Remote Sensing Cross-Modal Retrieval by Deep Image-Voice Hashing

Yichao Zhang, Xiangtao Zheng, *Senior Member, IEEE*, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*—Remote sensing image retrieval aims at searching remote sensing images of interest among immense volumes of remote sensing data, which is an enormous challenge. Direct use of voice for human–computer interaction is more convenient and intelligent. In this article, a *deep image-voice hashing* (DIVH) method is proposed for remote sensing image-voice retrieval. First, the whole framework is composed of the image and the voice feature learning subnetwork. Then, the hash code learning procedure will be leveraged in remote sensing image-voice retrieval to further improve the retrieval efficiency and reduce memory footprint. Hash code learning maps the deep features of images and voices into a common Hamming space. Finally, image-voice pairwise loss is proposed, which considers the similarity preservation and balance of hash codes. The similarity preserving and the balance controlling term of the loss function improve the similarity preservation from original data space to the Hamming space and the discrimination of binary code, respectively. This unified cross-modal feature and hash code learning framework significantly reduce the semantic gap between the two modal data. Experiments demonstrate that the proposed DIVH method can achieve a superior retrieval performance than other state-of-the-art remote sensing image-voice retrieval methods.**

***Index Terms*—Convolutional neural network (CNN), cross-modal retrieval, deep hashing, hash code.**

I. Introduction

REMOTE sensing image retrieval [1] tries to search the desired images with a specific query, which is a huge challenge. Remote sensing image retrieval can provide the original data required or restrict the search space for a wide range of application scenarios [2], such as resource investigating [3], [4], surveying and mapping [5], [6], and disaster relief [7], [8].

The remote sensing image retrieval task is mainly divided into single-modal retrieval task and cross-modal retrieval task. Single-modal retrieval task tries to retrieve the desired remote sensing image with semantic correlation in the immense database [9], [10]. However, the visual content in the images is complicated, and which is challenging to express the content of a remote sensing image. Cross-modal retrieval uses one modal data to retrieve other modal data with semantic relevance in a massive database. The image-text and image-voice retrieval are widely used to give detailed retrieval requirements. Thereby, the retrieved images are more in line with the operator's vision. How to model high-level semantics and associations between different modal data is challenging in cross-modal retrieval.

With the rapid advancement of deep learning [11], [12], some deep cross-modal retrieval methods have been proposed to learn the correlation among different modal data [13], [14]. Abdullah et al. [15] unified *long short-term memory network* (LSTM) and pretrained *convolutional neural network* (CNN) to design a deep bidirectional triplet network for remote sensing image-text retrieval. Rahhal et al. [16] designed a novel unsupervised loss for unsupervised remote sensing image-text retrieval. The image-voice retrieval shows greater value in remote sensing applications than image-text retrieval. In practical application, the operator provides the voice description, which is convenient for the human–computer interaction [17]. However, it is difficult to retrieve effective information between images and voices since the feature representations of voices and images are inconsistent.

Recently, some methods are proposed to achieve image-voice retrieval [18], [19], [20], [21]. Guo et al. [18] proposed deep networks to obtain the cross-modal similarity between remote sensing images and voices. Torfi et al. [19] proposed a deep framework with coupled 3-D CNN to enhance the correlation of image and voice features. Zhang et al. [20] explored the canonical correlation of image and voice features in the feature extraction procedure. However, these feature extraction procedures use high-dimensional real-valued features, which causes higher storage and computing costs.

In contrast, the hashing methods map different modal data into compact binary codes, which are low memory and high efficiency [22], [23]. The hashing methods show their remarkable performance in large-scale multimedia information retrieval.

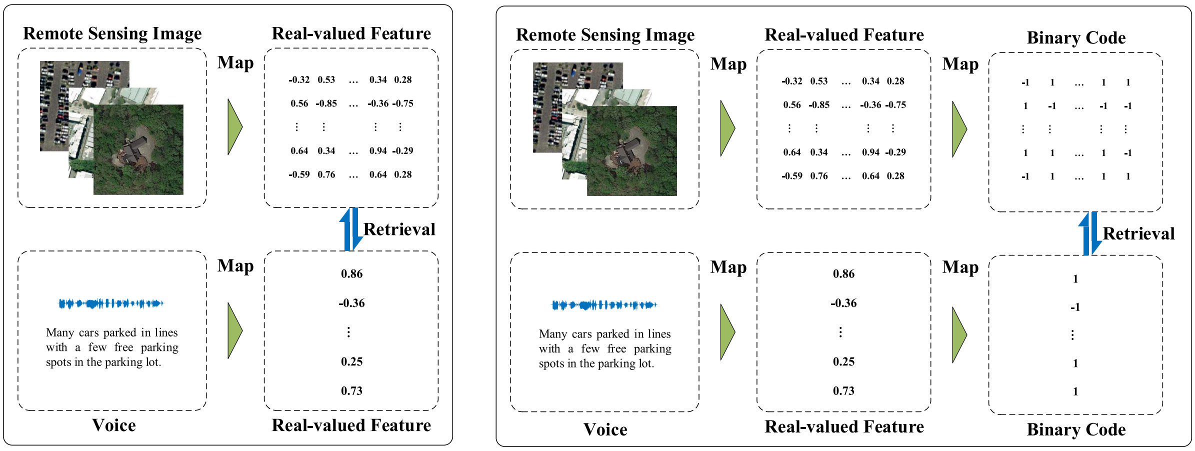
To improve retrieval efficiency, a *deep image-voice hashing* (DIVH) method is proposed to leverage the hash code learning procedure in remote sensing image-voice retrieval. The proposed method leverages two subnetworks to learn the remote sensing image and voice features. Fig. 1 shows the comparison between the proposed DIVH method and the real-valued retrieval method. Both remote sensing image and voice features are mapped into a common space. Furthermore, an image-voice pairwise loss is proposed to consider the similarity preserving

Manuscript received 24 July 2022; revised 12 September 2022; accepted 9 October 2022. Date of publication 25 October 2022; date of current version 8 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62271484, in part by the National Science Fund for Distinguished Young Scholars under Grant 61925112, and in part by the Innovation Capability Support Program of Shaanxi under Grant 2020TD-015. *(Corresponding author: Xiangtao Zheng.)*

Yichao Zhang is with the Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yczhang0819@gmail.com).

Xiangtao Zheng and Xiaoqiang Lu are with the Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xiangtaoz@gmail.com; luxq666666@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3216333


Remote Sensing Image
Real-valued Feature
Map
Retrieval
Voice
Many cars parked in lines with a few free parking spots in the parking lot.
Binary Code

(a) (b)

Fig. 1. Comparison between the retrieval real-valued method and the proposed DIVH method. (a) Real-valued retrieval method. (b) Proposed DIVH method. The real-valued method requires higher storage and computing costs.

and the balance of hash codes. The similarity preserving term of the loss function is designed to make paired similar images and voices as close as possible in the Hamming space. To improve the quality of the hash code, the balance controlling term is imposed on the value distribution of the binary code. Experiments demonstrate that the proposed DIVH method can achieve a superior retrieval performance than the state-of-the-art methods on the remote sensing image-voice retrieval task.

The remainder of this article is organized as follows. Section II briefly represents related work. Section III delves into the proposed DIVH method in depth. Section IV shows the experiments. Finally, Section V concludes this article.

II. Related Work

In this section, related work is discussed. The remote sensing image retrieval methods are discussed in Section II-A, while some remote sensing cross-modal retrieval methods are introduced in Section II-B.

A. Remote Sensing Image Retrieval

The remote sensing image retrieval tries to locate the desired remote sensing image from large amounts of remote sensing image data, which is divided into two categories: real-valued method and hashing method.

