# Deep Adversarial Cascaded Hashing for Cross-Modal Vessel Image Retrieval

Jiaen Guo and Xin Guan

***Abstract*—In recent years, cross-modal remote sensing image retrieval has attracted a lot of attention in remote sensing (RS) information processing. It is worth mentioning that land cover scenes, whether unimodal or cross-modal, are the primary research contents of remote sensing image retrieval, and there are few studies on vessel images captured by RS satellites, let alone cross-modal retrieval tasks. Vessel images have smaller scale, lower resolution, and less detailed information than land cover images, so it is difficult to retrieve the exact images we want. In this article, a hashing method called deep adversarial cascaded hashing (DACH) is proposed to address these problems. To extract the subtle and discriminative features contained in RS vessel images accurately, we build a deep cascaded network that fuses multilevel features boosted both in depth and width, and the self-attention mechanism can further enhance the fused features. Combined with hash learning, we also design a weighted quintuplet loss to supervise the transition of discrimination and similarity between different metric spaces, and reduce cross-modal discrepancy at the same time. In addition, we apply the deep adversarial constraint to both feature learning and hash learning, trying to bridge the modality gap and achieve a cross-modal retrieval as precise as unimodal retrieval. Comprehensive experiments on two public bimodal vessel image datasets compared with several excellent cross-modal retrieval methods are conducted to demonstrate the effectiveness of our DACH, and the results show that the proposed method is effective and competitive on cross-modal vessel image retrieval tasks, outperforming state-of-the-art methods.**

***Index Terms*—Cross-modal vessel image retrieval, deep adversarial learning, deep hash learning, multilevel feature fusion.**

## I. Introduction

RETRIEVING interesting target information from massive ocean monitoring data is a key point of marine situation awareness, and the rapid development of remote sensing (RS) technology provides powerful support for accurate observations and perceptions of the ocean. Under these circumstances, mining and analyzing vessel images can effectively facilitate the management of RS big data [1].

Nowadays, a large quantity of military and civilian RS satellites bring all kinds of high-quality RS images and promote the research on information retrieval indirectly. Emerging technologies, such as text-based retrieval [2], [3], [4], content-based retrieval [5], [6], unimodal retrieval [7], [8], [9] and cross-modal retrieval [10], [11], [12], have brought the RS big data analysis into a new era. Especially, the development of deep neural network (DNN) makes it possible to capture the deep semantic information hidden in RS images, so the retrieval accuracy has improved significantly. Take the widely-used unimodal RS dataset UC-Merced for example, Song et al. [13] used a deep hashing convolutional neural network (DHCNN) for retrieval and classification and achieve a mean average precision (MAP) of 98.08% on UC-Merced. Shan et al. [14] carried out retrieval experiments on UC-Merced and the MAP@20 even reached to 99.7%. In other words, we can get almost everything we want from huge databases accurately.

Recently, cross-modal retrieval tasks have drawn more attention and researchers start to explore the possibility of retrieving relevant scene images from different sources. [15] was the first attempt for this topic, which proposed source-invariant DHCNNs to cope with the modality differences, and a bimodal (panchromatic and multispectral) RS image dataset DSRSID is also constructed for evaluation. Xiong et al. [16] gave a basic mapping framework based on cycle-identity-generative adversarial network, which can generate images into the target domain, and the MAP on DSRSID is up to 97.55%. Furthermore, in [17], They tried to distill the information of the source domain images and then migrate the acquired knowledge to the target domain, so that cross-modal features can get closer with parameter transferring. The experimental results show that it can achieve a MAP of 98.98%, even higher than [16]. Another research [18] by Xiong's team gave research on cross-modal retrieval between SAR and optical RS images, and the highest MAP is up to 87.17%. Based on the above research we can see that the retrieval between multimodal RS images has developed quickly and achieved great retrieval results.

Although emerging methods have been proposed for RS image retrieval, the target level retrieval hasn't attracted much attention, and the retrieval for land cover scenes remains the main research topic. It is worth noting that vessel target is far different from terrain features, the main differences can be summarized as follows.

1) Vessel images have lower resolution, smaller size than land covers, and have more complex backgrounds such as ports, rocks, waves, clouds, etc.
2) The texture features of vessel images are not significant and the color information is poor, so there are not enough

Manuscript received 26 April 2022; revised 19 October 2022; accepted 13 January 2023. Date of publication 30 January 2023; date of current version 27 February 2023. This work was supported in part by the National Defense Science and Technology Excellence Youth Talent Foundation of China under Grant 02017-JCJQ-ZQ-003 and in part by the Taishan Scholar Engineering Special Foundation of China under Grant ts 201712072. *(Corresponding author: Xin Guan.)*

Jiaen Guo is with the Unit 91422 of PLA, Yantai 265200, China, and also with the Naval Aviation University, Yantai 264001, China (e-mail: guojiaen@163.com).

Xin Guan is with the Naval Aviation University, Yantai 264001, China (e-mail: gxtongwin@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3240414

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

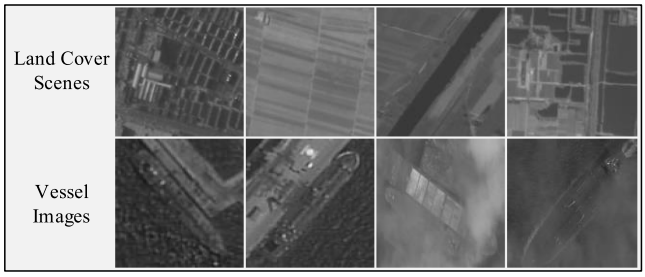



Fig. 1. Different images captured by panchromatic sources.

discriminative details for extraction, only the limited vessel contour and surface features can be used.

3) Vessels are moving targets, causing vessel images to deform easily, and different imaging mechanisms of RS satellites can also lead to misjudgment of vessels of different sizes.

Fig. 1 shows some example images of land covers and vessels captured by Panchromatic RS satellites. By contrast, we can see that information contained in vessel images is very poor, and there lie huge similarities between different vessel images. Whether it is to extract the features such as color and texture in advance or to extract the features in an end-to-end manner using DNN directly, it is very difficult to capture sufficient discriminative information for retrieval. Therefore, it is an arduous task to achieve an accurate vessel image retrieval. Current research mainly improves some RS image retrieval methods to adapt to the particularity of vessel image retrieval. In [19], Hu et al. used bag of visual words (BoVWs) to describe vessel images and the improved construction of visual dictionary effectively improved the efficiency of retrieval. Using BoVW as well, Tian et al. [20] utilized a convolutional neural network (CNN) to contract dictionary databases which consist of numerous convolution features, experiments indicated that the proposed method is much better than traditional methods. Aside from the research mentioned earlier, there is not much relevant research on vessel images retrieval. The poor condition makes it necessary to develop an appropriate retrieval system for marine vessel targets.

To meet the needs for cross-modal retrieval of RS vessel images, this article uses dual-stream CNN as the feature extractor and adopts hash learning to realize rapid retrieval. Firstly, based on Resnet50 [21], a feature cascaded subnetwork is designed to enhance the discrimination of extracted features both in width and depth. This branch can effectively solve the difficulty of feature extraction from RS vessel images by combining low-level color and texture features with high-level semantic features of vessel targets. Second, to realize the rapid cross-modal retrieval of vessel targets, a hash learning module is constructed to convert the fused features into high-quality hash codes, and different constraint ensures the complete transition of similarities and discrimination from feature space to hash space. Finally, to eliminate the retrieval difficulties caused by cross-modal differences, adversarial learning is introduced to model training, so that the features and hash codes of different modalities can compete in their respective semantic space, resulting in a free transformation between the source domain and target domain, which can effectively improve the accuracy of retrieval.

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

1) As far as we know, this is the first attempt at cross-modal RS vessel image retrieval. The study provides a basic framework DACH for the retrieval task at the target level, and the design of the feature learning module as well as the introduction of many competitive learning strategies can meet the needs of fine target image retrieval.
2) We propose a general cascaded fusion network for micro-target feature extraction, which can extract discriminative features from coarse to fine layer by layer. In addition, the self-attention mechanism can help to realize the adaptive fusion of different levels of features.
3) We design a hash learning module to transfer the metric space, combined with a new weighted quintuplet loss which can reduce information loss in the transferring process and enhance the discrimination of the learned hash codes.
4) To eliminate the modality gap, we add an adversarial training mechanism to both the feature learning part and hash learning part, which can realize the consistent expression of cross-modal features throughout the retrieval process.

The rest of this article is organized as follows. Section II summarizes the current research on cross-modal retrieval, hash learning, adversarial learning, and other related technologies. Section III introduces the framework of DACH in detail. In Section IV we analyze the retrieval effect of DACH and compare it with some state-of-the-art methods. Finally, a conclusion is given in Section V.

## II. Related Work

### A. Cross-Modal Retrieval

Unimodal image retrieval is a long-term hot spot and has attracted a lot of attention. There are two key points to solve this problem: feature extraction and similarity measurement. Xiong et al. [17] divided RS visual features into three categories: low-level; mid-level; and high-level. Before CNN is widely adopted in the field of computer vision, feature extraction has long relied on various hand-crafted low-level features, such as color, texture, and shape, as well as mid-level visual aggregation features, such as BoVW and fisher vector. However, the extraction needs to reasonably select appropriate visual description features from images with different characteristics, which is a great test for researchers' domain expertise and engineering skills, and it is difficult to ensure that the extracted features can fully describe the characteristics of images. Therefore, feature extraction of RS images entered a bottleneck before the advent of CNN, and the retrieval accuracy has not improved for a long time. CNN provides a new paradigm for visual understanding. Neural networks composed of a series of convolution layers and activation layers show strong visual feature description ability, and the convolution kernels are equivalent to a series of filters, which can fully extract the features such as color, edge, and

texture of the images. With the deepening of networks, the nonlinear fitting ability of networks is enhanced, and the deep semantic information is extracted as well. Therefore, compared with the low-level and mid-level features, high-level semantic features extracted by CNN have stronger representativeness and interpretability. Nowadays, DNN methods represented by CNN have been the primary choice for feature extraction [22].

After feature extraction, we need to measure the similarity to retrieve similar samples in metric space. Generally speaking, Minkowsky distance and histogram intersection method are often used for color features, whereas Euclidean distance and Mahalanobis distance are mainly used for texture features. A similarity measurement method based on weighted distance was proposed by calculating the weight of image categories [23], and the effect is better than the above conventional methods.

When it comes to cross-modal retrieval tasks, there is another challenge besides the above two key problems: the modality gap between heterogeneous data [24]. Satellites in different imaging mechanisms provide RS images with different characteristics. For example, multispectral (Mul) images usually have four channels, whereas panchromatic (Pan) images only have one. Synthetic aperture radar (SAR) images can work under any weather conditions, while visible images are easily affected by clouds, light, and seasons, etc. Chaudhuri et al. [25] tried to learn a discriminative common feature space for all modalities, and the proposed CMIR-NET can handle the Pan–Mul and image-audio cross-modal retrieval tasks adaptively. Based on hashing methods' powerful computati on efficiency, a fusion-based deep hashing method MsEspH [26] was proposed for retrieval between very-high-resolution (VHR) optical images and SAR images. Unlike conventional methods that used shared feature space to model the modality interactivity, MsEspH used Mul images generated by a generative adversarial network (GAN) to remove the spatial-spectral discrepancies, and an explicit semantic preserving-based function was used to preserve the intraclass similarity and interclass discrimination. What is more, a VHR-SAR bimodal dataset was constructed for evaluation.

As text and audio can also describe images, some predecessors conducted some research on visual-text and visual-audio retrieval. Ning et al. [27] considered that the intramodality and nonpaired intermodality representations also play an important role in semantic consistency modeling, so they built a consistency representation space to model these relationships, which is more effective than several excellent methods. DTBH [28] combined hash learning and relative semantic similarity relationship learning in an end-to-end network and improved the triplet loss with a selection strategy and a regularized method for visual-audio retrieval. In the domain of multimedia, the explosion of multimodal data also promotes the development of retrieval technologies, and a large number of advanced methods [29], [30], [31], [32] have emerged, which are roughly similar to the RS methods and greatly improve the accuracy of cross-modal retrieval. However, the target level retrieval hasn't attracted much attention yet, several published studies [33], [34], [35], [36], [37] mainly focus on unimodal tasks, and the cross-modal retrieval of target images needs further concentration.

