-pairs datasets are represented by $\mathbf{D} = \{(\mathbf{x}_i^{\mathrm{S}}, \mathbf{x}_i^{\mathrm{O}}, L_i) | i = 1, 2, \ldots, N\}$, where i denotes the index of patch-pairs, N denotes the SODMRSID volume, $\mathbf{x}_i^{\mathrm{S}} \in \mathrm{R}^{256\times256}$ indicates the SAR image, $\mathbf{x}_i^{\mathrm{O}} \in \mathrm{R}^{256\times256\times3}$ denotes the optical image, and L_i denotes the image label.

IV. Proposed Method

This section describes the proposed DCMHN processes, including image transformation, training with image pairs, and triplet hashing loss. The proposed method framework is shown in Fig. 3 and entails the following: First, Section IV-A describes how to increase the diversity of the input modality by transforming the three-channel optical images into four types of single-channel images; Second, Section IV-B presents how to conduct a paired training strategy to extract discriminative features of the image across the modalities; and Third, Section IV-C introduces the triplet hashing loss function to improve retrieval accuracy and deduce storage costing.

A. Image Transformation

Modality discrepancy is the key challenge restricting cross-modality retrieval task between SAR and optical sensors, since their respective images considerably differ in the same scene. Generally, optical images usually contain intensity information of multiple wave bands, which is convenient for target recognition and classification extraction. On the other hand, SAR images record echo information of only one wave band, in binary complex form. In addition, amplitude information contained in SAR images is less than the imaging level of optical images. As shown in Fig. 2, SAR images lack the rich color information in optical images. However, some scene and target-related rich information, which is contained in SAR images, such as geometric structure and material property, cannot be ignored.

Based on this observation, our network mainly focuses on the contour and texture information of SAR and optical images, and less on the color information. To achieve this goal, the network needs to learn the features of the same scene of images across different modalities. Therefore, a novel image transformation strategy, that can produce images of different spectrum channels, is proposed. By adding the cross-modalities images in the training process, the spectrum channels of the image have been disrupted, the network cannot focus on learning the color information of spectrum channels, it will pay more attention to the texture and contour part. Specially, for each optical image with RGB three channels, the transformed single-channel images of the different spectrum are obtained by selecting the corresponding spectrum channel from the original optical image. Furthermore, a grayscale image is also produced by transforming the optical image, to increase the diversity of the image modalities. In this article, the original optical image is denoted $\mathbf{x}_i^{\mathrm{O}}$, whereas it is corresponding single-channel images with red, green, and blue spectral channels are represented by $\mathbf{x}_i^{\mathrm{R}}$, $\mathbf{x}_i^{\mathrm{G}}$, $\mathbf{x}_i^{\mathrm{B}}$, respectively. The corresponding grayscale image is denoted $\mathbf{x}_i^{\mathrm{H}}$. A few examples can be found in Fig. 4.

Using the transformed images, as input in the network, significantly increases diversity of the training data, while the spectrum channels are disrupted by training with different

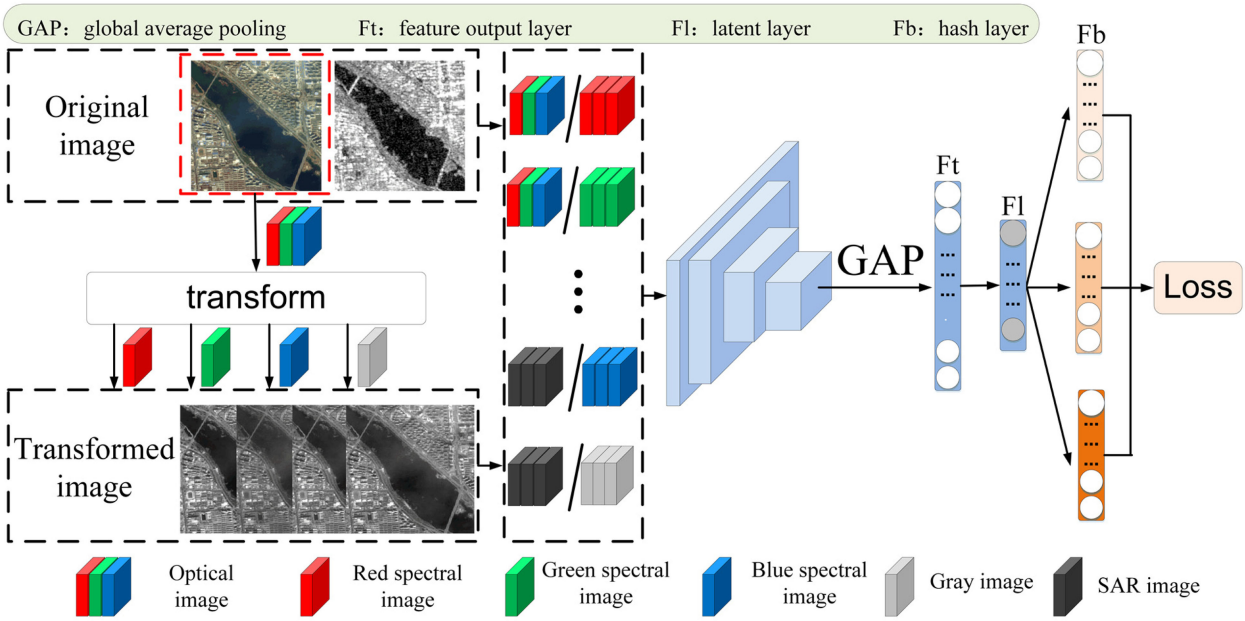


Fig. 3. Framework of our proposed method.