1) Real-Valued Method: Real-valued methods search for similar images by computing the similarity with real-valued features. Bosilj et al. [24] proposed a remote sensing image retrieval method, which used *vector of aggregate locally descriptor* as the feature representation. Li et al. [25] proposed a remote sensing image retrieval method with multiple feature representation and collaborative affinity metric fusion. Xiong et al. [26] proposed a deep framework with an attention module to reduce the background interference. Fan et al. [27] proposed a deep metric learning-based method to learn the image similarity and designed a distribution consistency loss. Although the real-value methods achieved acceptable results, they incurred high storage and similarity computation cost.

2) Hashing Method: Hashing methods received much attention because of their compact binary representation and efficient similarity computation. Lukač et al. [28] proposed a *kernelized supervised locality sensitive hashing* method for remote sensing image retrieval. Reato et al. [29] proposed a class-sensitive hashing method to express remote sensing images with a multihash code learning strategy. Li et al. [30] proposed a *deep hashing neural network* to introduce the pair-wise similarity constraint in an end-to-end strategy. Song et al. [31] proposed a *deep hashing convolutional neural network* to learn the compact hash codes with high-level feature.

B. Remote Sensing Cross-Modal Retrieval

Cross-modal retrieval can be divided into two categories: image-text retrieval method and image-voice retrieval method.

1) Image-Text Retrieval Method: Image-text retrieval method [15], [16] tries to search desired images (voices) according to the corresponding text description (images). Abdullah et al. [15] proposed a *deep bidirectional triplet network* (DBTN) to fuse the multiple text description with an average fusion strategy. Rahhal et al. [16] proposed a visual *big transfer* (BiT) Models and a *bidirectional LSTM* network to learn image and text features, respectively. Cheng et al. [32] proposed a *semantic alignment module* (SAM) to enhance the latent correlation between remote sensing images and text features. Yuan et al. [33] proposed an *asymmetric multimodal feature matching network* for image-text retrieval to achieve multiscale feature learning and target redundancy reduction. Although image-text retrieval methods can give more detailed descriptions of the desired images than image retrieval, text descriptions are still inconvenient and have too much subjectivity.

2) Image-Voice Retrieval Method: Image-voice retrieval methods [18], [34] search for desired images (voices) by using semantically related voice description (images). Guo et al. [18] integrated deep feature learning of different modal data into a unified framework for remote sensing image-voice retrieval.

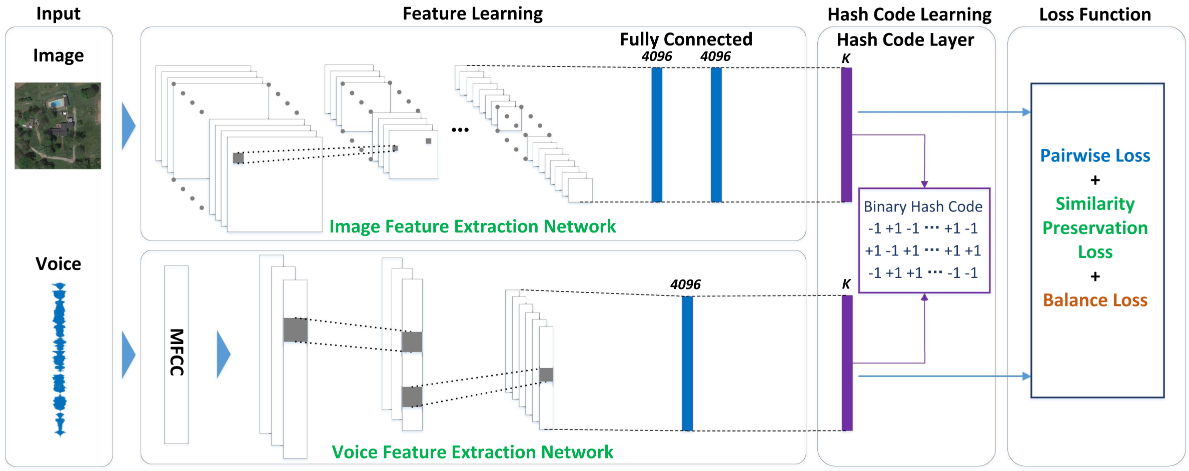


Fig. 2. Framework of the proposed DIVH method. An integrated deep framework is made up of image and voice branches. The binary hash code learning is used to increase the retrieval performance and minimize memory footprint. Hash code learning also can map the remote sensing images and voices into a common space.

Chen et al. [34] proposed a deep triplet-based network to learn the semantic similarity relationship between image and voice modal data. Chaudhuri et al. [35] proposed a *cross-modal information retrieval network* (CMIR-NET) to handle the multilabel unpaired image-voice retrieval. Yang et al. [36] proposed a cross-modal feature fusion retrieval method with intermodality adversarial learning and intramodality semantic discrimination to enhance the use of intramodal semantic information. Voice input can be more direct in human–computer interaction with remote sensing systems.

III. Proposed Method

The proposed *Deep Image-Voice Hashing* method is divided into three parts: image and voice feature extraction; pairwise similarity measurement; and hash code learning. The proposed DIVH method is shown in Fig. 2. The proposed DIVH method uses two subnetworks to learn the remote sensing image and voice features and map two modal features into common space. And image-voice pairwise loss is proposed to consider the similarity preserving and the balance of hash codes.

A. Image and Voice Feature Extraction

Let $\mathcal{I} = \{I_r\}_{r=1}^{N}$ represent a set of N images, and $\mathcal{V} = \{V_q\}_{q=1}^{M}$ is the corresponding set of M voice captions. Feature extraction includes two branches: remote sensing image branch and voice branch. The remote sensing image branch uses an image subnetwork to obtain convolutional features for raw remote sensing images. The voice branch first uses *Mel-frequency cepstral coefficients* (MFCC) [37], and then, refines the voice features for the subsequent hash code generation.

1) Image Feature Extraction: To express the image, the proposed DIVH method adopts the commonly used *Convolutional Neural Networks* (CNNs). The image features are gained from the last fully connected layer of the deep network, where the classification layer is obsoleted. The extracted feature of the raw remote sensing image I_r can be represented as follows:

$$H_i = f\left(I_r; \omega_i\right) \tag{1}$$

where H_i is the feature of remote sensing images, $f(\cdot)$ means the image feature extraction network, and ω_i represents deep network parameters for image feature extraction. A 4096-D feature vector is obtained by the image feature subnetwork.

2) Voice Feature Extraction: To express the voice, the proposed DIVH method adopts MFCC [37] and a feature refinement network. The raw voice data are mapped as a vector with the MFCC [37], which can be expressed as the following formula:

$$C_q = \mathcal{M}(V_q) \tag{2}$$

where V_q represents the qth voice, C_q represents the feature of the qth voice after MFCC procedure, and $\mathcal{M}(\cdot)$ represents the whole MFCC procedure. The input of the voice feature refinement network is the intermediate feature C_q after the MFCC procedure. The ultimate feature of voice modal data can be written as the following formula:

$$H_v = g\left(C_q; \omega_v\right) \tag{3}$$

where ω_v is the parameters of the CNN for voice modal data, H_v denotes the voice feature, and $g(\cdot)$ means the voice feature refinement network.