### B. Deep Hash Learning

The mapped features in Hamming space are all binary forms, so the feature transformation from Euclidean space to Hamm-ing space can reduce the storage costs and improve the retrieval speed tremendously, which received extreme attention during the last decade. Benefitting from the rapid development of DNN, deep hash learning shines brightly in retrieval tasks, and developed many branches such as supervised hashing [38], [39], semisupervised hashing [40], [41], unsupervised hashing [42], [43], asymmetric hashing [44], [45], discrete hashing [46], [47], and so on. To directly use hash codes instead of original features to achieve an accurate and efficient retrieval, the most fundamental thing is trying hard to ensure that the intraclass similarity and interclass discrimination hidden in original features can be transferred smoothly and completely into the generated hash codes. Many researchers have made many meaningful attempts to solve this. Cheng et al. [48] tried to adopt hash learning to multilabel RS image retrieval and proved the feasibility with the proposed semantic-preserving deep hashing model. Using labels as supervised information, Li et al. [49] built a similar hashing network to common hashing methods and added an evaluation on classification. Ji et al. [50] and Nie et al. [51] explored the effectiveness of multiscale features on retrieval, both of them extracted multiscale features based on DNN and embed the multiscale semantics into the hash codes, which are more expressive in feature extraction. The method in [52] mainly focused on the quantization loss caused by relax constraint of hash code and the proposed DADH directly learning the discrete binary codes without relaxation. What is more, Li et al. [53] and Meng et al. [54] learned hash codes in an asymmetric way, which only need to learn the hash function of query samples, and directly learn the hash code of the database, which further improves the training and retrieving speed. The above research shows that hashing methods' effectiveness in improving the retrieval speed is amazing and they are generally common in principles, which all rely on DNNs and appropriate loss constraints to ensure the consistency between feature space and hash space. Similarly, we can also migrate the above methods to RS vessel image retrieval tasks, but how to better model the similarity between different spaces still needs further exploration.

### C. Deep Adversarial Learning

Adversarial learning is derived from GAN, which is composed of a generator and a discriminator generally. The unsupervised architecture makes it possible to regenerate fake samples similar to real samples under an adversarial training strategy, realizing the confusion of samples finally. GAN's powerful ability of modeling data distributions makes it popular in many fields and developed greatly in recent years.

When replacing samples with modalities, we can apply adversarial learning to cross-modal retrieval tasks. Many papers draw lessons from the adversarial idea of GANs, and they train different modalities in an adversarial way that can compete with each other, and all of them can learn better representations of the opponent modality. Hu et al. [55] used a modality-specific discriminator to eliminate the cross-modal discrepancy and

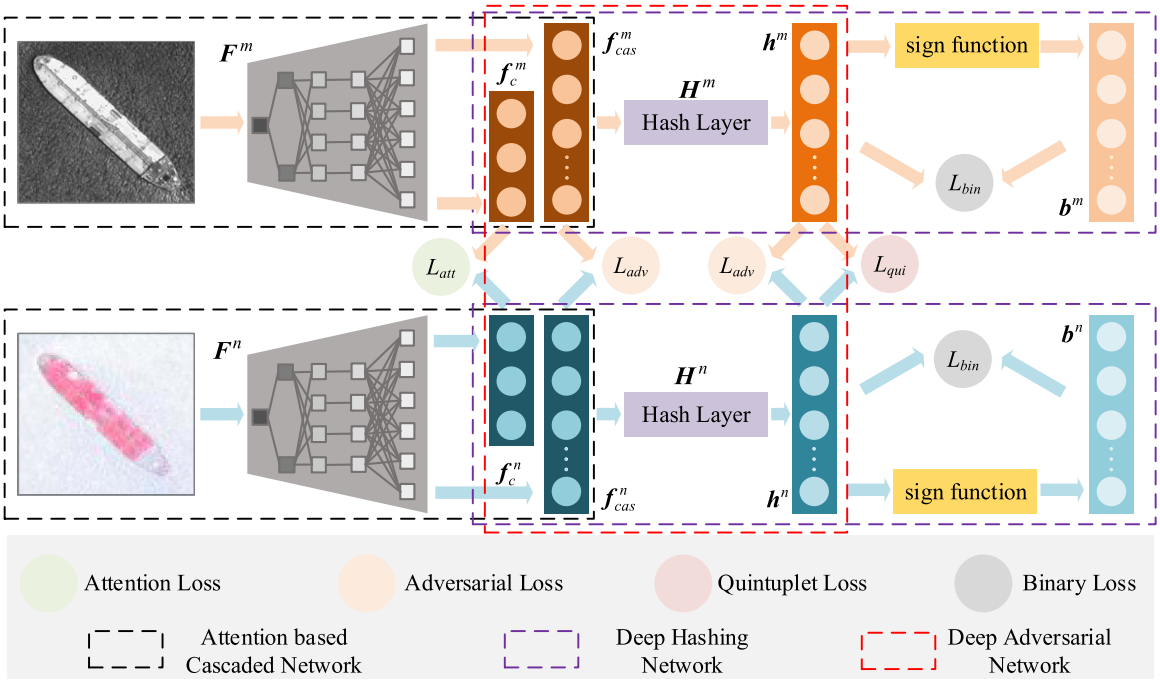



Fig. 2. Framework of our proposed DACH.

presented a multimodal discriminative analysis loss optimizing the above process. In another research, Hu et al. [56] designed two modality-specific generators, two modality-specific discriminators, and a cross-modal discriminative mechanism to achieve a confusion of modalities, and abundant experiments on several multimedia datasets proved the proposed method's effectiveness. Using GAN learning the shared feature semantic space of all modalities, Hong et al. [57] easily handled the fine-grained cross-modal retrieval tasks. Using GAN as well, the method in [58] paid more attention to feature semantic correlation in feature transformation, and enhanced the above process by sampling more semantically related and unrelated samples. Wu et al. [59] was one of the early attempts to bring adversarial learning to cross-modal hash methods. Specifically, it maximized the semantic relevance of different adversarial networks and used a self-supervised network generating label information to supervise the exploration of high-level semantic correlation. Huang et al. [60] designed a transfer network similar to distillation networks to transfer the knowledge from source to target domain, and the adversarial learning is used to improve the semantic consistency of cross-modal features. Besides, some other hashing methods [61], [62], [63], [64] adopted adversarial learning as well to eliminate the modality gap and enhance semantic correlation and consistency, but most of the above methods only impose adversarial learning on feature extraction or hash code generation separately, thus the similarity between feature space and hash space cannot be well maintained, causing information losing to some extent.

## III. Proposed Method

This section introduces the details of the proposed DACH. First, we make a preliminary of the cross-modal vessel image retrieval and give the formulation of the problem and some descriptions in math terms, then we introduce the feature learning part, hash learning part, and adversarial learning part successively, finally we discuss the optimization and some implementation details of the proposed DACH.

Fig. 2 shows the basic framework of our proposed method, which consists of three parts mentioned above. We call the feature learning part attention-based cascaded network (ABCN) and use it to extract discriminative features from RS vessel images. The deep hashing network (DHN) is designed for mapping the extracted features into compact hash codes, and the deep adversarial network (DAN) tries to enhance modal similarity in the above two subnetworks.

### A. Problem Formulation

Without losing generality, in this article, we mainly focus on the bimodal retrieval tasks. Assuming we have a collection of paired bimodal RS image samples, donated as $\boldsymbol{D} = \{\boldsymbol{d}_i\}_{i=1}^{R}$, $\boldsymbol{d}_i = \{m_i, n_i\}$, where $m_i$ donates the image of the first modality in $i$th paired instance, and $n_i$ donates the image of the second.

The bimodal dataset $\boldsymbol{D}$ assigns each bimodal image pair a semantic label additionally, donated as $\boldsymbol{L} = \{\boldsymbol{l}_i\}_{i=1}^{R}$, where $\boldsymbol{l}_i = [l_{i1}, l_{i2}, \ldots, l_{ic}]$ is the label vector of instance $i$, and $c$ is the total number of $\boldsymbol{D}$'s semantic categories. If instance $d_i$ belongs to the $j$th semantic category, $l_{ij} = 1$, otherwise $l_{ij} = 0$. The goal of

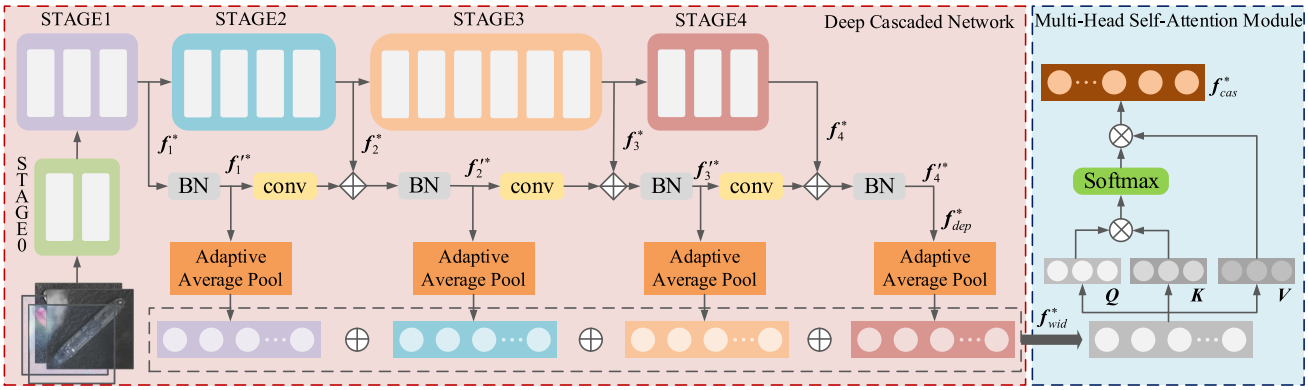



Fig. 3. Structure of DHCNN.

DACH is to learn the feature extraction functions $\boldsymbol{F}^*$ and hash code mapping functions $\boldsymbol{H}^*$

$$\begin{cases} \boldsymbol{f}_c^*, \boldsymbol{f}_{\text{cas}}^* = \boldsymbol{F}^*(*, \Theta^*) \\ \boldsymbol{h}^* = \boldsymbol{H}^*(\boldsymbol{f}_{\text{cas}}^*, \theta^*) \end{cases} \tag{1}$$

where $* \in m, n$, $\boldsymbol{f}_{\text{cas}}^* \in \mathbb{R}^d$, and $\boldsymbol{f}_c^* \in \mathbb{R}^c$, respectively, donates the learned cascaded features and predicted probability distribution of bimodal images, and $\boldsymbol{h}^* \in \mathbb{R}^K$ donates the learned continuous hash code, $\Theta^*, \theta^*$ are trainable parameters of corresponding functions. As the bimodal paired instances $\boldsymbol{d}_i = \{m_i, n_i\}$ are all images, we transform them to the same length $d$ and $K$ both in feature space and hash space, so that we can compare them directly in the metric space. The hash code in discrete form for retrieval is generated by an element-wise sign function $sign(\cdot)$

$$\boldsymbol{b}^* = \text{sign}(\boldsymbol{h}^*) \in \{-1, 1\}^K. \tag{2}$$

### *B. Attention-Based Cascaded Network for Discriminative Feature Extracting*

We have introduced the difficulties of the comprehensive feature extraction of RS vessel images in Section II-A, based on those analyses, the main challenge is to capture the complex information contained in images in a coarse-to-fine way completely. The structure of DNN exactly provides us with convenience to solve the above challenge, so we use DNN's special structure to construct the cascaded subnetwork to enhance the feature representation. What is more, the self-attention module is used to determine the weight of different levels of features and adjust them to the proper dimension. Now we introduce their detailed structures. Without special instructions, the following math terms all refer to the $i$th instance, and we omit the subscript $i$ for simplicity.