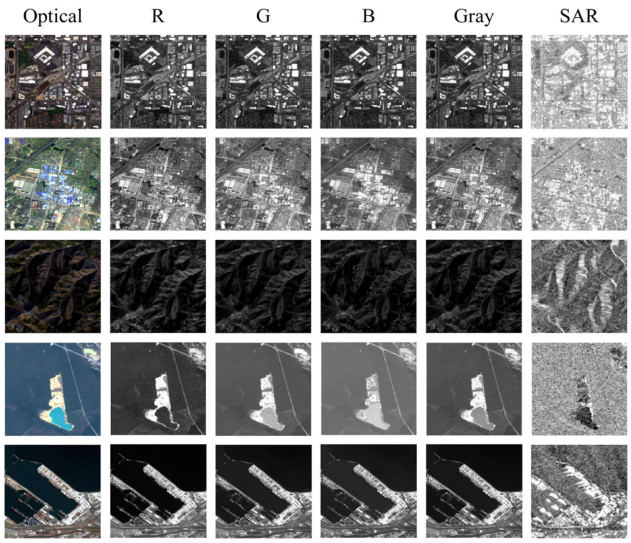


Fig. 4. Examples of the transformed images.

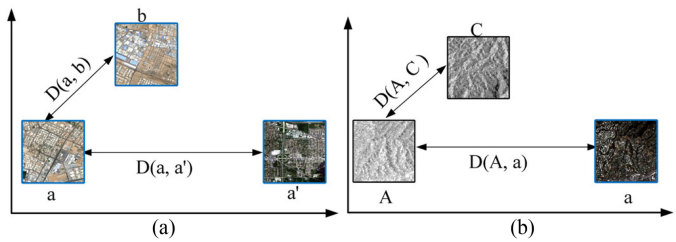


Fig. 5. Hard negatives caused by modality discrepancy. (a) "a" and "a′" represent the different images of the residential area category captured by an optical sensor, whereas "b" represents the image of the industrial category captured by optical sensor. (b) "a" and "A" represents the images of the residential area category captured by optical and SAR sensors, respectively. "C" represents the image of the mountain category captured by SAR sensors.

spectrum images. Thus, the network focuses on contour and texture information, but not color information. Moreover, adding the transformed images into the training data largely increases diversity of the modalities, making the networks to have more image modalities under the same scene. Thus, the network learns some shared features across different modalities, thereby significantly reducing the influence of the modality discrepancy.

B. Training With Image Pairs

Remote sensing images always suffer from hard negatives owing to the complex content of images and modality discrepancy [see Fig. 5(a)], where intraclass distance is often larger than interclass distance $(D(a, a') > D(a, b))$. Besides, it is more obvious when the images are from different modalities $(D(A, a) > D(A, C))$ [see Fig. 5(b)]. Therefore, these problems negatively impact the accuracy and efficiency of cross-modality retrieval, resulting in poor performance during the training process. Extraction of robust and discriminative features across different modalities requires an urgent solution.

To address this problem, we propose a novel paired training strategy (see Fig. 3). Specifically, for each input original image, we randomly choose one image from all the transformed images $\mathbf{x}_i^{\mathrm{R}}$, $\mathbf{x}_i^{\mathrm{G}}$, $\mathbf{x}_i^{\mathrm{B}}$, and $\mathbf{x}_i^{\mathrm{H}}$ for its corresponding optical image $\mathbf{x}_i^{\mathrm{O}}$ and SAR image $\mathbf{x}_i^{\mathrm{S}}$, as the image pairs before feeding them together into the network. Since the number of spectrum channels between optical and transformed images is different, channels for the transformed images and SAR images are triplicated to create a three-channel image, making them similar to optical images with three channels. Consequently, the networks for both modalities begin with a uniform architecture.

In addition, training with image pairs strategy significantly eliminates the intra- and inter-modality variations caused by modality discrepancy, thereby propelling the network to extract discriminative features across different modalities.

C. Triplet Hashing Loss

To extract powerful discriminative features and improve retrieval efficiency, we adopt the hashing-based triplet loss function to constrain and encode intraclass images as closely as possible, while encoding interclass ones as far apart in the feature space as possible.

During the training process, Resnet18 [20] and Resnet50 [20] are used as feature extractors, to obtain the feature maps of the input images after the last convolutional layer. Thereafter, a global average pooling (GAP) layer is applied to extract the unified features from the images. Subsequently, the latent layer $\mathrm{F_l}$ is introduced to construct the relationships between input images and binary codes, and preserve the rich semantic information of the images. For each image $\mathbf{x}_i$, the output of the latent layer is denoted $\mathbf{f}_i$, and $\mathbf{f}_{ik} \in [0,1](k=1,\ldots,K)$ represent the *k*th of the latent vector $\mathbf{f}_i$, which is activated by the sigmoid function.

Triplet loss [18], as a loss function, is then introduced to help the network to extract discriminative features. For each triplet image $\mathbf{I} = \{(\mathbf{x}_i^{\mathrm{a}}, \mathbf{x}_i^{\mathrm{p}}, \mathbf{x}_i^{\mathrm{n}})\}$, the deep features $\mathbf{T} = \{(\mathbf{f}_i^{\mathrm{a}}, \mathbf{f}_i^{\mathrm{p}}, \mathbf{f}_i^{\mathrm{n}})\}$ are obtained in the latent layer $\mathrm{F_l}$. The triplet loss is built with the intuition that an anchor image $\mathbf{x}_i^{\mathrm{a}}$ is closer to all positive images $\mathbf{x}_i^{\mathrm{p}}$, than to all negative images $\mathbf{x}_i^{\mathrm{n}}$ in the feature space. A triplet loss function is constructed as follows:

$$L_{\mathrm{Triplet}} = \sum_{i=1}^{m} \left[d\left(\mathbf{f}_i^{\mathrm{a}}, \mathbf{f}_i^{\mathrm{p}}\right) - d\left(\mathbf{f}_i^{\mathrm{a}}, \mathbf{f}_i^{\mathrm{n}}\right) + \alpha \right] \tag{1}$$

where m is the size of mini-batch, d is the similarity metric, and α is a margin that is enforced between positive and negative pairs.