B. Pairwise Similarity Measurement

Let $\mathcal{B}^{(i)} = \{b_r\} \in \{+1, -1\}^{K \times N}$ denote a set of N binary hash codes of images with K bits, and $\mathcal{B}^{(v)} = \{b_q\} \in \{+1, -1\}^{K \times M}$ denotes a set of M hash codes of voices with K bits. Two mappings $\mathcal{F} : \mathcal{I} \rightarrow \mathcal{B}^{(i)}$ and $\mathcal{G} : \mathcal{V} \rightarrow \mathcal{B}^{(v)}$ are learned by the hashing method. These two mappings correspond to image modality and voice modality, respectively. Θ_{rq} is used to denote the inner product between the hash code of the image $\mathbf{b}_r^{(i)}$ and the hash code of the voice $\mathbf{b}_q^{(v)}$:

$$\Theta_{rq} = \mathbf{b}_r^{(i)T} \mathbf{b}_q^{(v)}. \tag{4}$$

The following formula denotes the Hamming distance of the image hash code $\mathbf{b}_r^{(i)}$ and voice hash code $\mathbf{b}_q^{(v)}$:

$$\mathrm{dist}_H(\mathbf{b}_r^{(i)}, \mathbf{b}_q^{(v)}) = \frac{1}{2}(K - \Theta_{rq}). \tag{5}$$

The similarities of image and voice are expressed as: $\mathcal{S} = \{s_{rq}\} \in \{0,1\}^{n\times n}$, where s_{rq} equals to 1 if the hash code of the image $\mathbf{b}_r^{(i)}$ and hash code of voice $\mathbf{b}_q^{(v)}$ are corresponding in semantics. If not, s_{rq} equals to 0. By using these pairwise similarity measurements constructed, the maximum a posterior estimation will be expressed as the following formula:

$$\log p(\mathcal{B}\,|\mathcal{S}) = \sum_{r,q=1}^{n} \log p(s_{rq} \Big|(\mathbf{b}_r^{(i)}, \mathbf{b}_q^{(v)})p(\mathbf{b}_r^{(i)})p(\mathbf{b}_q^{(v)})) \tag{6}$$

where $p(\mathcal{B}|\mathcal{S})$ denotes the likelihood function. $p(s_{rq}|\mathbf{b}_r^{(i)}, \mathbf{b}_q^{(v)})$ is the conditional probability of the pairwise similarity measurement s_{iv} given the corresponding hash codes. The cross-modal pairwise similarity measurement between images and voices can be expressed as follows:

$$p(s_{rq}\Big|\mathbf{b}_r^{(i)}, \mathbf{b}_q^{(v)}) = \begin{cases} \sigma\left(\frac{1}{2}\Theta_{rq}\right), & s_{iv} = 1 \\ 1 - \sigma\left(\frac{1}{2}\Theta_{rq}\right), & \text{otherwise} \end{cases} \tag{7}$$

where $\sigma(x)$ denotes the sigmoid function. And $\sigma(x) = \frac{1}{1+e^{-x}}$. From (7), it can be reasonably concluded that the larger inner product between the hash code of image $\mathbf{b}_r^{(i)}$ and hash code of voice $\mathbf{b}_q^{(v)}$ is, the larger $p(s_{rq}|\mathbf{b}_r^{(i)}, \mathbf{b}_q^{(v)})$ will be, which represents that $\mathbf{b}_r^{(i)}$ and $\mathbf{b}_q^{(v)}$ would be regarded as corresponding ones.

C. Hash Code Learning

According to the maximum a posterior estimation in (6), the following formula can be leveraged to learn the hash code:

$$\mathcal{L}_1 = -\log p(\mathcal{S}\,|\mathcal{B}) = -\sum_{r,q=1}^{n}\left(s_{rq}\Theta_{rq} - \log\left(1 + e^{\Theta_{rq}}\right)\right) \tag{8}$$

where $\mathcal{L}_1$ denotes the cross-entropy loss. The $\mathcal{L}_1$ is used to make the Hamming distance of two similar samples as small as possible, and that of two dissimilar samples as large as possible simultaneously.

The $\mathbf{H}^{(i)}$ and $\mathbf{H}^{(v)}$ are extracted from the image subnetwork and the voice subnetwork, respectively. The binary hash codes $\mathbf{B}^{(i)}$ and $\mathbf{B}^{(v)}$ can be expected to preserve the cross-modal similarity.

$$\mathcal{L}_2 = \|\mathbf{B}^{(i)} - \mathbf{H}^{(i)}\|_I^2 + \|\mathbf{B}^{(v)} - \mathbf{H}^{(v)}\|_I^2 \tag{9}$$

where $\mathbf{B}^{(i)} = \mathrm{sign}(\mathbf{H}^{(i)})$, $\mathbf{B}^{(v)} = \mathrm{sign}(\mathbf{H}^{(v)})$, and $\mathbf{B} \in \{-1,+1\}^{N\times K}$.

A good balance of +1 and -1 is crucial for the generation of effective hash codes [38], [39], [40]. Let the appearance of +1 and that of -1 on hash codes to be almost the same, $\mathcal{L}_3$ can be designed as the following formula:

$$\mathcal{L}_3 = \sum_{z=1}^{K}\left|\mathrm{mean}(b_r^{(z)}) + \mathrm{mean}(b_q^{(z)})\right| \tag{10}$$

The overall loss function not only considers the preservation of the similarity between the image and the voice, but also considers the balance of the hash code. Finally, the aforementioned three parts are incorporated into the overall loss function:

$$\begin{aligned}\min_{\mathbf{B},\omega_i,\omega_v} \mathcal{L} = &-\sum_{r,q=1}^{n}\left(s_{rq}\Theta_{rq} - \log\left(1 + e^{\Theta_{rq}}\right)\right) \\ &+ \gamma\left(\|\mathbf{B}^{(i)} - \mathbf{H}^{(i)}\|_I^2 + \|\mathbf{B}^{(v)} - \mathbf{H}^{(v)}\|_I^2\right) \\ &+ \eta\sum_{z=1}^{K}\left|\mathrm{mean}(b_r^{(z)}) + \mathrm{mean}(b_q^{(z)})\right|\end{aligned} \tag{11}$$

where γ and η are the weights of $\mathcal{L}_2$ and $\mathcal{L}_3$, respectively. Both feature extraction and hash-code learning are integrated into a unified deep framework. The proposed DIVH method learns binary hash codes $\mathbf{B}^{(i)}$ and $\mathbf{B}^{(v)}$ and parameters of two feature extraction subnetworks ω_i and ω_v by minimizing the common loss.

IV. Experiments

This section briefly discusses the experimental settings (see Section IV-A), datasets (see Section IV-B), evaluation protocols (see Section IV-C), and hyperparameter analysis (see Section IV-D). Section IV-E shows the comparison results with state-of-the-art methods on three benchmark datasets. The ablation experiments and the complexity of the proposed DIVH method are analyzed in Section IV-F.

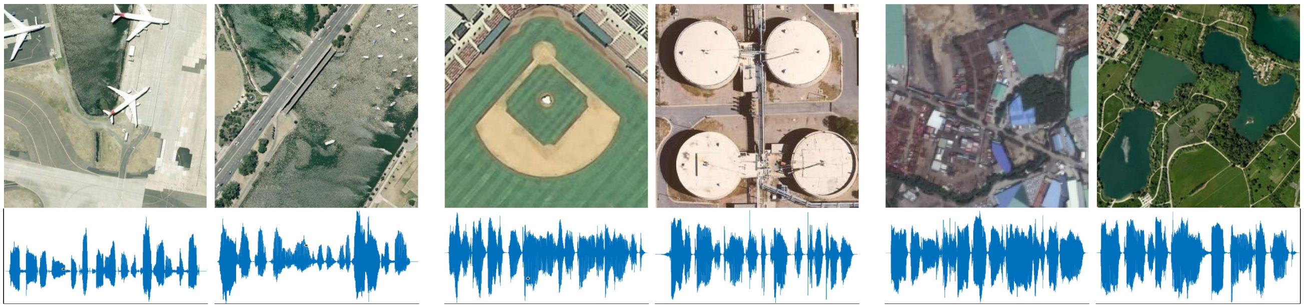
A. Experimental Settings

The proposed DIVH method is conducted on a server with an NVIDIA Quadro RTX 6000 GPU and 24 G of RAM. The loss function introduced previously is optimized utilizing *stochastic gradient descent* in the implementation process. A seven-layer CNN-F is leveraged as the image subnetwork for the remote sensing image branch. This CNN-F has been pretrained using the ImageNet Dataset. The hash code length is set to 64. Furthermore, γ and η are both set to 1. The selection of hyperparameters will be discussed in Section IV-D.