*1) Deep Cascaded Network:* To explore the multiscale information embedded in RS vessel images, we use Resnet50 as the backbone to extract features layer by layer. As is shown in Fig. 3, Resnet50 transforms images with one preprocessing stage and four transformation stages. The pre-processing stage converts the input image with a size of 3×224×224 into 64×56×56, and the following four stages can extract the discriminative features with the deepening of the network. It is known that the higher-level feature has more semantic information, whereas the lower-level contains more visual information like texture, shape, and color. The ABCN tries to integrate them to achieve a more discriminate and comprehensive feature representation.

Fig. 3 shows the specific structure of the ABCN and the left part is the deep cascaded network. We use the different processing stages in Resnet50 to build the cascaded network and integrate the features in two levels: depth and width. In specific, the four transformation stages in Resnet50 use the convolution kernels to halve the size and double the dimension of the feature maps from the previous stage, so we learn from the idea of Resnet50's shortcut connections and integrate the information extracted from the previous stage and the current stage with a batch normalization (BN) layer and a 3×3 convolution kernel [stride = (22)]. The feature extracted from the current stage can be re-expressed as follows:

$$\hat{\boldsymbol{f}}_{\text{cur}}^* = \text{conv}\left(\text{BN}\left(\boldsymbol{f}_{\text{pre}}^*\right)\right) + \boldsymbol{f}_{\text{cur}}^* \tag{3}$$

where conv indicates the convolution operation with a 3×3 convolution kernel and 2×2 stride, $\boldsymbol{f}_{\text{cur}}^*$ and $\boldsymbol{f}_{\text{pre}}^*$ indicates the feature map of the current stage and previous stage, respectively. $\hat{\boldsymbol{f}}_{\text{cur}}^*$ is the feature map of the current stage after summing. So, the final feature representation at the depth level $\boldsymbol{f}_{\text{dep}}^*$ can be formulated as follows:

$$\begin{aligned} \boldsymbol{f}_{\text{dep}}^* = \text{BN}\,(\text{conv}\,(\text{BN}\,(\text{conv}\,(\text{BN} \\ \times (\text{conv}\,(\text{BN}\,(\boldsymbol{f}_1^*)) + \boldsymbol{f}_2^*)) + \boldsymbol{f}_3^*)) + \boldsymbol{f}_4^*) \end{aligned} \tag{4}$$

where $\boldsymbol{f}_1^*, \boldsymbol{f}_2^*, \boldsymbol{f}_3^*, \boldsymbol{f}_4^*$ are the multiscale features extracted from the corresponding stage, and the final feature map size is 2048×7×7, equal to $\boldsymbol{f}_4^*$. The BNs and convs of different stages cascading with each other by orders constitute a cascaded subnetwork eventually. We can see that the final output feature map is fused with multiscale information, which enhances the feature representation greatly in depth.

Besides, we also enhance the feature representation in width. In specific, we use an adaptive average pool to compress the size of feature maps to 1×1 and restrain the useless features, and then we concatenate different features in channel. For clarity, the final feature representation in width level $\boldsymbol{f}_{\text{wid}}^*$ is formulated

as follows:

$$\boldsymbol{f}^*_{\text{wid}} = \text{AAP}\left(\boldsymbol{f'}^*_1\right) \oplus \text{AAP}\left(\boldsymbol{f'}^*_2\right) \oplus \text{AAP}\left(\boldsymbol{f'}^*_3\right) \oplus \text{AAP}\left(\boldsymbol{f'}^*_4\right) \tag{5}$$

where $\boldsymbol{f'}^*_1, \boldsymbol{f'}^*_2, \boldsymbol{f'}^*_3, \boldsymbol{f'}^*_4 (\boldsymbol{f'}^*_4 = \boldsymbol{f}^*_{\text{dep}})$ are the feature maps after BNs, "$\oplus$" indicates the concatenation operation, and the final number of feature channels is $3840(256 + 512 + 1024 + 2048)$.

The final feature representation at the width level contains not only low-level features but also features enhanced at depth, so we use it as the final enhanced feature and feed it into the self-attention module for further optimization.

*2) Multihead Self-Attention Module:* The core idea of the self-attention mechanism is to assign weights to different feature vectors and capture the internal correlation of features. The cascaded feature $\boldsymbol{f}^*_{\text{wid}}$ contains abundant multiscale information, which is captured in a balanced way. Whereas some of them are redundant and helpless, and features of different levels have different contributions to retrieval. Under these circumstances, the main idea of the application of the self-attention module is trying to focus on the important and discriminative information, and eliminate redundant information. It is generally accepted that the effect of the multihead self-attention mechanism is better than that of a single head because the former can capture more information, so we apply a multihead self-attention module to enhance the multiscale information even further.

As explained in [65], the attention module tries to map a query and a set of key-value pairs to an output, and the output is computed by weighting the values, where the weights are computed by the query and the corresponding key. The above process can be formulated as follows:

$$\text{Attention}\left(\boldsymbol{Q}^*, \boldsymbol{K}^*, \boldsymbol{V}^*\right) = \text{softmax}\left(\frac{\boldsymbol{Q}^* \boldsymbol{K}^{*\text{T}}}{\sqrt{d_k}}\right) \boldsymbol{V}^* \tag{6}$$

where $d_k$ is the dimension of queries, keys, and values, and $\boldsymbol{Q}^*, \boldsymbol{K}^*, \boldsymbol{V}^*$ are their packed matrix. Indeed, $\boldsymbol{Q}^*, \boldsymbol{K}^*, \boldsymbol{V}^*$ are linear transformation matrices of stacked cascaded feature vectors, where $\boldsymbol{Q}^*, \boldsymbol{K}^*$ are used to calculate the attention matrix to obtain the weight, which is then multiplied by $\boldsymbol{V}^*$ to obtain a more discriminative cascaded feature representation. The multihead self-attention module divides queries, keys, and values into several heads and calculates the attention vector for them, and then concatenates them together, which can be summarized in formulation as follows:

$$\begin{cases} \boldsymbol{f}^*_{\text{cas}} = \text{Concat}\left(\boldsymbol{head}^*_1, \ldots, \boldsymbol{head}^*_R\right) \boldsymbol{W}^{O*} \\ \boldsymbol{head}^*_r = \text{Attention}\left(\boldsymbol{Q}^* \boldsymbol{W}^{Q*}_r, \boldsymbol{K}^* \boldsymbol{W}^{K*}_r, \boldsymbol{V}^* \boldsymbol{W}^{V*}_r\right) \end{cases} \tag{7}$$

where $\boldsymbol{W}^{Q*}_r, \boldsymbol{W}^{K*}_r, \boldsymbol{W}^{V*}_r \in \mathbb{R}^{d_k \times (d_k/G)}$ and $\boldsymbol{W}^{O*}_r \in \mathbb{R}^{d_k \times d_k}$ are parameter matrices, $G$ is the number of heads, $\boldsymbol{f}^*_{\text{cas}} \in \mathbb{R}^d$ indicates the final learned cascaded features. As the attention mechanism can only deal with feature vectors, before seeding $\boldsymbol{f}^*_{\text{wid}}$ into the multihead self-attention module, we first need to flatten it into a vector, and then map it into the dimension we need through a linear layer when generating queries, keys, and values. The multihead self-attention module contains trainable parameters so that the final attention vector will get more and more discriminative under the constraints of various loss functions with training, then we get the enhanced feature representation for precise retrieval exactly.

To embed the semantics into $\boldsymbol{f}^*_{\text{cas}}$, We also add a category attention constraint to the network referring to [66], and utilize the cross-entropy loss function to increase the discrimination of different categories of samples and the similarity of similar samples. In specific, we send $\boldsymbol{f}^*_{\text{cas}}$ to the classifier which consists of two linear layers, the output of the classifier $\boldsymbol{f}^*_c$ is the probability distribution for categories. The math definition of attention loss is

$$L_{att} = -\frac{1}{R} \sum^{R} \left(\boldsymbol{l} \log\left(\boldsymbol{f}^m_c\right) + \boldsymbol{l} \log\left(\boldsymbol{f}^n_c\right)\right). \tag{8}$$

### C. Deep Hashing Network for Hash Code Generating

Compared with measuring the similarity between feature vectors in Euclidean space, converting continuous features into hash code significantly reduces the storage consumption and improve the correlation speed, so we try to convert the metric space to achieve a hashing retrieval. As is introduced in Section II-B, the key to realizing metric space transformation is to ensure that the similarity and discrimination of features in the original space can transit to the hash space completely. In Section II-B, we have designed an attention constraint to embed the semantics into $\boldsymbol{f}^*_{\text{cas}}$, and in this section, we try to transfer those similarities and embedded discrimination through the proposed weighted quintuplet loss, we also adopt binary constraints to make the real-valued hash codes follow the discrete uniform distribution as much as possible. Now, we introduce the DHN in detail.

As depicted in Fig. 2, the $\boldsymbol{f}^*_{\text{cas}}$ obtained in ABCN first transforms into real-valued hash codes $\boldsymbol{h}^*$ through DHN, which consists of a fully connected (FC) layer and the Tan*h* activation function. The linear layer is employed to resize the feature vector into the proper length we need, and Tan*h* can map the real-valued hash codes close to 0 and 1. It is no doubt that the hash module causes the information omission and the original feature loss, so we design a weighted quintuplet loss based on triplet loss to solve this, which comprehensively considers the samples of the same modality and cross-modality. Specifically, we build quintuplets of the first modality in the form $(\boldsymbol{b}^m, \boldsymbol{b}^m_+, \boldsymbol{b}^m_-, \boldsymbol{b}^n_+, \boldsymbol{b}^n_-)$ in hash space, where "+" indicates the instance is semantically relevant to $\boldsymbol{b}^m$ and "-" is on the contrary.

With the quintuplets we demonstrate the novel weighted quintuplet loss of the first modality as follows:

$$\begin{aligned} L^m_{qui} = \sum^{R} & \left(\max\left(\omega \boldsymbol{H}\left(\boldsymbol{b}^m, \boldsymbol{b}^m_+\right) - \mu \boldsymbol{H}\left(\boldsymbol{b}^m, \boldsymbol{b}^m_-\right) + p, 0\right)\right. \\ & \left. + \max\left(\omega \boldsymbol{H}\left(\boldsymbol{b}^m, \boldsymbol{b}^n_+\right) - \mu \boldsymbol{H}\left(\boldsymbol{b}^m, \boldsymbol{b}^n_-\right) + q, 0\right)\right) \end{aligned} \tag{9}$$

where $p, q$ are margin parameters, $\boldsymbol{H}(\cdot)$ denotes the Hamming distance, $\omega$ is a weight-control constant larger than 1, and $\mu$ is less than 1. As is shown in Fig. 4, with the same threshold $\omega$ and $\mu$, the distance between similar instances should be smaller while the distance between unsimilar instances should be larger, which can facilitate the discriminating ability of the network.