Activation by a sigmoid function in the latent layer results in a value of $\mathbf{f}_{ik}$ that lies between 0 and 1. Inspired by [21], we design a regularization loss function to constraint the feature representation to be close to either 0 or 1. The regularization loss function is defined as follows:

$$L_{\mathrm{Reg}} = \sum_{i=1}^{m} \left\| \mathbf{f}_i - 0.5\mathbf{e} \right\|_2^2 \tag{2}$$

where $\mathbf{e}$ represents the *K*-dimensional vector of all elements 1.

Apart from making the final feature approach binary, the balancing loss function is designed to push the feature vector across different modalities since data imbalance can hurt the retrieval performance during the training process. Constrained by balancing loss, the $\mathbf{f}_{ik}$ values have the same number of 0 and 1 for each bit *k*. The balancing loss function is then shown as follows:

$$L_{\mathrm{Balancing}} = \sum_{i=1}^{m} \left(\mathrm{mean}(\mathbf{f}_i) - 0.5 \right)^2 \tag{3}$$

where $\mathrm{mean}(\bullet)$ denotes the average value of elements in a vector. The total loss function is described by the following:

$$L_{\mathrm{Total}} = L_{\mathrm{Triplet}} + \beta L_{\mathrm{Reg}} + \gamma L_{\mathrm{Balancing}} \tag{4}$$

where β and γ denote the two hyperparameters.

Finally, after training with the total loss, we design the hash layer to transfer the high-dimensional deep features into compact *K*-bit hash codes. To obtain the binary hash codes, we quantize the unified feature using a simple threshold as follows:

$$\mathbf{b}_i = \left(\mathrm{sgn}\left(\mathbf{f}_i - 0.5 \right) + 1 \right) / 2 \tag{5}$$

where $\mathbf{b}_i$ denotes the binary code vectors, $\mathrm{sgn}(\bullet)$ denotes element-wise operations, i.e., $\mathrm{sgn}(x) = 1$ if $x > 0$ and -1 if otherwise.

V. Experiments and Analysis

To validate our proposed method, we consider an extensive series of experiments. Section V-A introduces the experimental setup and evaluation criteria, Section V-B describes validation of the proposed DCMHN, Section V-C analyzes the impacts of parameters α, β, and γ on the results for the cross-modality retrieval, whereas Section V-D presents a comparison between our results and some baselines.

A. Experimental Setup and Evaluation Criteria

During the training process, we adopt the proposed SODMRSID dataset for evaluating performance of our DCMHN for the CM-CBRSIR tasks. Specifically, SODMRSID is randomly split into two subsets and used to construct the training and testing sets, denoted as: $\mathbf{D}_{\mathrm{train}} = \{(\mathbf{S}_i, \mathbf{O}_i, L_i) | i = 1, 2, \ldots, V\}$ and $\mathbf{D}_{\mathrm{test}} = \{(\mathbf{S}_i, \mathbf{O}_i, L_i) \mid i = 1, 2, \ldots, Q\}$, respectively. Where V and Q are set to 11 000 and 1000, respectively, implying that 11 000 images pairs of the SODMRSID is used for training while the remainder is used for testing.

In the experiment, two feature extractors, Resnet18 and Resnet50, are introduced as shallow and deep networks, respectively. The DCMHN architecture is provided in Table II. Moreover, Adam optimizer with a learning rate of 0.001 is introduced to optimize loss of function. Furthermore, two evaluation metrics namely, the precision at k samples (P@k) and the mean average precision (mAP), are adopted for comparison.

All experiments are implemented under PyTorch deep learning framework on a 64-b station with Ubuntu16.04, 32GB of RAM, 8 Intel(R) Core(TM) i7-6770K CPU, and NVIDIA RTX 2080Ti.

B. Effective of the DCMHN

This section quantitatively evaluates the overall performance of our proposed method, under several loss function, to validate the DCMHN. Particularly, four loss function parameters are considered: β and γ are set to 0 to represent the adoption of triplet loss function; β and γ are set to 1 and 0, respectively, to represent the combination of triplet and regularization loss function; β is set to 0 and γ to 1 to represent the combination of triplet loss

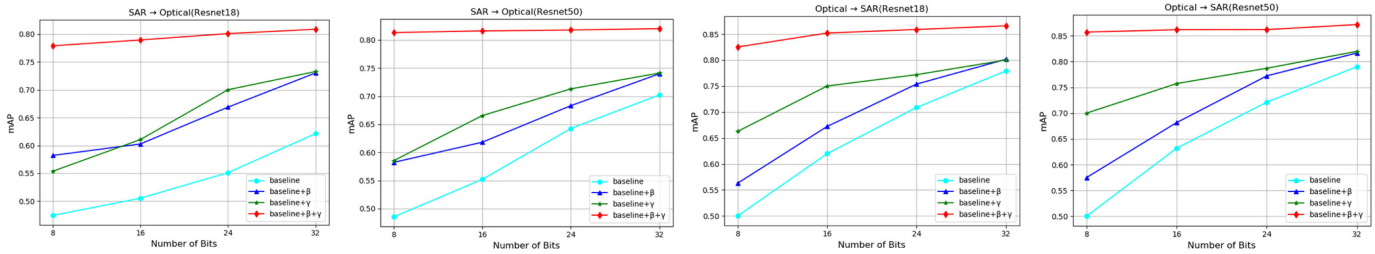


Fig. 6. mAP curves with respect to different code lengths under different loss functions: (a) SAR → Optical, Resnet18. (b) SAR → Optical, Resnet50. (c) Optical → SAR, Resnet18. (d) Optical → SAR, Resnet50.