B. Datasets

To verify the effectiveness of the proposed DIVH method, three image-voice benchmark datasets are introduced: Sydney Image-Voice Dataset, UCM Image-Voice Dataset, and RSICD Image-Voice Dataset. Table I shows some details of three image-voice benchmark datasets. Some image-voice pair examples in three benchmark datasets are shown in Fig. 3.

1) *Sydney Image-Voice Dataset [18]* is a cross-modal image-voice dataset that was created by expanding the original image dataset. The original dataset is a huge remote sensing image dataset of Sydney obtained from Google Earth. The size of the whole image is 18 000 × 14 000, with a resolution of 0.5 m/pixel. The overall image was cut into 1008 nonoverlapping subimages with a size of 500 × 500 containing 613 remote sensing images. The 613 images belong to seven categories, called industrial area,



Three same airplanes are parked on the airport with dark waters beside.

Many white boats on the wide river with a highway beside.

(a)

This is a baseball diamond with stands surround it.

Four storage tanks are connected to each other.

(b)

Many buildings and some green trees are in an industrial area.

Many green trees are around four ponds in a park.

(c)

Fig. 3. Some image-voice pair examples in three benchmark datasets. (a) UCM Image-Voice Dataset. (b) RSICD Image-Voice Dataset. (c) Sydney Image-Voice Dataset. Each remote sensing image corresponds to a voice description.

TABLE I
SOME DETAILS OF THREE IMAGE-VOICE BENCHMARK DATASETS LEVERAGED IN EXPERIMENTS

Dataset	UCM image-voice	Sydeney image-voice	RSICD image-voice
Categories	21	7	30
Images	2100	613	10 921
Voice caption	2100	613	10 921
Image-voice positive pairs	2 100	613	10 921
Image-voice negative pairs	2 100	613	10 921
Training	80%	80%	80%
Test	20%	20%	20%

ocean, lawn, river, airport, residential area, and railway. The dataset contains 613 image-voice sample pairs, and each image corresponds to a voice description. Similar to the previous literature [18], we constructed 613 positive sample pairs and 613 negative ones. Eighty percent of the image-voice sample pairs are chosen for training, while the remainder 20% is used for testing.

2) *UCM Image-Voice Dataset [18]* is a cross-modal image-voice dataset expanded from a single-modal remote sensing image dataset. It includes 2100 remote sensing images. There are 21 scene classes of remote sensing images (agricultural, airplane, beach, buildings, chaparral, dense residential, forest, etc.) in the original dataset, and each class contains 100 remote sensing images. These images in the UCM dataset are 256 × 256 pixels in size. Each remote sensing image is given a voice description to compile this dataset.

3) *RSICD Image-Voice Dataset [18]* is an extension of the remote sensing caption dataset [41]. The original dataset contains 10 921 images belonging to 30 categories. The expanded dataset has 10 921 positive and negative image-voice sample pairs, and each remote sensing image corresponds to a voice. Like the experimental settings of the other two datasets, eighty percent of the image-voice sample pairs are chosen for training, with the remainder used for testing.

C. Evaluation Protocols

Two representative metrics are leveraged as evaluation protocols for the image retrieval performance.

1) *Mean average precision (mAP)* is the mean of average precision (AP) for each query in query set. The following formula is used to calculate the AP:

$$\mathrm{AP} = \frac{1}{R}\sum_{i}^{N}\frac{R_i}{i} \times \mathrm{rel}_i \tag{12}$$

where N denotes the number of the samples. The number of relevant images in the dataset is denoted by R. The number of relevant images in the top i results is denoted by Ri. If the sample ranked ith place is relevant, rel_i will be 1 and otherwise 0.

2) P@m is the precision of top m nearest images from the same class or with semantic consistency in a certain Hamming space.

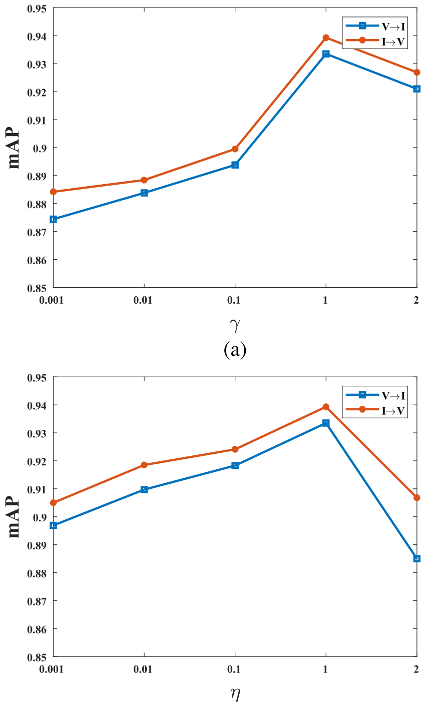
$$\mathrm{P@}m = \frac{\sum_{i=1}^{m}\mathrm{rel}_i}{m}. \tag{13}$$

The greater the value of these indicators, the better the performance is. In this article, the aforementioned two evaluation protocols are used on three image-voice retrieval datasets.

D. Hyperparameter Analysis

Two hyperparameters in the loss function (11) utilized for hash code learning. The hyperparameter γ determines the contribution of cross-modal similarity $\mathcal{L}_2$ in (9). The hyperparameter η determines the contribution of the hash code balance $\mathcal{L}_3$ in (10). This experiment modifies one of the two hyperparameters while leaving the other alone.

The value of η is fixed to 1, and the value of γ is set to 0.001, 0.01, 0.1, 1, and 2, respectively. The image-voice and voice-image retrieval experiments were carried out using the settings described previously. As may be observed from Fig. 4(a), whether it is image-voice or voice-image retrieval, when the value of γ is set to 1, the algorithm performs best in



(a)

(b)

Fig. 4. Hyperparameter experiments results. (a) mAP corresponding to the different γ values on the Sydney Image-Voice Dataset (the η value is fixed to 1). (b) mAP corresponding to the different η values on the Sydney Image-Voice Dataset (the γ value is fixed to 1).

both image-voice and voice-image retrieval. In the experiment of image-voice retrieval, the value of γ is set to 1, which is nearly 6% higher than when the γ is 0.001. In the voice-image retrieval experiment, the value of γ is set to 1, which is nearly 6% higher than when the γ is 0.001. It is worth noting that when the γ value is set to 2, the performance is reduced. This is mainly because the γ value is too large, affecting the contribution of other loss items.

The value of γ is fixed to 1, and the value of η is set to 0.001, 0.01, 0.1, 1, and 2, respectively. The image-voice and voice-image retrieval experiments were carried out according to the aforementioned settings, and the results are shown in Fig. 4(b). In the experiment of image-voice retrieval, the value of η is set to 1, which is about 3.5% higher than when η is 0.001. In the voice-image retrieval experiment, the value of η set to 1 is about 3.2% higher than when η is 0.001. When the value of η is set to 2, the performance is greatly reduced. When the η and γ values are set to extremely small, γ has a greater impact on performance. The aforementioned two phenomena show that the contribution of $\mathcal{L}_2$ to the entire loss function is higher.

E. Comparison Results on Benchmark Datasets

In this subsection, nine current state-of-the-art cross-modal retrieval methods are used, including: SIFT+M [42], CMFH [43], DBLP [44], CNN [45], DVAN [18], CMIR-NET [35], DIVR [46], SePHklr [47], and DTBH [34]. The experiment was implemented on three benchmark datasets, which are Sydney Image-Voice Dataset, UCM Image-Voice Dataset, and RSICD Image-Voice Dataset.