Equation (9) is a function in discrete form, which is hard to optimize. To avoid the NP-hard problem, we relax the discrete

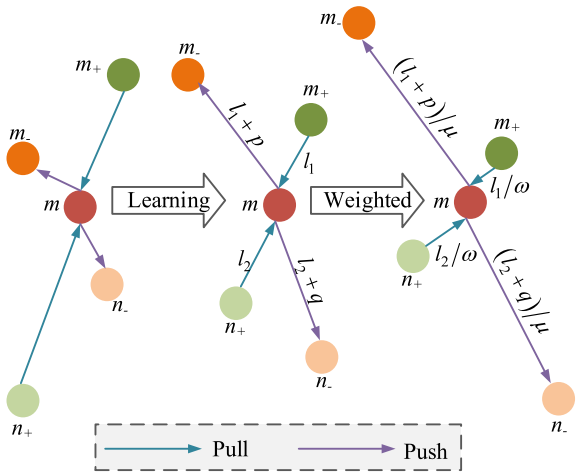



Fig. 4. Weighted quintuplets we built in hash space.

hash code to be continuous, and replace $\boldsymbol{b}^*$ with $\boldsymbol{h}^*$, so that we can directly optimize (9). The replaced objective function is

$$L_{qui}^{m} = \sum^{R} \left(\max\left(\omega \boldsymbol{E}\left(\boldsymbol{h}^{m}, \boldsymbol{h}_{+}^{m}\right) - \mu \boldsymbol{E}\left(\boldsymbol{h}^{m}, \boldsymbol{h}_{-}^{m}\right) + p, 0\right) + \max\left(\omega \boldsymbol{E}\left(\boldsymbol{h}^{m}, \boldsymbol{h}_{+}^{n}\right) - \mu \boldsymbol{E}\left(\boldsymbol{h}^{m}, \boldsymbol{h}_{-}^{n}\right) + q, 0\right)\right) \tag{10}$$

where $\boldsymbol{E}(\cdot)$ indicates the Euclidean distance. Following the same manner, the quintuplet loss of the other modality can be formulated as follows:

$$L_{\text{qui}}^{n} = \sum^{R} \left(\max\left(\omega \boldsymbol{E}\left(\boldsymbol{h}^{n}, \boldsymbol{h}_{+}^{n}\right) - \mu \boldsymbol{E}\left(\boldsymbol{h}^{n}, \boldsymbol{h}_{-}^{n}\right) + p, 0\right) + \max\left(\omega \boldsymbol{E}\left(\boldsymbol{h}^{n}, \boldsymbol{h}_{+}^{m}\right) - \mu \boldsymbol{E}\left(\boldsymbol{h}^{n}, \boldsymbol{h}_{-}^{m}\right) + q, 0\right)\right) \tag{11}$$

and the total quintuplet loss is the combination of (10) and (11)

$$L_{\text{qui}} = L_{\text{qui}}^{m} + L_{\text{qui}}^{n}. \tag{12}$$

In (12), continuous hash code is used to approximately replace the binary codes. In fact, we use binary codes when retrieving, the quantization loss in the construction of discrete hash code with the sign function cannot be ignored. In other words, we need to make the real-valued hash codes as close as binary form, and we use the binary loss to achieve this, which is defined as follows:

$$L_{\text{bin}} = \sum^{R} \left(\|\boldsymbol{b}^{m} - \boldsymbol{h}^{m}\|_{F}^{2} + \|\boldsymbol{b}^{n} - \boldsymbol{h}^{n}\|_{F}^{2}\right). \tag{13}$$

### D. Deep Adversarial Network for Modality Confusing

In Section III-B and Section III-C, we mainly explore the maintenance of discrimination and the transfer of similarities, although the weighted quintuplet loss considers samples from all modalities, there are still great differences between cross-modal features and hash codes. Increasing the similarities of cross-modal samples with the same semantic label and eliminating the modality gap can improve the effect of cross-modal RS vessel image retrieval, and following this, we add the deep adversarial constraint to both ABCN and DHN, trying to achieve a consistent representation across modalities.

In specific, we modify the generator and discriminator in GAN and design two discriminators for cross-modal features and hash codes separately, which are used to distinguish the belonging modality, and the networks of different modalities are used as generators. Take the deep adversarial feature learning as an example, the output cascaded feature $\boldsymbol{f}_{\text{cas}}^{m}$ is considered as the real feature whereas $\boldsymbol{f}_{\text{cas}}^{n}$ is fake, and the feature discriminator $\boldsymbol{D}_f$ tries to distinguish which one is true. When $\boldsymbol{D}_f$ cannot distinguish features from different modalities, the modalities confusion is achieved. In other words, the discriminator can distinguish the differences between different modality features, and the hash discriminator $\boldsymbol{D}_h$ acts in a similar way to $\boldsymbol{D}_f$. The above process can be considered as a classification problem, so we use cross-entropy loss to realize it, and the deep adversarial loss is formulated as follows:

$$\begin{aligned} L_{\text{adv}} &= L_{\text{adv}}^{f} + L_{\text{adv}}^{h} \\ &= -\frac{1}{R}\Big[\sum \left(\log\left(\boldsymbol{D}_f\left(\boldsymbol{f}_{\text{cas}}^{m}; \Phi_f\right)\right) + \log\left(1 - \boldsymbol{D}_f\left(\boldsymbol{f}_{\text{cas}}^{n}; \Phi_f\right)\right)\right) + \sum \left(\log\left(\boldsymbol{D}_h\left(\boldsymbol{h}_{\text{cas}}^{m}; \Phi_h\right)\right) \right. \\ &\quad \left. + \log\left(1 - \boldsymbol{D}_h\left(\boldsymbol{h}_{\text{cas}}^{n}; \Phi_h\right)\right)\right)\Big] \end{aligned} \tag{14}$$

where $L_{\text{adv}}^{f}, L_{\text{adv}}^{h}$ donate the deep adversarial loss of features and hash codes and $\Phi_f, \Phi_h$ are trainable parameters of discriminators.

By optimizing (14), the generators can generate modality-invariant representations of features and hash codes, when the boundaries between different modalities disappear, it is much easier to achieve a precise retrieval.

### E. Optimization

In summary, the final objective function includes $L_{\text{att}}$, $L_{\text{qui}}$, $L_{\text{bin}}$ and $L_{\text{adv}}$ four parts and there are ABCN, DHN, and DAN three modules need to be optimized with training., We combine ABCN and DHN and optimize them in an end-to-end way, and the combined loss function is defined as follows:

$$L_{\text{com}} = \alpha L_{\text{qui}} + \beta L_{\text{att}} + \gamma L_{\text{bin}} \tag{15}$$

where $\alpha, \beta$, and $\gamma$ are hyper-parameters to control the contributions to the $L_{\text{com}}$, and $L_{\text{adv}}$ is used to optimize DAN separately.

After the combination, there remain two parts to optimize, and the overall objective function can be formulated as follows:

$$\begin{aligned} &\min_{\Theta^*, \theta^*, \Phi_f, \Phi_h} \left(L_{\text{com}} + L_{\text{adv}}\right) \\ &= \min_{\Theta^*, \theta^*, \Phi_f, \Phi_h} \left(\alpha L_{\text{qui}} + \beta L_{\text{att}} + \gamma L_{\text{bin}} + L_{\text{adv}}\right). \end{aligned} \tag{16}$$

We adopt an alternating strategy to optimize the parameters in both two parts. In specific, we only optimize the parameters in one part at a time with the others fixed, and parameters in different parts are optimized alternatively. The detailed optimization algorithm of the DACH is summarized in Algorithm 1, where $\tau$

**Algorithm 1:** Optimization for DACH.

**Input:** The training dataset $D$ and label set $L$
**Output:** Parameters $\Theta^*, \theta^*, \Phi_f, \Phi_h$ of networks
**Initialization:** Initialize the network parameters and hyper-parameters
**repeat:**
  **for** iteration $= 12, \ldots, [R/b]$ do
    Update $\Theta^*, \theta^*$ by descending their gradients:
    $\Theta^*, \theta^* \leftarrow \Theta^*, \theta^* - \tau \nabla_{\Theta^*, \theta^*} L_{com}$
    Update $\Theta^*, \theta^*$ by descending their gradients:
    $\Phi_f, \Phi_h \leftarrow \Phi_f, \Phi_h - \tau \nabla_{\Phi_f, \Phi_h} L_{adv}$
  **end**
**until** *convergence* or reach the maximal training epoch $E$

is the learning rate, $b$ is the batch size, and the whole network uses the backpropagation algorithm to update the gradient.

## IV. Experiment and Analysis

To verify the effectiveness of our proposed DACH on cross-modal vessel image retrieval tasks, we conduct extensive evaluations on two published bimodal vessel image datasets, VAIS [67] and MPSC [68], the detailed information of the two datasets is given in Section IV-A, and in Section IV-B we introduce the implementation details of the DACH and the evaluation metrics we use in experiments, the following parts are the overall performance of DACH and some further analysis.

### A. Dataset Introduction

1) *MPSC*: MPSC is acquired by the GF-2 satellite which can capture panchromatic and multispectral images simultaneously. The resolution of the panchromatic image is 1 m and that of the multispectral image is 4 m. Besides, the multispectral images contain four-band spectral of near-infrared, R, G, and B, whereas panchromatic images only have one. Li et al. [68] sliced the images containing vessel targets from the images obtained by the GF-2 satellite and paired them, they collected 2632 paired vessel target images in total and divided them into six categories. They also gave the official division of the training set and the test set, and 500 paired images are used for testing.

2) *VAIS*: VAIS is the world's first published bimodal vessel image dataset in visible (VIS) and infrared (IR) used for autonomous sea surface vessels, which contains 2865 images (1623 VIS and 1242 IR) and there are 1088 paired images. The whole dataset includes 264 uniquely named vessels in total and the collectors divided them into six coarse-grained categories as well. Among the VAIS, there are 154 nighttime IR images and the bounding boxes of the images range from hundreds of pixels to millions, which makes the feature extraction more challenging in retrieval tasks. It should also be noted that VAIS is not captured by RS satellite and we only use the 1088 paired images for complemental evaluation of the effectiveness of DACH. We select some representative example vessel images in MPSC and VAIS shown in Fig. 5, and the detailed statistics of the two biomodal datasets are given in Table I.

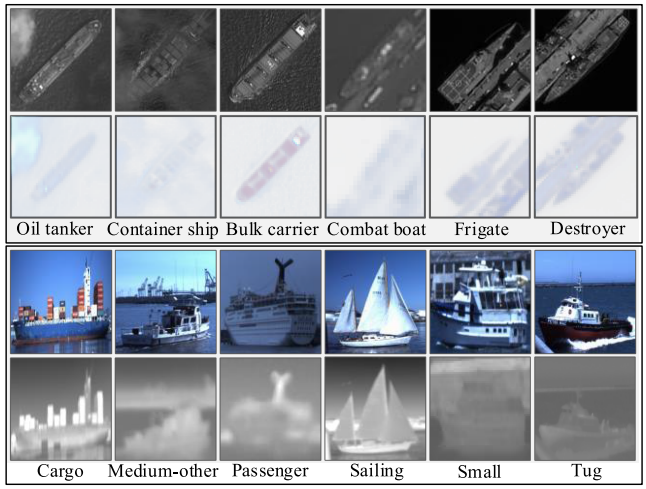



Fig. 5. Examples of the two datasets.

TABLE I
General Statistics of the Two Datasets

| Dataset | Categories | Amount | Train/Test |
|---|---|---|---|
| MPSC | Destroyer | 99 | 2132/500 |
| | Frigate | 179 | |
| | Combat boat | 261 | |
| | Bulk carrier | 919 | |
| | Container ship | 387 | |
| | Oil tanker | 787 | |
| VAIS | Cargo | 146 | 539/549 |
| | Medium-other | 138 | |
| | Passenger | 117 | |
| | Sailing | 284 | |
| | Small | 353 | |
| | Tug | 50 | |

TABLE II
Configurations of Different Modules

| Module name | Layer name | Shape | Active function |
|---|---|---|---|
| $\boldsymbol{D}_f$ | FC1-1 | 2048→1024 | ReLU |
| | FC1-2 | 1024→512 | ReLU |
| | FC1-3 | 512→1 | None |
| $\boldsymbol{D}_h$ | FC2-1 | $K$→$K/2$ | ReLU |
| | FC2-2 | $K/2$→$K/4$ | ReLU |
| | FC2-3 | $K/4$→1 | None |
| ***Classifier*** | FC3-1 | 2048→1024 | None |
| | FC3-2 | 1024→6 | Sigmoid |
| ***Hash Layer*** | FC4-1 | 2048→$K$ | Tanh |

### B. Experiment Settings

*1) Implementation Details:* Before evaluation, the DACH still needs some preparation to complement the whole framework. Since we use Resnet50 as the backbone for feature extraction, we firstly resize all images into 224×224 and initialize Resnet50 with weights pre-trained by ImageNet. As for the two discriminators $\boldsymbol{D}_f$ and $\boldsymbol{D}_h$ and the classifier, hash layer, we use FC layers with active functions to build them, the details are given in Table II. Regarding the hyper-parameters in DACH, we use ten-fold cross validation in training and set $\alpha = 1.0$, $\beta = 1.0$,

$\gamma = 0.1$, $G = 16$, $p = 0.3$, $q = 0.3$, $\omega = 1.04$, $\mu = 0.95$ by default, and the influence of them is analyzed in Section IV-E.