TABLE II
ARCHITECTURE OF THE PROPOSED NETWORK

Layer name	Output size	18-layer	50-layer
Conv1	128*128	7*7, 64, stride2	
Conv2	64*64	3*3 maxpool, stride2	
		$\begin{bmatrix} 3*3,64 \\ 3*3,64 \end{bmatrix}$*2	$\begin{bmatrix} 1*1,64 \\ 3*3,64 \\ 1*1,256 \end{bmatrix}$*3
Conv3	32*32	$\begin{bmatrix} 3*3,128 \\ 3*3,128 \end{bmatrix}$*2	$\begin{bmatrix} 1*1,128 \\ 3*3,128 \\ 1*1,512 \end{bmatrix}$*4
Conv4	16*16	$\begin{bmatrix} 3*3,256 \\ 3*3,256 \end{bmatrix}$*2	$\begin{bmatrix} 1*1,256 \\ 3*3,256 \\ 1*1,1024 \end{bmatrix}$*6
Conv5	8*8	$\begin{bmatrix} 3*3,512 \\ 3*3,512 \end{bmatrix}$*2	$\begin{bmatrix} 1*1,512 \\ 3*3,512 \\ 1*1,2048 \end{bmatrix}$*3
	1*1	Global average pool	
Ft	1*1	512	2048
Fl	1*1	K[1]	K
Fb	1*1	K	K

function and balancing loss function; and finally,[2] β and γ are set to 1 to represent the combination of triplet loss function, regularization loss function, and balancing loss function.

Results from two evaluation protocols (precision at top-200 retrieved images and mAP) are reported under various loss functions and hashing feature coding lengths in Tables III and IV. According to Table III, when the network structure is fixed, the combination of three loss functions results in a more favorable performance relative to other loss functions under the same hashing feature coding length for the SAR → Optical retrieval task. Similarly, the best performance is also achieved by combining the three-loss functions for the Optical → SAR retrieval task (see Table IV). Besides, a higher accuracy is recorded in the Optical → SAR retrieval task under similar conditions than with SAR → Optical retrieval task. This is mainly because the speckle noise contained in SAR images makes the feature representation not accurate enough. Moreover, the deeper network achieves better performance than the shallower network for both SAR → Optical and Optical → SAR retrieval tasks, although the improvement is not apparent.

The superiority of the proposed method is intuitively captured in Fig. 6, as evidenced by the impact of the hash code lengths on the result mAP value. Particularly, the "baseline" means the only adoption of the triplet loss (i.e.,β and γ are set to 0) (see Fig. 6). Based on these comparisons, it is clear that a combination of three loss functions has a consistent advantage over individual loss functions. Besides, combining the triplet loss with balancing loss or regularization loss results in a relatively better performance. The total loss also provides a stable and most favorable performance for the different code lengths. This may be attributed to the fact that the total loss, resulting from a combination of the three losses, may produce more effective binary codes to represent the discriminative features. In addition, all advocated loss functions may gradually improve performance of the proposed methods along with the increase of the hash code lengths. Results from the analysis of varying tendency of precision, as the number of top retrieved images changes based on 32-b hash codes, are illustrated in Fig. 7. Generally, the precision curves for total loss are considerably above other losses, which is sufficient to prove the superiority of retrieval ability of the total loss.

To enable feature visualization, we employ the *t*-distributed stochastic neighbor embedding (*t*-SNE) algorithm [30] to obtain the 2-D representation of the feature vectors under various loss function, under a fixed network structure (see Fig. 8). Specifically, principal component analysis is introduced to compress the high dimensional features to two-dimensions. This intuitively reveals feature distributions for SAR and optical images under triplet loss, triplet loss with regularization loss, triplet loss with balancing loss, a combination of triplet loss, regularization loss, and balancing loss. Furthermore, the illustrations indicate that the feature distribution in Fig. 8(d) are more compact than those in Fig. 8(a)–(c). In conclusion, the binary codes produced by the proposed total losses are discriminative enough to enable

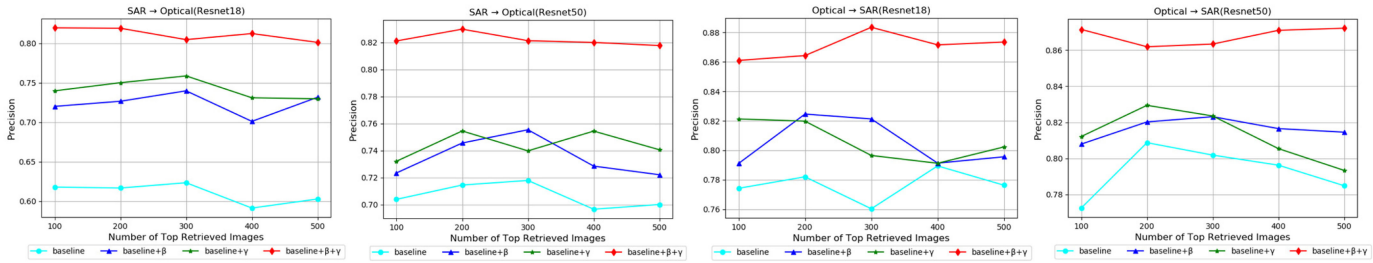
[2]K: hash code length.

TABLE III
COMPARED RESULTS OF DIFFERENT LOSS FUNCTION ON SAR → OPTICAL RETRIEVAL TASK

Feature Extractor	β	γ	Precision@200				MAP			
			K=8	K=16	K=24	K=32	K=8	K=16	K=24	K=32
Resnet18	0	0	0.4756	0.5088	0.5675	0.6167	0.4745	0.5045	0.5512	0.6212
	1	0	0.5802	0.6055	0.6712	0.7267	0.5821	0.6027	0.6687	0.7301
	0	1	0.5598	0.6212	0.7189	0.7502	0.5538	0.6106	0.7000	0.7326
	1	1	0.7923	0.8020	0.8066	0.8192	0.7788	0.7892	0.8007	0.8084
Resnet50	0	0	0.4865	0.5506	0.6478	0.7145	0.4854	0.5521	0.6426	0.7023
	1	0	0.6006	0.6198	0.7023	0.7456	0.5823	0.6182	0.6832	0.7399
	0	1	0.6043	0.6786	0.7196	0.7545	0.5854	0.6656	0.7131	0.7412
	1	1	0.8194	0.8230	0.8252	0.8298	0.8132	0.8161	0.8177	0.8201