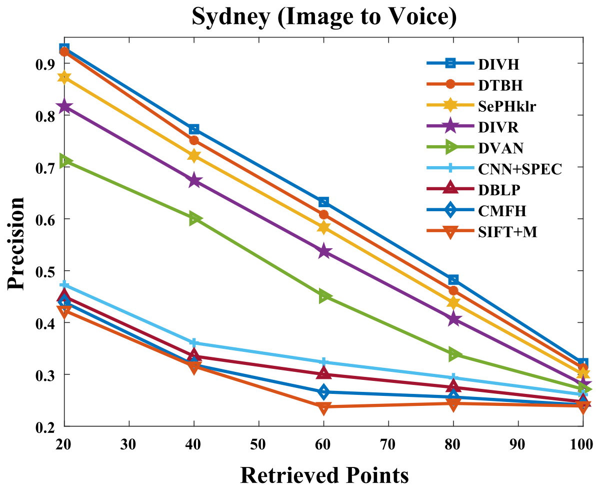


Fig. 5. Precision results with the different numbers of retrieved points (image to voice) on the Sydney Image-Voice Dataset.

1) Retrieval Results on Sydney Image-Voice Dataset: Table II reports the image-to-voice retrieval results on the Sydney Image-Voice Dataset, including mAP and P@m (P@1, P@5, and P@10). The significance of bold entities is best experimental result in each metric. It can be noted that the performance of the proposed DIVH method is better than other methods in the mAP, especially more than 60% higher than the SIFT+M method, demonstrating the superiority that the DIVH method leverages the end-to-end framework to learn both the features and hash codes. The feature learning procedures of image and voice are combined into a single framework capable of learning more efficient hash codes. Furthermore, a robust semantic association is established for the hash codes of voices and images. This significantly enhances the image-voice retrieval model's performance. Although the proposed DIVH method is slightly lower than the DTBH model on the P@m indicator, there may be the following two reasons:

1) the unpredictability of data division;
2) the DTBH method uses the triplet-wise training that is more difficult to train.

Although the DTBH method improves the performance of the model on the Sydney Image-Voice Dataset, it is not easy to train. On the contrary, the proposed DIVH method does not need to consider this problem, can be effectively trained, and does not require too much fine-tuning and additional construction of triplet datasets. Fig. 5 shows the precision curves with the different numbers of samples retrieved on the Sydney Image-Voice Dataset. It can be seen that the proposed DIVH method is superior to other methods in all returned neighbors. Fig. 5 further illustrates the effectiveness of the proposed DIVH method. Fig. 8 shows some visual results of image-to-voice retrieval. The first row is the result on the Sydney Image-Voice Dataset. There are runways and lawns in the image, indicating that the scene is a runway scene. The retrieved voice contains this valid information. For example, different words with the same concept were found, including "runway" and "lawn." This

TABLE II
IMAGE-VOICE RETRIEVAL RESULTS (MEAN AVERAGE PRECISION AND P@m) ON THE SYDNEY IMAGE-VOICE DATASET

Methods	Image to voice (Sydney)				Voice to image (Sydney)			
	mAP	P@1	P@5	P@10	mAP	P@1	P@5	P@10
SIFT+M [42]	31.67	11.21	35.00	37.59	26.50	34.48	24.48	23.28
CMFH [43]	40.62	44.27	41.66	39.35	31.58	29.83	28.16	27.32
DBLP [44]	44.38	56.51	52.65	49.68	34.87	21.63	26.78	30.94
SPEC [45]	46.67	58.62	55.00	51.64	35.72	17.24	27.76	31.21
DVAN [18]	71.77	75.86	73.62	72.93	63.88	67.24	63.34	67.07
CMIR-NET [35]	78.44	84.68	82.54	81.04	71.28	76.69	74.52	71.60
DIVR [46]	81.35	88.26	86.35	84.47	75.97	80.44	78.05	76.27
SePHklr [47]	88.18	86.71	87.31	86.60	85.93	85.96	83.40	84.53
DTBH [34]	92.45	**97.41**	**95.63**	**93.78**	87.49	92.18	90.36	88.82
DIVH	**93.93**	95.23	94.65	93.07	**93.35**	**95.20**	**93.84**	**93.55**

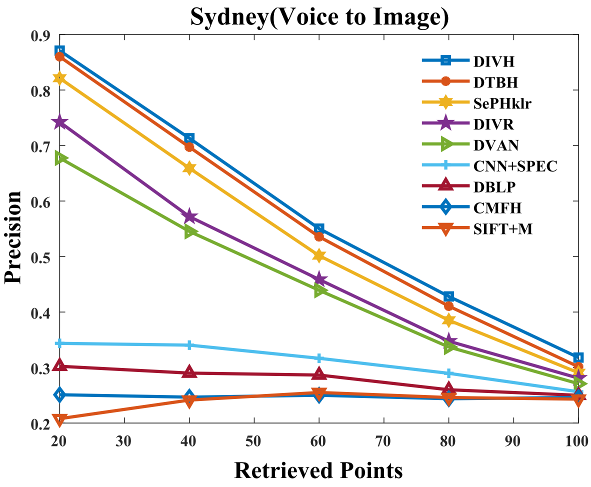
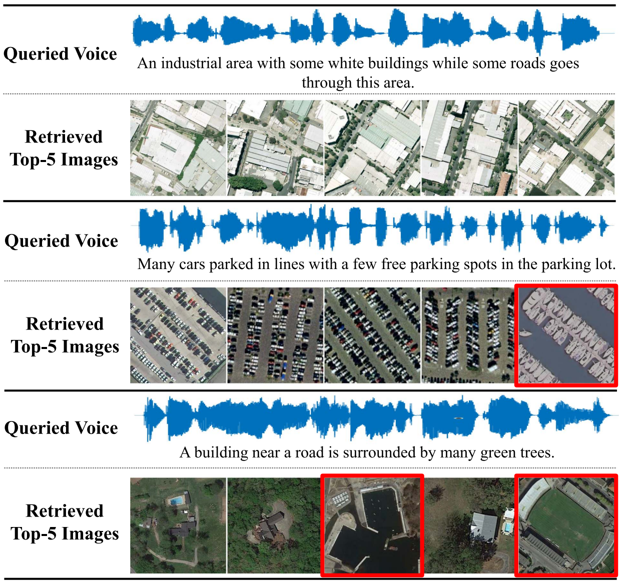


Fig. 6. Precision results with the different numbers of retrieved points (voice to image) on the Sydney Image-Voice Dataset.



An industrial area with some white buildings while some roads goes through this area.

Many cars parked in lines with a few free parking spots in the parking lot.

A building near a road is surrounded by many green trees.

Fig. 7. Some voice-image retrieval results of the DIVH method on three image-voice datasets. The first row corresponds to Sydney dataset. The second row corresponds to UCM dataset. And the third row corresponds to RSICD Image-Vocie Datasets. The first row correspond to Sydney dataset. The red boxes indicate incorrect retrieval results.

shows that the proposed DIVH method can effectively identify the effective information in the images and retrieve the relevant voices.

Table II reports the results of the proposed DIVH method and state-of-the-arts in the voice-to-image retrieval on the Sydney Image-Voice dataset. The significance of bold entities is best experimental result in each metric. As may be observed, the proposed DIVH method is higher than other methods in all indicators (mAP, P@1, P@5, and P@10). This is because of the unified framework of DIVH that integrates both two modal feature learning and hash code learning. Through end-to-end training, efficient hash code can be learned, and a semantic association is established simultaneously between the two modalities, thereby improving the performance of cross-modal retrieval. It is worth noting that even though the state-of-the-art DTBH method uses a more difficult-to-train triplet-wise way to improve the performance of the model, the proposed method still has a certain improvement in various indicators. This shows the simplicity and effectiveness of the proposed DIVH method. Fig. 6 shows the precision results with different retrieved points on the Sydney Image-Voice Dataset. As may be observed, the proposed method is superior to other methods in all returned neighbors. This also shows the effectiveness of our method. Fig. 7 shows some visual results of voice-to-image. The first row is the result on the Sydney Image-Voice Dataset. These voices contain words such as “white building,” “roads,” “industrial area,” etc., indicating that it is describing an industrial scene. It can be seen that the retrieved images also include concepts such as “white buildings” and “roads.” This shows that the proposed DIVH method can identify the effective information of the voices and retrieve the relevant remote sensing scene images.