The entire network is trained on a Windows 10 Workstation with Intel i9-11900K CPU, Nvidia GTX 3090 GPU, and 64GB RAM in PyTorch, and the Adam(Adaptive momentum) is employed to optimize the network with a learning rate $\tau = 0.0001$ and a maximal training epoch $E = 200$, the batch size $b$ is set to 96. The default hash code length $K$ is set to 256 unless otherwise stated.

To demonstrate our DACH's effectiveness on cross-modal vessel image retrieval tasks, we select several hashing and nonhashing, adversarial and nonadversarial methods for a comprehensive comparison. Since there is no relevant research on this topic and no published retrieval results on the two datasets, we mainly choose those cross-modal methods that work well or have open source code in the field of multimedia and RS, which include DADH [52], AGAH [44], DCMHN [18], MIAN [70], DSCMR [69], GASAnet [57], DCMR [71], Distillation [17], MAN [55], SDML [31], and replace all the feature extractors with Resnet50 pretrained by ImageNet for a fair comparison.

*2) Evaluation Metrics:* We conduct the following four kinds of cross-modal vessel image retrieval tasks on the two public datasets in the experiments.

1) Retrieving PAN samples using MS as queries (M2P).
2) Retrieving MS samples using PAN as queries (P2M).
3) Retrieving IR samples using VIS as queries (V2I).
4) Retrieving VIS samples using IR as queries (I2V).

To verify the retrieval effect of different methods in full views, we adopt three commonly used evaluation metrics namely, the MAP, the precision-recall (PR) curves, and the precision at $k$ (P@$k$). Given a query instance, p@$k$ indicates the precision of the top $k$ returned images, before introducing MAP, we first give the definition of average precision (AP) based on p@$k$

$$\mathrm{AP} = \frac{1}{Q}\sum_{k=1}^{R}\delta(k)\,\mathrm{p@}k \tag{17}$$

where $Q$ is the number of entities and $Q'$ is the number of semantic relevant instances of the query in the database, If $k$th instance belongs to the same semantic category with the query, $\delta(k) = 1$, otherwise $\delta(k) = 0$. Given a query set with $W$ queries, MAP is defined as follows:

$$\mathrm{MAP} = \frac{1}{W}\sum_{i}^{W}\mathrm{AP}_i. \tag{18}$$

P@$k$ and R@$k$ can be obtained through the following expressions:

$$\begin{cases}\mathrm{P@}k = \frac{1}{W}\sum\limits_{i}^{W}(\mathrm{p@}k)_i \\ \mathrm{R@}k = \frac{1}{W}\sum\limits_{i}^{W}(\mathrm{r@}k)_i\end{cases} \tag{19}$$

where r@$k$ indicates the recall of the top $k$ returned images. The PR curve can be obtained by varying $k$.

The three evaluation metrics measure the effectiveness of the methods in different views, and the higher they are, the more effective the method is.

TABLE III
COMPARED RESULTS OF DIFFERENT ABCN SETTINGS

| Task | M2P | P2M | V2I | I2V |
|---|---|---|---|---|
| ABCN-1 | 0.621 | 0.590 | 0.443 | 0.426 |
| ABCN-2 | 0.608 | 0.586 | 0.449 | 0.428 |
| ABCN-3 | 0.640 | 0.601 | 0.446 | 0.420 |
| ABCN-4 | 0.601 | 0.585 | 0.432 | 0.410 |
| ABCN | **0.681** | **0.645** | **0.474** | **0.450** |

### C. Effective of Attention-Based Cascaded Network

Before the overall evaluation of DACH, we first quantitatively evaluates the effectiveness of the ABCN in the following four forms particularly.

1) *ABCN*-1: ABCN without feature in depth(directly concatenating features without "+", i.e., $\hat{\boldsymbol{f}}^*_{\mathrm{cur}} = \boldsymbol{f}^*_{\mathrm{cur}}$).
2) *ABCN*-2: ABCN without feature in width(directly using $f^*_{\mathrm{dep}}$ without concatenating).
3) *ABCN*-3: ABCN without self-attention module (replacing with linear layer).
4) *ABCN*-4: ABCN without deep cascaded network (replacing with ResNet50).

We select MAP as the evaluation protocol and report the results on two datasets in Table III, where the terms in bold indicate the best performance.

According to Table III, we can easily conclude that each component we design in ABCN can contribute to the final retrieval performance, and the combination of them leads to optimal results in the above four retrieval tasks. In contrast with Resnet50, the deep cascaded network brings an 8.0%, 6.0%, and 4.2%, 4.0% improvement in MAP in four retrieval tasks, respectively. The other components also have consistent advantages over ResNet50, which proves the superiority of ABCN. That is mainly attributed to the ABCN integrating the features of different levels in a cascaded structure and the self-attention module further facilitates the feature fusion. In this way, abundant discriminative information can be integrated into the cascaded feature, which helps a precise retrieval.

### D. Overall Performance

In this section, we compare our proposed method DACH with several state-of-the-art methods under the three evaluation metrics mentioned above. The P@$k$ values are computed and drawn in Fig. 6, and the MAP results of the two datasets are summarized in Tables IV and V. As for the PR curve, we collect experiment data under different retrieval tasks and give an exhibition in Figs. 7–10.

As given in Tables IV and V, DACH achieves the best results in all the four retrieval tasks across different hash code lengths compared with hashing methods, and is superior to the nonhashing methods as well. The peak value on MPSC can reach to 68.1% and 64.5% when the length of the hash code is set to 256, 5.4%, and 3.7% higher than the second-best DSCMR, and much higher than the hashing methods. Since VAIS has lower resolutions and fewer training images, it is difficult to achieve a retrieval as precise as MPSC. However, DACH still has significant advantages over the compared methods, and the

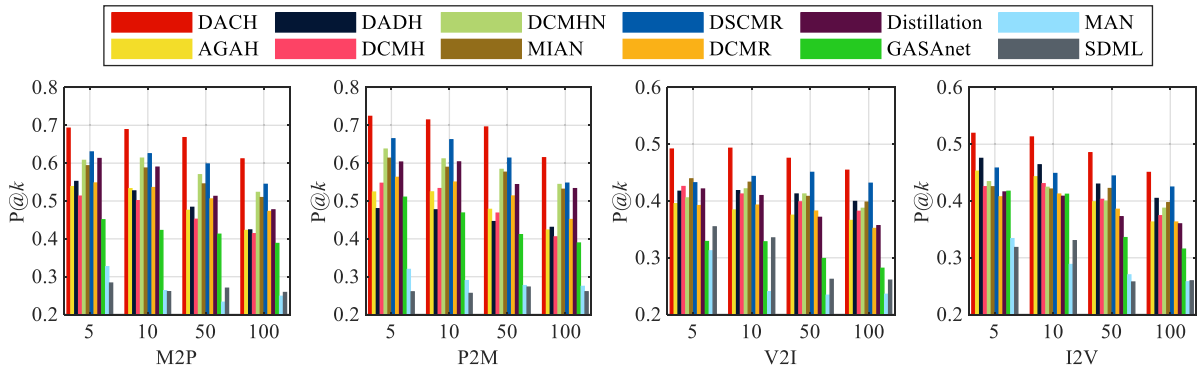



Fig. 6. P@k values of different methods.

TABLE IV
MAP OF HASHING METHODS

| Dataset | Method | Task | 32bit | 64bit | 128bit | 256bit |
|---|---|---|---|---|---|---|
| MPSC | DACH | M2P | **0.655** | **0.658** | **0.674** | **0.681** |
| | | P2M | **0.634** | **0.635** | **0.644** | **0.645** |
| | AGAH | M2P | 0.437 | 0.444 | 0.432 | 0.437 |
| | | P2M | 0.446 | 0.457 | 0.443 | 0.446 |
| | DADH | M2P | 0.455 | 0.458 | 0.446 | 0.432 |
| | | P2M | 0.453 | 0.461 | 0.470 | 0.439 |
| | DCMH | M2P | 0.378 | 0.400 | 0.332 | 0.440 |
| | | P2M | 0.346 | 0.370 | 0.268 | 0.422 |
| | DCMHN | M2P | 0.598 | 0.589 | 0.601 | 0.599 |
| | | P2M | 0.563 | 0.561 | 0.593 | 0.568 |
| | MIAN | M2P | 0.571 | 0.584 | 0.611 | 0.625 |
| | | P2M | 0.577 | 0.592 | 0.601 | 0.593 |
| VAIS | DACH | V2I | **0.468** | **0.467** | **0.463** | **0.474** |
| | | I2V | **0.439** | **0.443** | **0.440** | **0.450** |
| | AGAH | V2I | 0.390 | 0.401 | 0.387 | 0.368 |
| | | I2V | 0.369 | 0.390 | 0.383 | 0.361 |
| | DADH | V2I | 0.389 | 0.398 | 0.401 | 0.413 |
| | | I2V | 0.386 | 0.392 | 0.387 | 0.388 |
| | DCMH | V2I | 0.401 | 0.404 | 0.403 | 0.396 |
| | | I2V | 0.384 | 0.368 | 0.384 | 0.372 |
| | DCMHN | V2I | 0.402 | 0.399 | 0.411 | 0.428 |
| | | I2V | 0.387 | 0.379 | 0.402 | 0.404 |
| | MIAN | V2I | 0.431 | 0.421 | 0.429 | 0.435 |
| | | I2V | 0.377 | 0.347 | 0.383 | 0.390 |

TABLE V
MAP OF NON-HASHING METHODS

| Dataset | Task | DSCMR | GASAnet | DCMR | Distillation | MAN | SDML |
|---|---|---|---|---|---|---|---|
| MPSC | M2P | **0.627** | 0.391 | 0.508 | 0.599 | 0.262 | 0.269 |
| | P2M | **0.608** | 0.379 | 0.459 | 0.585 | 0.287 | 0.271 |
| VAIS | V2I | **0.433** | 0.304 | 0.399 | 0.408 | 0.255 | 0.273 |
| | I2V | **0.420** | 0.304 | 0.375 | 0.393 | 0.263 | 0.274 |

MAP values are 4.1% and 3.0% higher than the second-best. We can also find that the MAP of M2P is relatively superior to P2M, and higher values are achieved in V2I than in I2V.

Besides, longer hash codes provide higher retrieval accuracy generally. In general, due to the information loss in the binarization process, nonhashing methods usually achieve better performance than hashing methods, that is why DSCMR outperforms the compared hashing methods. Different from conventional hashing methods mentioned above, DACH tries to improve the retrieval performance throughout the whole network, including the component evaluated in Section IV-C and the two-stage adversarial learning, and all these improvements help DACH to be competent for vessel image retrieval task.

Moreover, extensive evaluations are conducted and we can capture the superiority of DACH in Fig. 6 intuitively. Fig. 6 shows the P@$k$ values of all the hashing and nonhashing methods. Similarly, we can find that the DACH outperforms the compared methods in all retrieval tasks and the retrieval accuracy decreases slowly with the increasing of $K$, while most of the compared methods decrease sharply or maintain a low precision.