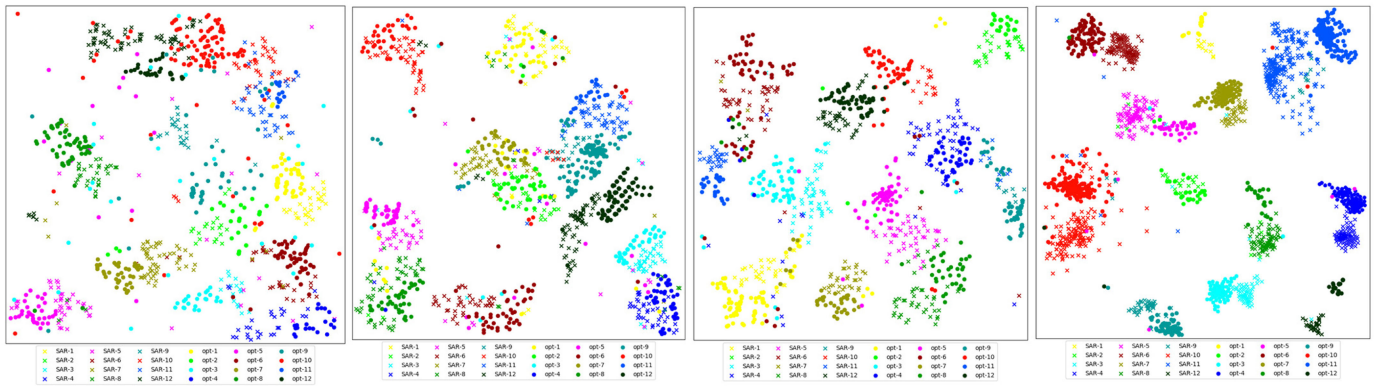
TABLE IV
COMPARED RESULTS OF DIFFERENT LOSS FUNCTIONS ON OPTICAL → SAR RETRIEVAL TASK

Feature Extractor	β	γ	Precision@200				MAP			
			K=8	K=16	K=24	K=32	K=8	K=16	K=24	K=32
Resnet18	0	0	0.5120	0.6203	0.7032	0.7820	0.5001	0.6199	0.7088	0.7792
	1	0	0.6098	0.7023	0.7613	0.8246	0.5632	0.6723	0.7538	0.8016
	0	1	0.6692	0.7582	0.7902	0.8198	0.6628	0.7502	0.7719	0.8002
	1	1	0.8418	0.8503	0.8684	0.8643	0.8253	0.8521	0.8589	0.8659
Resnet50	0	0	0.5038	0.6692	0.7412	0.8087	0.5003	0.6321	0.7213	0.7901
	1	0	0.6001	0.6887	0.7612	0.8202	0.5752	0.6818	0.7723	0.8167
	0	1	0.7009	0.7612	0.7992	0.8294	0.7001	0.7574	0.7871	0.8198
	1	1	0.8621	0.8598	0.8705	0.8619	0.8572	0.8618	0.8620	0.8717



(a) (b) (c) (d)

Fig. 7. Precision curves with respect to different number of top retrieved images under different loss functions: (a) SAR → Optical, Resnet18. (b) SAR → Optical, Resnet50. (c) Optical → SAR, Resnet18. (d) Optical → SAR, Resnet50.



(a) (b) (c) (d)

Fig. 8. Feature visualization of the learned features under different loss functions: (a) baseline. (b) baseline + β. (c) baseline + γ. (d) baseline + β + γ.

TABLE V
QUANTIFYING THE EFFECTIVE OF PROPOSED NETWORK UNDER DIFFERENT ALPHA

Cross-modality retrieval tasks	Feature Extractor	α=0.1	α=0.3	α=0.5	α=0.7	α=0.9
SAR → Optical	Resnet18	0.8025	0.8084	0.8031	0.8028	0.8017
	Resnet50	0.8198	0.8201	0.8169	0.8154	0.8133
Optical → SAR	Resnet18	0.8641	0.8659	0.8652	0.8643	0.8636
	Resnet50	0.8706	0.8717	0.8714	0.8708	0.8702

TABLE VI
QUANTIFYING THE EFFECTIVE OF PROPOSED NETWORK UNDER DIFFERENT BETA

Cross-modality retrieval tasks	Feature Extractor	β=0	β=0.5	β=1	β=2	β=4
SAR → Optical	Resnet18	0.7326	0.7843	0.8084	0.7967	0.7845
	Resnet50	0.7412	0.7895	0.8201	0.8169	0.8105
Optical → SAR	Resnet18	0.8002	0.8567	0.8659	0.8443	0.8326
	Resnet50	0.8198	0.8603	0.8717	0.8606	0.8488

TABLE VII
QUANTIFYING THE EFFECTIVE OF PROPOSED NETWORK UNDER DIFFERENT GAMMA

Cross-modality retrieval tasks	Feature Extractor	γ=0	γ=0.5	γ=1	γ=2	γ=4
SAR → Optical	Resnet18	0.7267	0.7825	0.8084	0.8074	0.7932
	Resnet50	0.7456	0.7903	0.8201	0.8180	0.8043
Optical → SAR	Resnet18	0.8016	0.8243	0.8659	0.8534	0.8512
	Resnet50	0.8167	0.8298	0.8717	0.8602	0.8565

clustering of samples in the same category and separate those under different categories on the cross-modality retrieval task.

C. Parameter Analysis

In this section, the effect of the hyperparameters α, β, and γ on mAP values are mainly explored with the code length set to 32. γ and β are set to 1. Table V shows the influence of the margins α with values ranging between 0.1 and 0.9 on the two cross-modality retrieval tasks. It can be seen that the best mAP value is obtained when the α is set to 0.3 for both Optical → SAR and SAR → Optical retrieval tasks, and the mAP value decreases gradually as α increases. Therefore, a reasonable margin value increases the ability of DCMHN to discriminate features. Table VI shows the influence of β with values ranging between 0 and 4 for the two cross-modality retrieval tasks with α set to 0.3 and γ set to 1. Table V demonstrates that DCMHN obtains competitive results when $\beta = 1$. By contrast, the results obtained at $\beta = 0$ are poor, indicating that an appropriate proportion level of regularization loss is required to produce efficient binary hash codes. Table VII shows the influence of γ with values ranging between 0 and 4 on the two cross-modality retrieval tasks with α set to 0.3 and β set to 1. These data indicate that DCMHN achieves slightly better results when $\gamma = 1$ than when the value is 2 and 4.