2) *Retrieval Results on UCM Image-Voice Dataset:* Table III shows the image-to-voice retrieval results of our method and some comparison methods on the UCM image-voice dataset, including mAP and P@m (P@1, P@5, and P@10). The significance of bold entities is best experimental result in each metric. As may be observed, the performance of the proposed DIVH method is better than other methods in all indicators, especially almost 70% higher than the SIFT+M method and 8.84% higher

TABLE III
IMAGE-VOICE RETRIEVAL RESULTS (MEAN AVERAGE PRECISION AND P@m) ON THE UCM IMAGE-VOICE DATASET

Methods	Image to voice (UCM)				Voice to image (UCM)			
	mAP	P@1	P@5	P@10	mAP	P@1	P@5	P@10
SIFT+M [42]	8.55	4.56	4.65	4.56	6.66	3.58	4.41	4.68
CMFH [43]	16.17	18.33	18.00	18.10	17.93	19.76	19.57	19.40
DBLP [44]	25.48	24.18	23.87	23.24	19.33	17.12	17.62	16.31
SPEC [45]	26.25	29.50	25.52	23.65	21.79	19.42	19.86	19.23
DVAN [18]	36.79	32.37	33.29	33.74	32.28	32.37	33.91	34.34
CMIR-NET [35]	45.82	52.92	49.74	43.38	40.37	46.74	43.75	39.62
DIVR [46]	50.94	59.34	54.17	50.12	45.34	52.23	48.76	44.98
SePHklr [47]	61.47	68.26	66.15	64.81	57.50	66.84	64.13	62.03
DTBH [34]	64.24	73.10	69.69	65.63	60.13	70.26	66.63	61.73
DIVH	**67.87**	**73.66**	**74.99**	**74.47**	**65.24**	**84.10**	**81.20**	**79.77**

than the DTBH in P@10. The feature learning processes of image and voice modal data are unified into a framework that can learn more efficient hash codes. And a strong semantic association is established for the hash codes of voices and images. This effectively improves the performance of the proposed DIVH method. In comparison to the DTBH, the proposed DIVH method does not need to consider this problem, can be simply and effectively trained, and does not require too much fine-tuning and additional construction of triplet-wise datasets. Fig. 9 shows the precision results with different numbers of retrieved points on the UCM Image-Voice Dataset. The proposed method outperforms the others in all returned neighbors. Fig. 9 further illustrates the effectiveness of the proposed DIVH method. The second row of Fig. 8 is the result on UCM Image-Voice Dataset. There are buildings in the image, indicating that the scene is a building scene. It is worth noting that although the third voice retrieved is wrong, it contains the word "gray," indicating that the model can learn useful semantic information. This shows that the proposed DIVH method can identify the effective information in the images and retrieve the relevant voices.

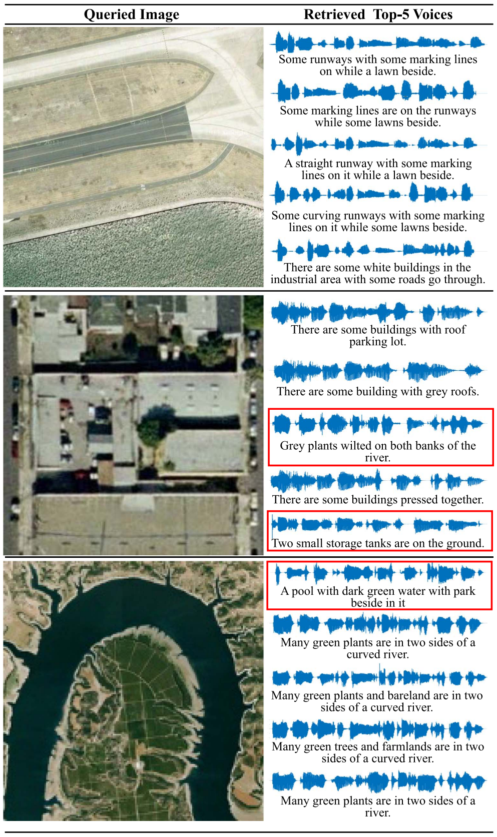

Queried Image
Retrieved Top-5 Voices
Some runways with some marking lines on while a lawn beside.
Some marking lines are on the runways while some lawns beside.
A straight runway with some marking lines on it while a lawn beside.
Some curving runways with some marking lines on it while some lawns beside.
There are some white buildings in the industrial area with some roads go through.
There are some buildings with roof parking lot.
There are some building with grey roofs.
Grey plants wilted on both banks of the river.
There are some buildings pressed together.
Two small storage tanks are on the ground.
A pool with dark green water with park beside in it
Many green plants are in two sides of a curved river.
Many green plants and bareland are in two sides of a curved river.
Many green trees and farmlands are in two sides of a curved river.
Many green plants are in two sides of a river.

Fig. 8. Some image-voice retrieval results of the DIVH method on three image-voice datasets. The first row corresponds to Sydney dataset. The second row corresponds to UCM dataset. And the third row corresponds to RSICD Image-Voice Datasets. The first row correspond to Sydney dataset. The red boxes indicate incorrect retrieval results.

Table III reports the results of the proposed DIVH method and state-of-the-arts in the voice-to-image retrieval on the UCM Image-Voice Dataset. The significance of bold entities is best experimental result in each metric. As may be observed, the proposed DIVH method is higher than other methods in all indicators (mAP, P@1, P@5, and P@10). This is because of the unified framework of DIVH that integrates two modal feature learning and hash code learning. Through end-to-end training, efficient hash code can be learned, and a semantic association is established simultaneously between the two modal data, thereby improving the performance of cross-modal retrieval. It is worth noting that even though the current state-of-the-art DTBH method uses a more difficult-to-train triplet-wise way to improve the performance of the model, the proposed method still has a certain improvement in all indicators. This shows the simplicity and effectiveness of the proposed method. Fig. 10 shows the precision results with different retrieved points on the UCM Image-Voice Dataset. As may be observed, the proposed method is superior to other methods in all returned neighbors. The second row of Fig. 7 is the result on UCM Image-Voice Dataset. The voices contain words such as "many," "cars," "park," etc., indicating that it is describing a parking lot scene. It can be seen that the retrieved images also include concepts such as "cars," "parking," and "many." The fifth image is a wrong retrieval result, but it contains "park" and "many." This shows that the proposed method can identify the effective information of the voices and retrieve the relevant scene images.

3) Retrieval Results on RSICD Image-Voice Dataset: The image-to-voice retrieval results of our method and other methods on the RSICD Image-Voice Dataset are shown in Table IV,

TABLE IV
IMAGE-VOICE RETRIEVAL RESULTS (MEAN AVERAGE PRECISION AND P@m) ON THE RSICD IMAGE-VOICE DATASET

Methods	Image to voice (RSICD)				Voice to image (RSICD)			
	mAP	P@1	P@5	P@10	mAP	P@1	P@5	P@10
SIFT+M [42]	5.04	6.22	5.34	4.50	4.85	3.66	3.60	3.54
CMFH [43]	9.72	10.57	10.16	9.81	7.83	6.92	6.16	5.94
DBLP [44]	12.70	15.32	15.21	14.22	8.14	6.21	6.08	6.76
SPEC [45]	13.24	16.82	16.62	15.69	9.96	7.13	7.00	7.44
DVAN [18]	16.29	22.49	22.56	21.70	15.71	16.18	15.10	14.76
CMIR-NET [35]	17.78	24.11	23.52	22.54	17.25	17.94	16.58	15.36
DIVR [48]	19.62	25.43	24.84	24.20	18.58	19.76	18.31	17.59
SePHklr [47]	21.52	24.36	23.81	23.14	20.91	22.14	19.48	17.52
DTBH [34]	23.46	27.58	26.84	25.49	22.72	23.30	22.48	21.17
DIVH	**27.42**	**30.81**	**28.36**	**27.04**	**29.18**	**32.42**	**31.83**	**30.28**