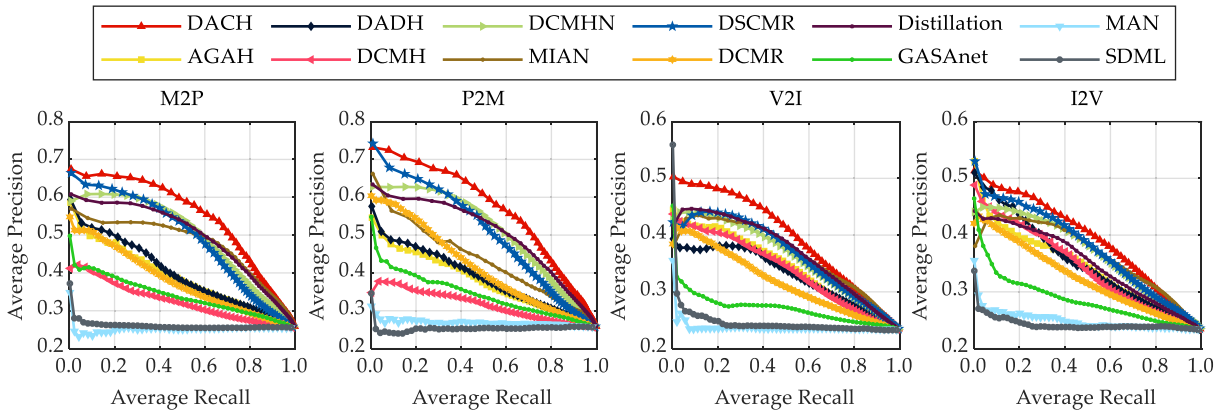



Fig. 7. PR curves of different methods in 32 b.

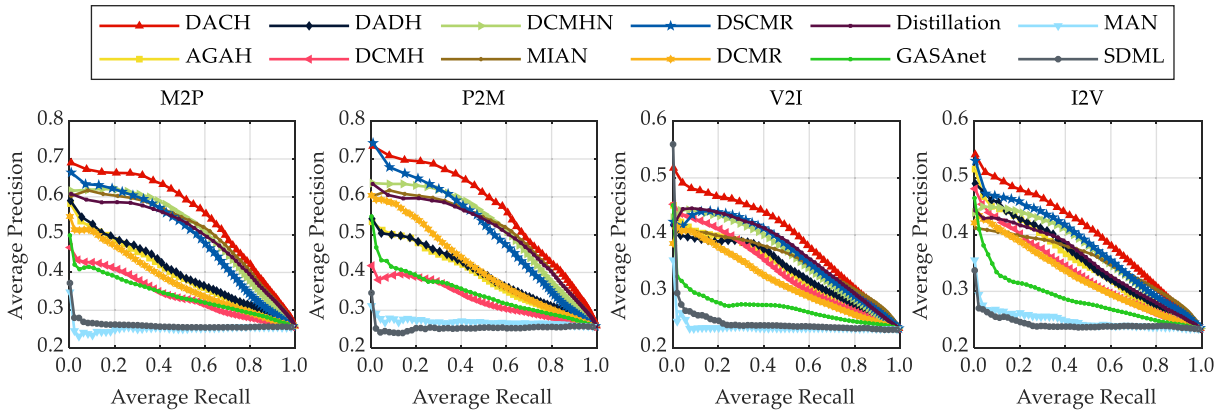



Fig. 8. PR curves of different methods in 64 b.

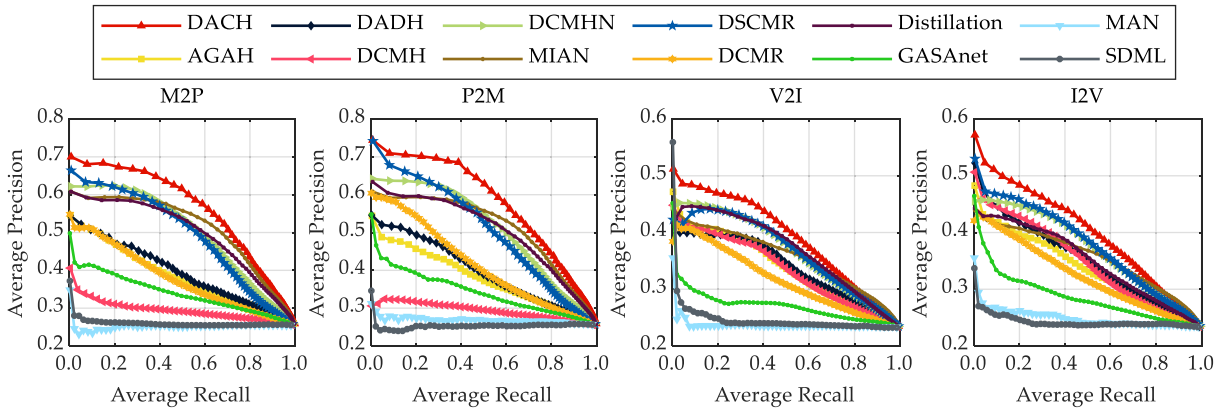



Fig. 9. PR curves of different methods in 128 b.

It is clear that more retrieval quantity means higher retrieval difficulty, the results show that DACH can adapt to the need for accurate retrieval of different numbers of images. Although it decreases when the $K$ rises, it still maintains an obvious advantage compared with other methods.

To further observe the retrieval behavior of different methods when $K$ varies, we calculate the PR values and give the PR curves in Figs. 7–10. The PR curves show that DACH has a higher retrieval precision than that of the compared methods under the same recall values in all hash code lengths, and the advantage

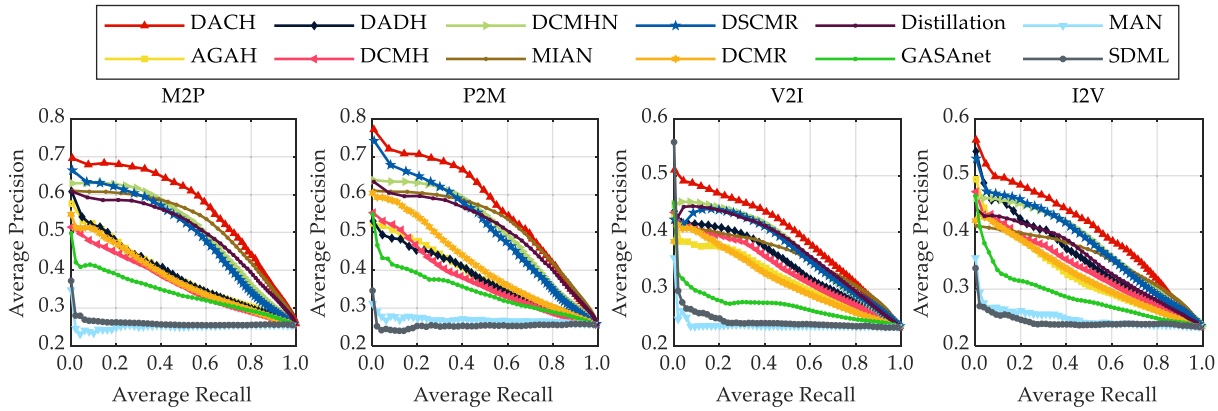



Fig. 10. PR curves of different methods in 256 b.

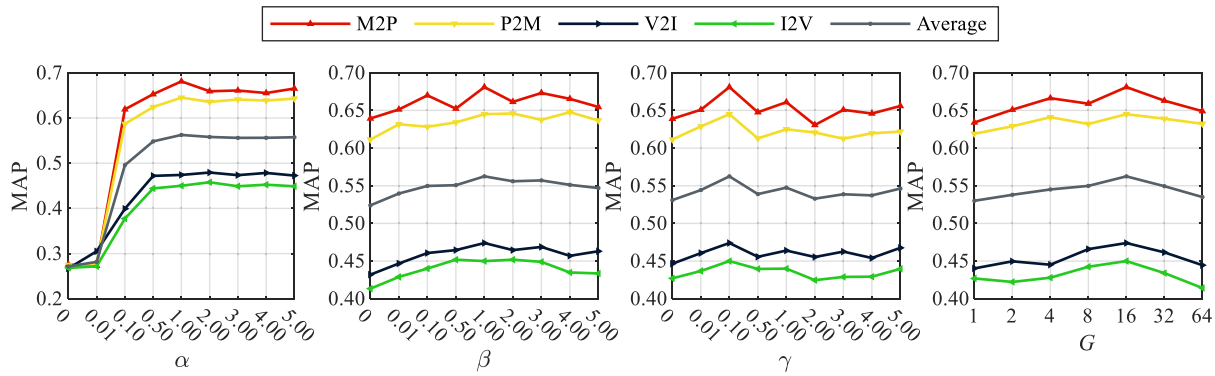



Fig. 11. MAP on parameters $\alpha$, $\beta$, $\gamma$, and $G$.

is much more obvious when $K$ rises. Take the PR curve on I2 V as an example, when $K$ is set to 32, the AP of DACH is lower than that of DSCMR at the beginning, while DACH surpasses the DSCMR gradually with $K$ rising. The encouraging results of DACH show its stable performance across different hash code lengths and consistent superiority across different quantities of retrieval samples, which demonstrate DACH's advantage in effective feature and hash learning.

In order to better visualize the training process, we also give the loss change of the two biomodal vessel image dataset in Fig. 13.

### E. Parameters Analysis

In this section, we conduct experiments on four retrieval tasks to comprehensively illustrate the effect of the hyperparameters we set on the performance. For clarity, we examine the effect of each hyperparameter on the results separately and divide them into two parts, the final MAP values under different settings are exhibited in Figs. 11 and 12 and the line in gray is the average MAP of the four tasks. As for the hyperparameters $\alpha$, $\beta$, and $\gamma$ in (16), we set them to 0, 0.01, 0.1, 0.5, 1, 2, 3, 4, 5 uniformly in the experiment and illustrate their effect on retrieval.

First of all, we can see that different hyperparameter settings have the same influence on the four retrieval tasks basically, and the optimal performance is acquired according to the average MAP when $\alpha$, $\beta$, and $\gamma$ are set to 1.0, 1.0, and 0.1, respectively. When $\alpha$ is less than 0.5, we can see that the retrieval performance is terrible in all the four tasks. When $\alpha$ is larger than 1, it achieves a high and stable retrieval accuracy. As defined in (15), $\alpha$ controls the contribution of the weighted quintuplets loss to the $L_{\text{com}}$, and a proper value can significantly strengthen the ability of the intramodality similarity and intermodality discrimination modeling. In contrast with $\alpha$, different settings of $\beta$ and $\gamma$ have relatively small fluctuations, but the difference between different results can still reach as much as 3.85% and 2.96% in average MAP, which cannot be ignored. To achieve the best performance, we use the hyperparameter setting based on the best average MAP to conduct the other experiments. $G$ is the number of heads in multihead self-attention modules, and it has been verified in previous experiments that the higher value brings better performance. Based on the average MAP as well, we set the final value of $G$ as 16.

Fig. 12 intuitively shows the MAP changes under different settings of $\omega$, $\mu$, $p$, and $q$, which are all hyperparameters in the weighted quintuplets loss we designed, and we can get the best parameter settings by analyzing the results. As is shown in

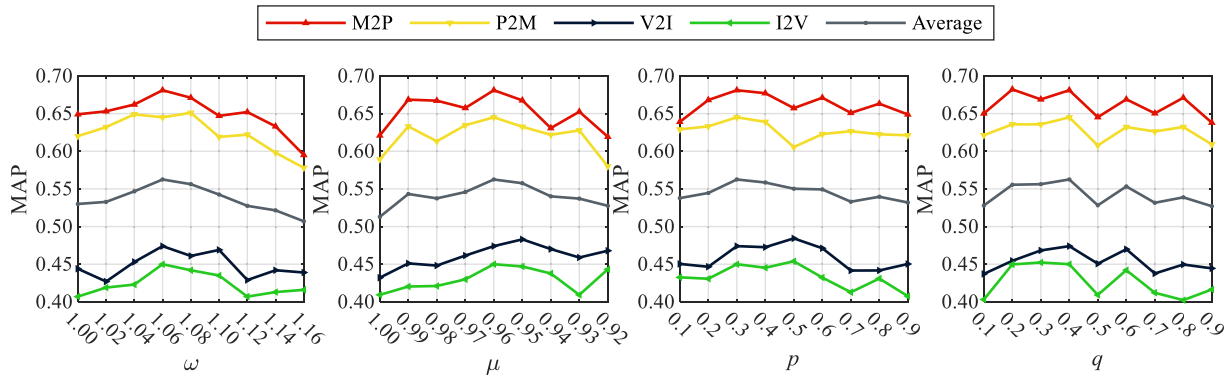



Fig. 12. MAP on parameters $\omega$, $\mu$, $p$, and $q$.