To intuitively show the influence of parameters on retrieval accuracy, a bar graph showing changes in mAP value with parameters is shown in Fig. 9. Notably, there is no significant difference between Fig. 9(c)–(f) and Fig. 9(a) and (b). We conclude that the performance of DCMHN is more sensitive to β and γ than α.

D. Comparison With Several Baselines

The performance of the proposed DCMHN is determined by comparing the mAP values of our method with various baselines under different code lengths on the proposed SODMR-SID dataset as shown in Table VIII. DCMHN_18 denotes the proposed method with Resnet18 as the architecture network, and DCMHN_50 is the method with Resnet50 as the architecture network. All methods shown in Table VIII adopt deep features except DCH [49] and SCM [52]. The performance of the method based on handcrafted features is inferior in the two cross-modality retrieval tasks compared with deep features.

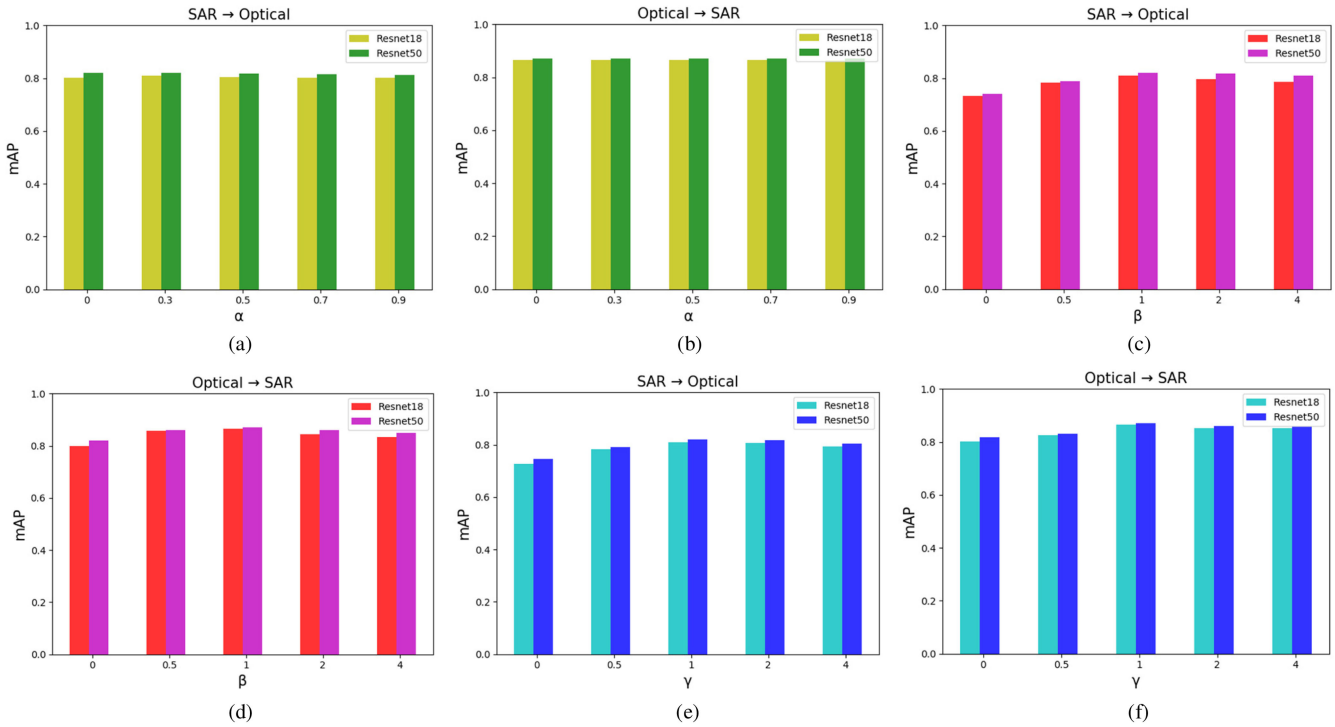


Fig. 9. The influence of different parameters on retrieval accuracy: (a) The influence of α on retrieval accuracy, SAR $\rightarrow$ Optical. (b) The influence of α on retrieval accuracy, Optical $\rightarrow$ SAR. (c) The influence of β on retrieval accuracy, SAR $\rightarrow$ Optical. (d) The influence of β on retrieval accuracy, Optical $\rightarrow$ SAR. (e) The influence of γ on retrieval accuracy, SAR $\rightarrow$ Optical. (f) The influence of γ on retrieval accuracy, Optical $\rightarrow$ SAR.

TABLE VIII
COMPARISON OF MAP VALUES WITH DIFFERENT METHODS UNDER DIFFERENT CODE LENGTHS

Cross-modality retrieval tasks	Feature Extractor	K=8	K=16	K=24	K=32
SAR → Optical	DCH [49]	0.1325	0.1742	0.1754	0.1788
	SCM [52]	0.1862	0.1893	0.1942	0.1976
	DCMH [54]	0.3623	0.3629	0.3657	0.3677
	DVSH [53]	0.3729	0.3765	0.3772	0.3783
	SIDHCNNs [14]	0.4123	0.4134	0.4216	0.4233
	DCMHN_18	0.7788	0.7892	0.8007	0.8084
	DCMHN_50	0.8132	0.8161	0.8177	0.8201
Optical → SAR	DCH [49]	0.2213	0.2314	0.2363	0.2376
	SCM [52]	0.1921	0.1954	0.2013	0.2019
	DCMH [54]	0.4123	0.4142	0.4145	0.4178
	DVSH [53]	0.4353	0.4432	0.4486	0.4491
	SIDHCNNs [14]	0.4812	0.4844	0.4873	0.4876
	DCMHN_18	0.8253	0.8521	0.8589	0.8659
	DCMHN_50	0.8572	0.8618	0.8620	0.8717