TABLE V
ABLATION EXPERIMENTS ABOUT SIMILARITY PRESERVING TERM AND BALANCE CONTROLLING TERM

Datasets	Similarity preserving term	Balance controlling term	Results (Image to voice)	Results (Voice to image)
Sydney Image-Voice	×	×	0.8533	0.8452
	✓	×	0.9014	0.8915
	×	✓	0.8821	0.8703
	✓	✓	**0.9393**	**0.9335**
UCM Image-Voice	×	×	0.6128	0.6015
	✓	×	0.6453	0.6284
	×	✓	0.6401	0.6270
	✓	✓	**0.6787**	**0.6524**
RSICD Image-Voice	×	×	0.1953	0.1872
	✓	×	0.2326	0.2261
	×	✓	0.2247	0.2182
	✓	✓	**0.2742**	**0.2918**

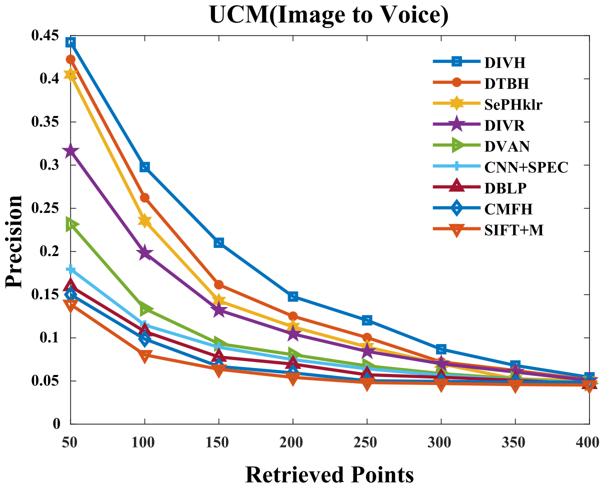


Fig. 9. Precision results with the different numbers of retrieved points (image to voice) on the UCM Image-Voice Dataset.

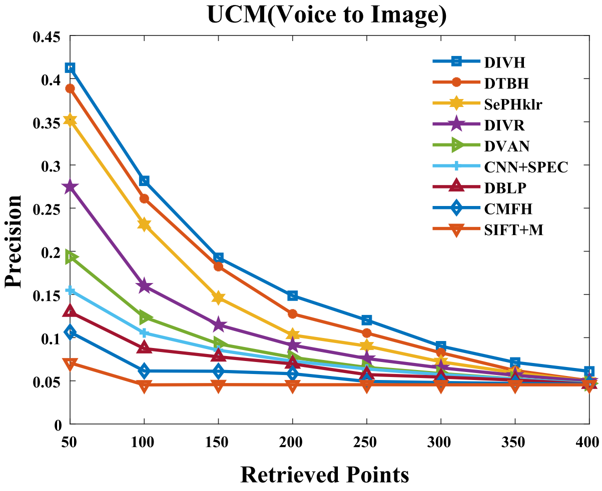


Fig. 10. Precision results with the different numbers of retrieved points (voice to image) on the UCM Image-Voice Dataset.

TABLE VI
RUNNING EFFICIENCY AND COMPLEXITY OF THE PROPOSED DIVH METHOD

	Inference time complexity	Running time
DIVH	$O(N_1 \cdot N_2 \cdot K)$	0.28 ms

including mAP and P@m (P@1, P@5, and P@10). The significance of bold entities is best experimental result in each metric. As may be observed, the performance of our method is better than other methods in all indicators, especially more than 20% higher than the SIFT+M method and 3.96% higher

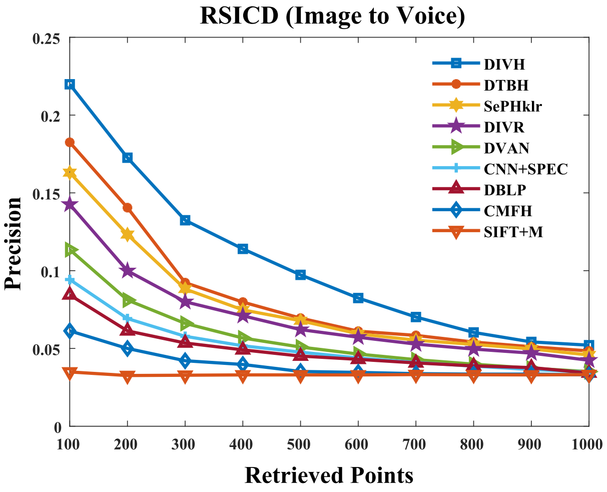


Fig. 11. Precision results with the different numbers of retrieved points (image to voice) on the RSICD Image-Voice Dataset.

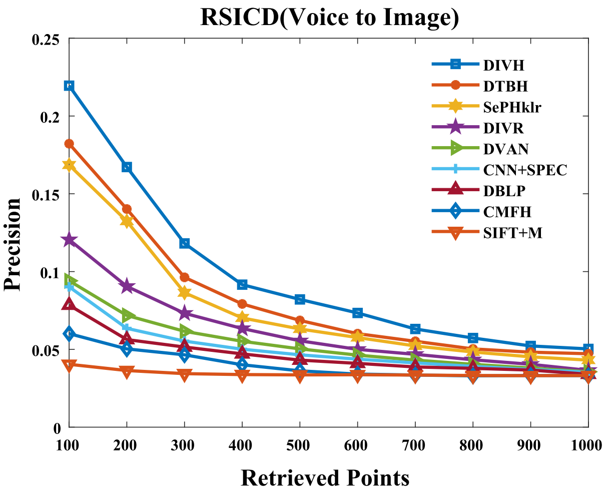


Fig. 12. Precision results with the different numbers of retrieved points (voice to image) on the RSICD Image-Voice Dataset.

than the DTBH in P@1, which shows the superiority that DIVH method leverages the end-to-end framework to extract the image and voice features and produce hash codes. The feature learning processes of image modality and voice modality are unified into a framework that can learn more efficient hash codes. And a strong semantic association is established for the hash codes of voices and images. This considerably improves the performance of the image-voice retrieval model. In comparison to the DTBH, our method does not need to consider this problem, can be simply and effectively trained, and does not require too much fine-tuning and additional construction of triplet-wise datasets. Fig. 11 shows the precision results with the different numbers of retrieved points on the RSICD Image-Voice Dataset. As may be observed, the proposed method is superior to other methods in all returned neighbors. Fig. 11 further illustrates the effectiveness of our method. The third row of Fig. 8 is the result on RSICD Image-Voice Dataset. There are buildings in the image, indicating that the scene is a building scene. It is worth noting that although the first voice retrieved is wrong, it contains the word "water" and "green," indicating that the algorithm can learn useful semantic information. This shows that the proposed method can identify the effective information in the images and retrieve the relevant voices.

The results of the proposed DIVH method and comparison methods in voice-to-image retrieval on the RSICD Image-Voice dataset are reported in Table IV. The significance of bold entities is best experimental result in each metric. The proposed DIVH method outperforms all other methods in all evaluation protocols (mAP, P@1, P@5, and P@10). This is because of the unified framework of DIVH that integrates two modal feature learning and hash code learning. End-to-end training allows efficient hash code to be learned while concurrently establishing a semantic association between the two modal data, boosting the cross-modal retrieval performance. It is worth noting that even though the state-of-the-art DTBH method uses a more difficult-to-train triplet-wise way to improve the performance of the model, the proposed method still has a certain improvement in all indicators. This shows the simplicity and effectiveness of our method. Fig. 12 shows the precision results with different samples retrieved on the RSICD Image-Voice Dataset. It can be seen that the proposed method is superior to other methods in all returned neighbors. This also shows the effectiveness of our method. The second row of Fig. 7 is the result on RSICD Image-Voice Dataset. The voices contain words such as "many," "car," "park," etc., indicating that it is describing a parking lot scene. As may be observed, the retrieved images also include concepts such as "building," "road," "green," and "tree." The third image is a wrong retrieval result, but it contains "road." The fifth image is also a wrong retrieval result, but it contains "green." This shows that the proposed method can identify the effective information of the voices and retrieve the relevant scene images.