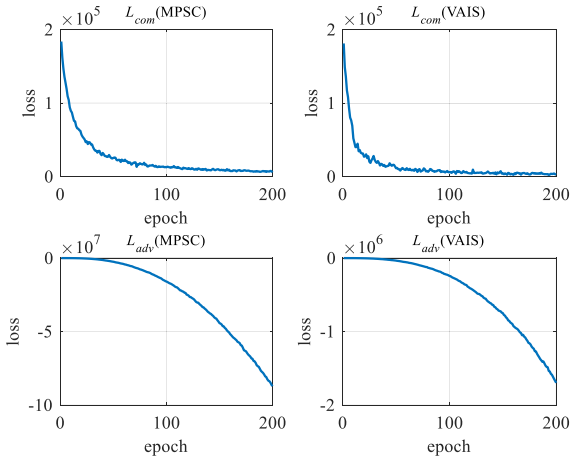



Fig. 13. Loss change of different dataset.

Fig. 12, the best average MAP value is achieved when $\omega$ and $\mu$ are set to 1.06 and 0.96, respectively, better than the performance when $\omega$ and $\mu$ are both set to 1.00. When $\omega$ further rises or $\mu$ continues to decrease, the performance gets worse. The results confirm that the design of $\omega$ and $\mu$ is helpful to a precise retrieval performance. The margin parameters play a similar role as they can control the distances between positive pairs and negative pairs across different modalities. According to the average MAP values, the setting of 0.3 and 0.4 on $p$ and $q$ can yield the best performance, which is slightly better than the other settings.

### F. Time Cost Discussion

As a hashing method, it is necessary to discuss the time cost of training and retrieving of DACH to examine its computational complexity. In this section, we record the training time of our DACH and the compared methods in Table VI, and make a comparison of retrieving speed between hash codes and continuous features, the detailed results are given in Table VII. For a fair comparison, we conduct experiments of DACH and all the compared methods on the same computing platform and only count the training time of the networks ignoring the time of data preprocessing.

TABLE VI
TRAINING TIME OF DIFFERENT METHODS

| Method | MPSC | VAIS |
|---|---|---|
| DACH | 61m 54s | 27m 5s |
| AGAH | 39m 24s | 10m 25s |
| DADH | 38m 33s | 9m 39s |
| DCMH | 31m 1s | 7m 2s |
| DCMHN | 34m 38s | 8m 57s |
| DSCMR | 48m 43s | 23m 23s |
| DCMR | 29m 22s | 6m 38s |
| Distillation | 50m 16s | 23m 44s |
| GASAnet | 40m 55s | 21m 32s |
| MAN | 178m 27s | 45m 10s |
| SDML | 87m 24s | 28m 7s |
| MIAN | 122m 21s | 43m 11s |

TABLE VII
RETRIEVAL SPEED OF OUR PROPOSED DACH (UNIT: SECOND)

| Task | | M2P | P2M | V2I | I2V |
|---|---|---|---|---|---|
| Real-valued hash codes | 32bit | 0.0688 | 0.0698 | 0.0698 | 0.0768 |
| | 64bit | 0.0698 | 0.0708 | 0.0828 | 0.0808 |
| | 128bit | 0.0738 | 0.0748 | 0.0848 | 0.0848 |
| | 256bit | 0.0748 | 0.0768 | 0.0878 | 0.0908 |
| Binary hash codes | 32bit | 3.0359 | 3.0179 | 3.6463 | 3.6493 |
| | 64bit | 3.1017 | 3.0709 | 3.6482 | 3.6881 |
| | 128bit | 3.1426 | 3.1466 | 3.7460 | 3.7689 |
| | 256bit | 3.2752 | 3.2782 | 3.7580 | 3.8477 |
| ResNet50 features | | 3.8956 | 3.8707 | 4.5329 | 4.5967 |

From Table VI we can see that hashing methods generally have a lower time cost in training than nonhashing methods, which means the binary form of features can contribute to faster construction of the retrieval models. Besides, among the several well-performed methods, the longer training time brings the better performance combined with the results in Tables IV and V. Since DACH consists of three enhanced modules needed to be optimized, the time cost in training is relatively higher than the other hashing methods, but the encouraging retrieval performance makes it acceptable in retrieval tasks.

The advantage of hashing method is mainly reflected in the retrieval process since the calculation of Hamming distance is much faster than Euclidean distance. To give a clear comparison, we count the time consumption of binary hash codes and real-valued hash codes, what is more, we also conduct time cost

TABLE VIII
RETRIEVAL PERFORMANCE OF DIFFERENT DACH SETTINGS

| Dataset | Network | Task | P@5 | P@10 | P@50 | P@100 | MAP |
|---|---|---|---|---|---|---|---|
| MPSC | DACH-1 | M2P | 0.628 | 0.648 | 0.645 | 0.593 | 0.656 |
| | | P2M | 0.515 | 0.618 | 0.653 | 0.610 | 0.620 |
| | DACH-2 | M2P | 0.662 | 0.659 | 0.651 | 0.598 | 0.660 |
| | | P2M | 0.700 | 0.676 | 0.673 | 0.599 | 0.636 |
| | DACH-3 | M2P | 0.683 | 0.677 | 0.651 | 0.592 | 0.666 |
| | | P2M | 0.655 | 0.690 | 0.678 | 0.596 | 0.641 |
| | DACH-4 | M2P | 0.662 | 0.657 | 0.643 | 0.588 | 0.642 |
| | | P2M | 0.627 | 0.636 | 0.655 | 0.574 | 0.613 |
| | DACH-5 | M2P | 0.653 | 0.665 | 0.646 | 0.594 | 0.669 |
| | | P2M | 0.683 | 0.696 | 0.692 | 0.597 | 0.639 |
| | DACH | M2P | **0.692** | **0.688** | **0.667** | **0.611** | **0.681** |
| | | P2M | **0.724** | **0.714** | **0.695** | **0.614** | **0.645** |
| VAIS | DACH-1 | V2I | 0.444 | 0.446 | 0.439 | 0.420 | 0.438 |
| | | I2V | 0.497 | 0.490 | 0.434 | 0.409 | 0.414 |
| | DACH-2 | V2I | **0.502** | 0.492 | 0.484 | 0.465 | 0.471 |
| | | I2V | 0.476 | 0.477 | 0.475 | **0.456** | 0.450 |
| | DACH-3 | V2I | 0.481 | 0.489 | 0.474 | 0.450 | 0.469 |
| | | I2V | 0.502 | 0.511 | 0.505 | 0.467 | 0.453 |
| | DACH-4 | V2I | 0.473 | 0.481 | 0.472 | 0.439 | 0.458 |
| | | I2V | 0.467 | 0.471 | 0.472 | 0.443 | 0.445 |
| | DACH-5 | V2I | 0.487 | 0.487 | **0.479** | 0.461 | 0.471 |
| | | I2V | 0.479 | 0.491 | 0.486 | 0.470 | 0.458 |
| | DACH | V2I | 0.491 | **0.493** | 0.475 | **0.454** | **0.474** |
| | | I2V | **0.519** | **0.513** | **0.485** | 0.450 | **0.450** |

experiments on ResNet50 features to further demonstrate the superiority of DACH. Since the valid set is small, here we use the whole valid dataset as the query set and count the overall retrieval time.

As is clearly given in Table VII, the binary hash codes have the lowest time consumption among the three forms, which embodies its significant advantage in retrieval, and the longer hash codes have higher time consumption, while much less than the increase of the length of the hash codes. In contrast, the ResNet50 features with a dimension of 2048 have a much higher time cost, this is because the longer features need more time for similarity measuring, and the calculation of the Euclidean distance further increases the time consumption. So that we can shorten the features into a proper length and convert them into binary form to achieve a fast retrieval.

### G. Ablation Study

In this section, extensive ablation experiments are conducted to demonstrate the effect of the modules and strategies we designed in DACH and figure out their contributions. We have already illustrated the impact of different components in ABCN on image retrieval in Section IV-C, we mainly focus on the other two modules. In particular, we redesign the DACH in five novel forms according to the components in DHN and DAN, which can be summarized as follows.

1) *DACH1:* Replacing weighted quintuplet loss with triplet loss.
2) *DACH2:* Only imposing deep feature adversarial learning on training.
3) *DACH3:* Only imposing deep hash adversarial learning on training.
4) *DACH4:* Removing both deep hash adversarial learning and deep feature adversarial learning in training.
5) *DACH5:* Removing the category attention constraint.

For a comprehensive comparison, we select MAP and P@$k$ as evaluation indicators to numerically assess these networks' retrieval effects on different tasks. The retrieval results are given in Table VIII.

By analyzing the results we can find that the original DACH remains obvious advantages over the redesigned DACHs and achieves the best result in almost all the retrieval tasks under different evaluation indicators. For further contrast, we can conclude the several following points. First, the adversarial learning does bring an encouraging improvement on retrieval, and the combination of deep adversarial hash learning and deep adversarial feature learning makes a better performance than the single one when retrieving and the removal of DAN both reduce the value of MAP and P@$k$, which demonstrate that our improvement on deep adversarial learning is acceptable and effective. Take the MAP value on MPSC for example, the implemention of DAN can bring a 3.9% improvement on M2P and a 3.2% improvement on P2M. Second, the design of weighted quintuplet loss has superiority compared with the idiomatic triplet loss, and a 2.5% improvement is achieved both on M2P and P2M. Since the samples of the same modality and cross-modality are comprehensively considered when training, the samples belonging to the same semantic category can get closer both in feature space and hash space, thus bringing a better performance in retrieving. What is more, the category attention constraint also plays an active role in retrieval, as given in Table VIII, the DACH-5's performance is slightly inferior to DACH. This is because the category attention constraint can embed the semantic information into learned features, which

is more discriminative in metric space in contrast with features without embedding, and this can help us retrieve the exact image we want to a certain extent. Considering all these analyses and positive results, we can conclude that all the three modules in DACH can contribute to a better retrieval performance, and the combination of them makes DACH more competitive in various retrieval tasks.

## V. Conclusion

This article presents a DHN DACH based on adversarial learning to tackle the vessel image retrieval tasks. To overcome the shortcoming of the conventional DNN in multilevel feature extraction, ABCN is developed to capture the multiscale and complex features comprehensively, which can integrate the abundant discriminate information in a course-to-fine way and weight the information with a self-attention mechanism to achieve a robust feature presentation. The DHN is presented to improve the retrieval efficiency and the novel weighted quintuplet loss can strengthen the semantic discrimination and cross-modal consistency to a great extent. What is more, the adversarial learning imposed on both ABCN and DHN ensures the smooth transition of the above information. Extensive experiments on the only two public bimodal vessel image datasets demonstrate that our proposed DACH has superiority over many state-of-the-art cross-modal retrieval methods in both hashing and nonhashing, and can achieve competitive performance under various evaluation metrics.

However, the DACH still has some shortcomings that cannot be ignored, such as the high time-consuming and the heavy dependency on large training samples, and these are the urgent future works for us to solve.

## References


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

[2] Q. Cheng, Y. Zhou, P. Fu, Y. Xu, and L. Zhang, "A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 14, pp. 4284–4297, 2021.

[3] Z. Yuan et al., "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022, Art. no. 4404119, doi: 10.1109/TGRS.2021.3078451.

[4] Z. Yuan et al., "A lightweight multi-scale crossmodal text-image retrieval method in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022, Art. no. 5612819, doi: 10.1109/TGRS.2021.3124252.