This is because the hash methods based on handcrafted features cannot effectively preserve the semantic information and learn the discriminative hash codes. For this reason, the performance of DCMH [54] and DVSH [53] improve significantly, attributed to the fact that deep network produces more effective binary hash codes than the handcrafted ones. However, the methods based on deep features facilitate the retrieval of natural images and texts. The complexity of CM-CBRSIR tasks between SAR and optical sensors limit the performance of the two deep hash methods. Of note, SIDHCNNs [14] protects the modality-specific information by adopting two different deep architectures. However, the purpose of SIDHCNNs [14] is to solve CM-CBRSIR between panchromatic and multispectral sensors, but this success cannot be transferred to the CM-CBRSIR between SAR and optical sensors tasks. The complexity of the feature representation of the content of SAR images makes it hard for SIDHCNNs to perform well. The proposed DCMHN_18 consistently outperforms all the baselines on both SAR $\rightarrow$ Optical and Optical $\rightarrow$ SAR

TABLE IX
TIME COMPARISON WITH DIFFERENT METHODS

Cross-modality retrieval tasks	Feature Extractor	Time (train)	Time (test)
SAR →Optical	DCMH [54]	3.1h	11.2s
	DVSH [53]	6.5h	39.2s
	SIDHCNNs [14]	1.9h	3.9s
	DCMHN_18	2.7h	6.1s
Optical → SAR	DCMH [54]	3.1h	71.1s
	DVSH [53]	6.5h	214.2s
	SIDHCNNs [14]	1.9h	19.2s
	DCMHN_18	2.7h	31.4s

retrieval tasks with different code lengths. The DCMHN_50 described in this study yields the best retrieval results among all methods.

The computational complexity of the proposed method is determined by comparing the training and testing time of our method with various baselines as shown in Table IX. Considering that DCH [49] and SCM [52] are based on the hand-crafted features, it is unfair to compare computational complexity both for the training and testing phase. SIDHCNNs [14] is the faster one with its light network structure, but it cannot obtain a satisfactory retrieval accuracy. We can find that the proposed DCMHN_18 can achieve a competitive performance within a short time.

VI. CONCLUSION

In this article, we have proposed a novel deep cross-modality hashing network for CM-CBRSIR between SAR and optical sensors. To the best of our knowledge, this is the first work to solve the problem of CM-CBRSIR allowing the sensor between SAR and optical. In the proposed method, an image transforming strategy is introduced to convert optical images with three channels to four different types of single channel images. In this way, the diversity of the modalities is considerably increased, making the network pay more attention to the texture information of the SAR and optical images. Afterward, the paired training strategy is employed to extract the discriminative features across different modalities. Finally, triplet loss combined with hash codes is conducted to reduce the dimension of feature and produce the efficient binary codes which further increases the retrieval accuracy and efficiency.

The dataset named SODMRSID is first proposed to evaluate the effectiveness of the proposed method. The results demonstrate that the proposed method is superior to several baselines.

REFERENCES

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

[2] 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.

[3] M. Lin, R. Ji, S. Chen, X. Sun, and C. Lin, "Similarity-preserving linkage hashing for online image retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 5289–5300, 2020.

[4] J. Yang, J. Liang, H. Shen, K. Wang, P. L. Rosin, and M. Yang, "Dynamic match kernel with deep convolutional features for image retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5288–5302, Nov. 2018.

[5] G. S. Xia *et al.*, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," 2017, *arXiv:1707.07321*.

[6] Y. Li, Y. Zhang, H. Xin, Hu 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.

[7] W. Zhou *et al.*, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 489.

[8] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2019.

[9] M. Yang *et al.*, "Multitask learning for cross-domain image captioning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1047–1061, Apr. 2019.

[10] X. Min, G. Zhai, J. Zhou, X. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Trans. Image Process.*, vol. 29, pp. 3805–3819, 2020.

[11] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Pérez, and G. Richard, "Weakly supervised representation learning for audio-visual scene analysis," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 416–428, 2020.

[12] X. Xiang, N. Lv, Z. Yu, M. Zhai, and A. El Saddik, "Cross-modality person re-identification based on dual-path multi-branch network," *IEEE Sens. J.*, vol. 19, no. 23, pp. 11706–11713, Dec. 2019.

[13] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5390–5399.

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

[15] W. Xiong, Z. Xiong, Y. Cui, and Y. Lv, "A discriminative distillation network for cross-source remote sensing image retrieval," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1234–1247, 2020.

[16] W. Xiong, Y. Lv, X. Zhang, and Y. Cui, "Learning to translate for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4860–4874, Jul. 2020,

[17] G. Healey, and A. Jain, "Retrieving multispectral satellite images using physics-based invariant representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 842–848, Aug. 1996.

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

[19] 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.

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

[21] H. F. Yang, K. Lin, and C. S. Chen, "Supervised learning of semantic-spreserving hash via deep convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 437–451, Feb. 2018.

[22] X. Jin, F. Yi, Z. Fan, and E. Wong, "DeepShape: Deep learned shape descriptor for 3D shape matching and retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1275–1283.

[23] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput.Vision*, Sep. 1999, pp. 1150–1157.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

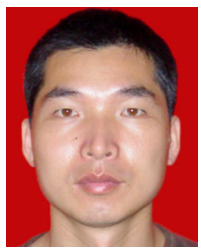
[25] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vision*, May 2006, pp. 404–417.

[26] 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. 8, pp. 273–292, 2015.

[27] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. Comput. Vision Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[28] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2007, pp. 1–8.