[5] O. E. Dai, B. Demir, B. Sankur, and L. Bruzzone, "A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 11, no. 7, pp. 2473–2490, Jul. 2018.

[6] 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.

[7] L. Han, P. Li, X. Bai, C. Grecos, X. Zhang, and P. Ren, "Cohesion intensive deep hashing for remote sensing image retrieval," *Remote Sens.* vol. 12, no. 1, pp. 101, Dec. 2019, doi: 10.3390/rs12010101.

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

[9] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020, doi: 10.1109/JSTARS.2019.2961634.

[10] G. Mao, Y. Yuan, and L. Xiaoqiang, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens.*, 2018, pp. 1–7.

[11] 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. Topics Appl. Earth Observ. Remote Sens.,* vol. 12, no. 11, pp. 4644–4654, Nov. 2019.

[12] Y. Chen, X. Lu, and S. Wang, "Deep cross-modal image–voice retrieval in remote sensing," *IEEE Trans. Geosci. Remote Sens.,* vol. 58, no. 10, pp. 7049–7061, Oct. 2020.

[13] 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.

[14] 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, pp. 2789, Aug. 2020, doi: 10.3390/rs12172789.

[15] 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.

[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] W. Xiong, Z. Xiong, Y. Cui, and Y. Lv, "A discriminative distillation network for cross-source remote sensing image retrievala," in *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1234–1247, 2020, doi: 10.1109/JSTARS.2020.2980870.

[18] W. Xiong, Z. Xiong, Y. Zhang, Y. Cui, and X. Gu, "A deep cross-modality hashing network for SAR and optical remote sensing images retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 13, pp. 5284–5296, 2020, doi: 10.1109/JSTARS.2020.3021390.

[19] R. Hu, J. Yang, B. Zhu, and Z. Guo, "Research on ship image retrieval based on BoVW model under hadoop platform," in *Proc. Int. Conf. Inf. Sci. System*, 2018, pp. 156–160, doi: 10.1145/3209914.3209948.

[20] C. Tian, J. Xia, J. Tang, and H. Yin, "Deep image retrieval of large-scale vessels images based on BoW model," *Multimedia Tools Appl.,* vol. 79, no. 13-14, pp. 9387–9401, 2019.

[21] 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.

[22] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Trans. Geosci. Remote Sens.,* vol. 56, no. 2, pp. 950–965, Feb. 2018.

[23] 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.

[24] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3270–3278.

[25] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "CMIR-NET : A deep learning based model for cross-modal retrieval in remote sensing," *Pattern Recognit. Lett.,* vol. 131, pp. 456–462, 2020.

[26] Y. Sun et al., "Multisensor fusion and explicit semantic preserving-based deep hashing for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5219614, doi: 10.1109/TGRS.2021.3136641.

[27] 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, 2022, Art. no. 4700614, doi: 10.1109/TGRS.2021.3060705.

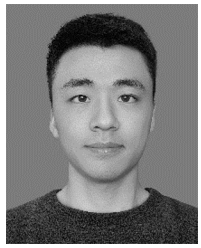
[28] Y. Chen and X. Lu, "A deep hashing technique for remote sensing image-sound retrieval," *Remote Sens.*, vol. 12, no. 1, p. 84, Dec. 2019, doi: 10.3390/rs12010084.

[29] K. Ding, B. Fan, C. Huo, S. Xiang, and C. Pan, "Cross-modal hashing via rank-order preserving," *IEEE Trans. Multimedia,* vol. 19, no. 3, pp. 571–585, Mar. 2017.

[30] X. Dong, L. Liu, L. Zhu, L. Nie, and H. Zhang, "Adversarial graph convolutional network for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1634–1645, Mar. 2022.

[31] P. Hu, L. Zhen, D. Peng, and P. Liu, "Scalable deep multimodal learning for cross-modal retrieval," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 635–644, doi: 10.1145/3331184.3331213.

[32] 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, May 2017.
[33] D. Qiao, G. Liu, F. Dong, S.-X. Jiang, and L. Dai, "Marine vessel re-identification: A large-scale dataset and global-and-local fusion-based discriminative feature learning," *IEEE Access*, vol. 8, pp. 27744–27756, 2020.
[34] A. Ghahremani, T. Alkanat, E. Bondarev, and P. H. N. de With, "Maritime vessel re-identification: Novel VR-VCA dataset and a multi-branch architecture MVR-net," *Mach. Vis. Appl.*, vol. 32, no. 71, 2021, doi: 10.1007/s00138-021-01199-1.
[35] A. Ghahremani, Y. Kong, E. Bondarev, and P. H. N. de With, "Towards parameter-optimized vessel re-identification based on IORnet," in *Proc. Comput. Sci., Int. Conf. Comput. Sci. (Lecture Notes Comput. Sci.)*, 2019, pp. 125–136.
[36] A. Ghahremani, Y. Kong, E. Bondarev, and P. H. N. de With, "Re-identification of vessels with convolutional neural networks," in *Proc. 5th Int. Conf. Comput. Technol. Appl.*, 2019, pp. 93–97, doi: 10.1145/3323933.3324075.
[37] S. Mukherjee, S. Cohen, and I. Gertner, "Content-based vessel image retrieval," in *Proc. Autom. Target Recognit. XXVI*, May 12, 2016, Art. no. 984412, doi: 10.1117/12.2234847.
[38] Z. Chen, F. Zhong, G. Min, Y. Leng, and Y. Ying, "Supervised intra- and inter-modality similarity preserving hashing for cross-modal retrieval," *IEEE Access*, vol. 6, pp. 27796–27808, 2018.
[39] X. Wang, M. Liang, X. Cao, and J. Du, "Dual-pathway attention based supervised adversarial hashing for cross-modal retrieval," in *Proc. IEEE Int. Conf. Big Data Smart Comput.*, 2021, pp. 168–171.
[40] J. Zhang, Y. Peng, and M. Yuan, "SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 489–502, Feb. 2020.
[41] 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, Sep. 2019, doi: 10.3390/rs11172055.
[42] P.-F. Zhang, Y. Li, Z. Huang, and X.-S. Xu, "Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval," *IEEE Trans. Multimedia,* vol. 24, pp. 466–479, 2022, doi: 10.1109/TMM.2021.3053766.
[43] J. Zhang and Y. Peng, "Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval," *IEEE Trans. Multimedia,* vol. 22, no. 1, pp. 174–187, Jan. 2020.
[44] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retrieval*, 2019, pp. 159–167, doi: 10.1145/3323873.3325045.
[45] Y. Wang, X. Luo, L. Nie, J. Song, W. Zhang, and X.-S. Xu, "BATCH: A scalable asymmetric discrete cross-modal hashing," *IEEE Trans. Knowl. Data Eng.,* vol. 33, no. 11, pp. 3507–3519, Nov. 2021.
[46] C.-X. Li, T.-K. Yan, X. Luo, L. Nie, and X.-S. Xu, "Supervised robust discrete multimodal hashing for cross-media retrieval," *IEEE Trans. Multimedia,* vol. 21, no. 11, pp. 2863–2877, Nov. 2019.
[47] T. Yao, Z. Zhang, L. Yan, J. Yue, and Q. Tian, "Discrete robust supervised hashing for cross-modal retrieval," *IEEE Access*, vol. 7, pp. 39806–39814, 2019.
[48] Q. Cheng, H. Huang, L. Ye, P. Fu, D. Gan, and Y. Zhou, "A semantic-preserving deep hashing model for multi-label remote sensing image retrieval," *Remote Sens.*, vol. 13, no. 24, p. 4965, Dec. 2021, doi: 10.3390/rs13244965.
[49] P. Li, X. Zhang, X. Zhu, and P. Ren, "Online hashing for scalable remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 5, p. 709, May 2018, doi: 10.3390/rs10050709.
[50] Z. Ji, W. Yao, W. Wei, H. Song, and H. Pi, "Deep multi-level semantic hashing for cross-modal retrieval," *IEEE Access*, vol. 7, pp. 23667–23674, 2019.
[51] X. Nie, B. Wang, J. Li, F. Hao, M. Jian, and Y. Yin, "Deep multiscale fusion hashing for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 401–410, Jan. 2021.
[52] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, "Deep adversarial discrete hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retrieval*, 2020, pp. 525–531, doi: 10.1145/3372278.3390711.
[53] J. Li, B. Zhang, G. Lu, and D. Zhang, "Dual asymmetric deep hashing learning," *IEEE Access*, vol. 7, pp. 113372–113384, 2019.
[54] M. Meng, H. Wang, J. Yu, H. Chen, and J. Wu, "Asymmetric supervised consistent and specific hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 30, pp. 986–1000, 2021, doi: 10.1109/TIP.2020.3038365.
[55] P. Hu, D. Peng, X. Wang, and Y. Xiang, "Multimodal adversarial network for cross-modal retrieval," *Knowl., Based Syst.*, vol. 180, pp. 38–50, 2019.
[56] P. Hu et al., "Cross-modal discriminative adversarial network," *Pattern Recognit.*, vol. 112, 2021, Art. no. 107734.
[57] J. Hong, H. Luo, Y. Yao, and Z. Tang, " Generative adversarial and self-attention based fine-grained cross-media retrieval," in *Proc. 4th Int. Conf. Vis., Image Signal Process.*, 2020, pp. 1–8, doi: 10.1145/3448823.3448825.
[58] Y. Wu, S. Wang, G. Song, and Q. Huang, "Augmented adversarial training for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 23, pp. 559–571, 2021, doi: 10.1109/TMM.2020.2985540.
[59] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4242–4251.
[60] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1047–1059, Mar. 2020.
[61] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 3626–3637, 2020, doi: 10.1109/TIP.2020.2963957.
[62] X. Ma, T. Zhang, and C. Xu, "Multi-level correlation adversarial hashing for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3101–3114, Dec. 2020.
[63] M. Li and H. Wang, "Deep semantic adversarial hashing based on autoencoder for large-scale cross-modal retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2020, pp. 1–6.
[64] X. Liu, Y.-M. Cheung, Z. Hu, Y. He, and B. Zhong, "Adversarial tri-fusion hashing network for imbalanced cross-modal retrieval," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 4, pp. 607–619, Aug. 2021.
[65] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[66] 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.
[67] M. M. Zhang, J. Choi, K. Daniilidis, M. T. Wolf, and C. Kanan, "VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 10–16.
[68] M. Li, W. Sun, X. Du, X. Zhang, and L. Yao, "Ship classification by the fusion of panchromatic image and multi-spectral image based on pseudo siamese lightweight network," *J. Phys., Conf. Ser.*, vol. 1757, no. 1, Oct. 24-25, 2020, doi: 10.1088/1742-6596/1757/1/012022.
[69] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10386–10395.
[70] Z. Zhang, H. Luo, L. Zhu, G. Lu, and H. T. Shen, "Modality-invariant asymmetric networks for cross-modal hashing," *IEEE Trans. Knowl. Data Eng.*, to be published, doi: 10.1109/TKDE.2022.3144352.
[71] G. Mao, Y. Yuan, and L. Xiaoqiang, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens.*, 2018, pp. 1–7.



**Jiaen Guo** received the B.S. and M.S. degrees in information and communication engineering from Naval Aviation University, Yantai, China, in 2020 and 2022, respectively.

His research interests include track correlation and multimodal information fusion.

**Xin Guan** received the B.E. degree in communication engineering from Liaoning University, Liaoning, China, in 1999, and the M.E. and Ph.D. degrees in information and communication engineering from Naval Aeronautical Engineering Institute, Yantai, China, in 2002 and 2006, respectively.

She is currently a Full Professor and a Doctor Tutor with PLA Naval Aviation University. She is an author of four books, more than 100 articles, and more than 10 inventions. Her research interests include evidence reasoning, signal processing, and target recognition.

Miss Guan was the recipient of the Program for New Century Excellent Talents by the Minister of Education in 2011 and the Taishan Scholar in 2017. She is an Active Journal Reviewer, including the *Chinese Journal of Aeronautics*, the *Chinese Journal of Electronics, Science China*, and so on.