[29] 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.
[30] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 8, pp. 2579–2605, 2008.
[31] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4062–4076, Jun. 2019.
[32] Z. Shao *et al.*, "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 12, 2020, Art. no. 1050.
[33] Z. Shao, L. Wang, Z. Wang, and J. Deng, "Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2663–2674, Aug. 2019.
[34] Y. Ge, S. Jiang, Q. Xu, C. Jiang, and F. Ye, "Exploiting representations from pre-trained convolutional neural networks for high-resolution remote sensing image retrieval," *Multimedia Tools Appl.*, pp. 1–27, 2017.
[35] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sens.* vol. 9, no. 12, 2017, Art. no. 1330.
[36] Z. Shao, L. Wang, Z. Wang, W. Du, and W. Wu, "Saliency-aware convolution neural network for ship detection in surveillance video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 781–794, Mar. 2020.
[37] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 117–126, Jan. 2018.
[38] Q. Liu, R. Hang, H. Song, and F. Zhu, "Adaptive deep pyramid matching for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, Nov. 2016.
[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. 4, 2019, Art. no. 281.
[40] 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. 1–16, Feb. 2018.
[41] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, and W. Min, "Remote sensing image retrieval using convolutional neural network features and weighted distance," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1535–1539, Oct. 2018.
[42] T. Reato, B. Demir, and L. Bruzzone, "An unsupervised multicode hashing method for accurate and scalable remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 276–280, Feb. 2019.
[43] D. Espinoza-Molina, J. Chadalawada, and M. Datcu, "SAR image content retrieval by speckle robust compression based methods," in *Proc. EUSAR 10th Eur. Conf. Synthetic Aperture Radar*, 2014, pp. 1–4.
[44] L. Jiao, X. Tang, B. Hou, and S. Wang, "SAR images retrieval based on semantic classification and region-based similarity measure for earth observation," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 3876–3891, Aug. 2015.
[45] X. Tang and L. Jiao, "Fusion similarity-based reranking for SAR image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 242–246, Feb. 2017.
[46] X. Tang, L. Jiao, and W. J. Emery, "SAR image content retrieval based on fuzzy similarity and relevance feedback," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1824–1842, May 2017.
[47] K. Zhang, B. Li, and R. Tao, "SAR image retrieval based-on fly algorithm," in *Proc. 10th Int. Conf. Adv. Comput. Intell.*, 2018, pp. 502–507.
[48] F. Ye, W. Luo, M. Dong, H. He, and W. Min, "SAR image retrieval based on unsupervised domain adaptation and clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1482–1486, Sep. 2019.
[49] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, 2017.
[50] X. Luo, X.-Y. Yin, L. Nie, X. Song, Y. Wang, and X.-S. Xu, "SDMCH: Supervised discrete manifold-embedded cross-modal hashing," *IJCAI*, pp. 2518–2524, 2018.
[51] Q. Y. Jiang and W. J. Li, "Discrete latent factor model for cross-modal hashing," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3490–3501, Jul. 2019.
[52] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014.
[53] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1445–1454.
[54] Q. Y. Jiang and W. J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3232–3240.
[55] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.
[56] Z. Zheng, J. Chen, X. Zheng, and X. Lu, "Remote sensing image generation from audio," *IEEE Geosci. Remote Sens. Lett.* to be published, doi: 10.1109/LGRS.2020.2992324.
[57] M. Gou, Y. Yuan, and X. Lu, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens.*, 2018, pp. 1–7.
[58] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.
[59] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1985–2000, Mar. 2020.
[60] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 256–270, 2020.
[61] 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.
[62] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogram. Remote Sens.*, vol. 14 no. 5, pp. 197–209, 2018.
[63] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 964.
[64] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," in *Proc. ISPRS Ann. Photogram., Remote Sens. Spatial Inform. Sci.* , 2018, 2018.
[65] R. Torres *et al.*, "GMES Sentinel-1 mission," *Remote Sens. Environ.*, vol. 120, pp. 9–24, 2012.
[66] M. Drusch *et al.*, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, 2012.
[67] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020.
[68] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th Int. Conf. Adv. Geogr. Inf. Syst.*, 2010, pp. 270–279.
[69] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *Geosc. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, 2015.



Wei Xiong received the B.S., M.S., and Ph.D. degrees from Naval Aviation University, Yantai, China, in 1998, 2001, and 2005, respectively.

From 2007 to 2009, he was a Postdoctoral Researcher with the Department of Electronic Information Engineering, Tsinghua University, Tsinghua, Beijing.

He is currently a Full Professor with the Naval Aviation University, Yantai, China, where he teaches random signal processing and information fusion. He is one of the Founders and the Directors with the Research Institute of information Fusion, Naval Aviation University. His research interests include pattern recognition, remote sensing, and multisensor information fusion.

Dr. Xiong is the Member and the Director General of Information Fusion Branch of Chinese Society of Aeronautics and Astronautics.

Zhenyu Xiong received the B.S. degrees from Naval Aviation University, Yantai, China, in 2018. He is currently working toward the M.S. degree in information and communication engineering with Naval Aviation University.

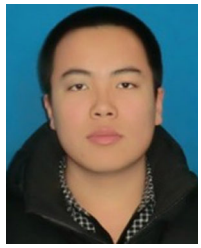
His research interests include information fusion, and deep learning with their applications in remote sensing.

Yang Zhang received the M.S. and Ph.D. degrees from Naval Aviation University, Yantai, China, in 2007 and 2011, respectively.

His research interests include information fusion and tactics of defense and attack.

Yaqi Cui received the B.S., M.S., and Ph.D. degrees in information and communication engineering from Naval Aviation University, Yantai, China, in 2008, 2011, and 2014, respectively.

From 2014, he was a Lecturer with Naval Aviation University. His research interests include information fusion, machine learning, and deep learning with their applications in information fusion.



Xiangqi Gu received the B.S. degree from Changchun Institute of Technology, Changchun, China, in 2017, the M.S. degree from Naval Aviation University, Yantai, China, in 2019. He is currently working toward the Ph.D. degree in information and communication engineering with Naval Aviation University.

His research interests include information fusion, and reinforcement learning with their applications in target tracking.