F. Further Analysis

1) Ablation Experiments: The ablation experiments are implemented to further verify the effectiveness of the similarity preserving and the balance controlling term of the loss function. Table V shows the ablation experiment results about similarity preserving term and balance controlling term. The significance of bold entities is best experimental result in each metric. First, we analyze the situation with only balance controlling term. For the image-to-voice retrieval task, the similarity preserving term improves performance by 5.72%, 3.86%, and 4.95% on Sydney Image-Voice, UCM Image-Voice, and RSICD Image-Voice dataset, respectively. For the voice-to-image retrieval task, the similarity preserving term improves performance by 6.32%, 2.54%, and 7.36% on Sydney Image-Voice, UCM Image-Voice, and RSICD Image-Voice dataset, respectively. Then, we analyze the situation with only similarity preserving term. For the image-to-voice retrieval task, the balance controlling term improves performance by 3.79%, 3.34%, and 3.16% on Sydney Image-Voice, UCM Image-Voice, and RSICD Image-Voice dataset, respectively. For the voice-to-image retrieval task, the balance controlling term improves performance by 4.25%, 2.4%, and 6.57% on Sydney Image-Voice, UCM Image-Voice, and RSICD

Image-Voice dataset, respectively. The aforementioned results demonstrate the effectiveness of the similarity preserving and the balance controlling term in the loss function.

2) Complexity Analysis: Theoretical inference time complexity is further analyzed. Suppose there are N_1 images and N_2 voices in total, and the length of hash code is K. The theoretical inference time complexity of the proposed framework is $O(N_1 \cdot N_2 \cdot K)$ approximately [49]. The proposed DIVH framework is implemented on a server with an NVIDIA Quadro RTX 6000 GPU and 24 G of RAM. The single query runtime is tested. The single query runtime indicates the time to compare a single voice query to all of images in the test set of the Sydney Image-Voice dataset. The single query runtime and complexity of the proposed DIVH method are reported in Table VI.

V. Conclusion

In this article, a deep remote sensing image-voice retrieval method is proposed, called DIVH. The image and voice feature learning are combined into a single framework capable of learning efficient hash codes. Furthermore, the image-voice pairwise loss is presented by considering hash code similarity preservation and hash code balance controlling. Experiments demonstrate that the proposed DIVH method obtains superior retrieval performance than other state-of-the-art cross-modal methods. The existing cross-modal remote sensing retrieval still focuses on one most important content in the small remote sensing image. In the future, the retrieval challenge for multiple contents in the same huge remote sensing image will be the new research focus.

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] X. Qian, S. Lin, G. Cheng, X. Yao, H. Ren, and W. Wang, "Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 143.

[3] S. Rivest, Y. Bédard, M.-J. Proulx, M. Nadeau, F. Hubert, and J. Pastor, "Solap technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 60, no. 1, pp. 17–33, 2005.

[4] X. Yao, Q. Cao, X. Feng, G. Cheng, and J. Han, "Scale-aware detailed matching for few-shot aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2021, Art. no. 5611711.

[5] C. Ivancsits and M.-F. R. Lee, "Visual navigation system for small unmanned aerial vehicles," *Sensor Rev.*, vol. 33, no. 3, pp. 267–291, 2013.

[6] Q. Yu et al., "Full-parameter vision navigation based on scene matching for aircrafts," *Sci. China Inf. Sci.*, vol. 57, no. 5, pp. 1–10, 2014.

[7] G. Panteras and G. Cervone, "Enhancing the temporal resolution of satellite-based flood extent generation using crowdsourced data for disaster monitoring," *Int. J. Remote Sens.*, vol. 39, no. 5, pp. 1459–1474, 2018.

[8] S. S. Durbha, R. L. King, V. P. Shah, and N. H. Younan, "Image information mining for coastal disaster management," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2007, pp. 342–345.

[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, Jan. 2020.

[10] J. Zhang, L. Chen, L. Zhuo, X. Liang, and J. Li, "An efficient hyperspectral image retrieval method: Deep spectral-spatial feature extraction with DCGAN and dimensionality reduction using t-SNE-based NM hashing," *Remote Sens.*, vol. 10, no. 2, 2018, Art. no. 271.

[11] X. Qian, Y. Zeng, W. Wang, and Q. Zhang, "Co-saliency detection guided by group weakly supervised learning," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2022.3167805.

[12] X. Feng, J. Han, X. Yao, and G. Cheng, "TCANet: Triple context-aware network for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6946–6955, Aug. 2021.

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

[14] Q. Jiang and W. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3232–3240.

[15] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. Zuair, "TextRS: Deep bidirectional triplet network for matching text to remote sensing images," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 405.

[16] M. M. A. Rahhal, Y. Bazi, T. Abdullah, M. L. Mekhalfi, and M. Zuair, "Deep unsupervised embedding for remote sensing image retrieval using textual cues," *Appl. Sci.*, vol. 10, no. 24, 2020, Art. no. 8931.

[17] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2021, Art. no. 5606514.

[18] 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.

[19] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3D convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22081–22091, 2017.

[20] H. Zhang, Y. Zhuang, and F. Wu, "Cross-modal correlation learning for clustering on image-audio dataset," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 273–276.

[21] H. Ning, X. Zheng, X. Lu, and Y. Yuan, "Disentangled representation learning for cross-modal biometric matching," *IEEE Trans. Multimedia*, vol. 24, pp. 1763–1774, Apr. 2021.

[22] 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, Sep. 2020.

[23] W. Hu, L. Wu, M. Jian, Y. Chen, and H. Yu, "Cosine metric supervised deep hashing with balanced similarity," *Neurocomputing*, vol. 448, pp. 94–105, 2021.

[24] P. Bosilj, E. Aptoula, S. Lefèvre, and E. Kijak, "Retrieval of remote sensing images with pattern spectra descriptors," *ISPRS Int. J. Geo- Inf.*, vol. 5, no. 12, 2016, Art. no. 228.

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

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

[27] L. Fan, H. Zhao, and H. Zhao, "Distribution consistency loss for large-scale remote sensing image retrieval," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 175.

[28] N. Lukač, B. Žalik, S. Cui, and M. Datcu, "GPU-based kernelized locality-sensitive hashing for satellite image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1468–1471.

[29] T. Reato, B. Demir, and L. Bruzzone, "A novel class sensitive hashing technique for large-scale content-based remote sensing image retrieval," in *Proc. Image Signal Process. Remote Sens. XXIII*, 2017, vol. 10427, pp. 307–315.

[30] 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.

[31] 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.

[32] 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, Apr. 2021.

[33] Z. Yuan et al., "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4404119.

[34] Y. Chen and X. Lu, "A deep hashing technique for remote sensing image-sound retrieval," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 84.

[35] 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.

[36] R. Yang et al., "Cross-modal feature fusion retrieval for remote sensing image-voice retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2855–2858.

[37] M. Müller, *Information Retrieval for Music and Motion*, vol. 2. Berlin, Germany: Springer, 2007.

[38] J. He, S.-F. Chang, R. Radhakrishnan, and C. Bauer, "Compact hashing with joint optimization of search accuracy and time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 753–760.

[39] L. Zhang, Y. Zhang, J. Tang, X. Gu, J. Li, and Q. Tian, "Topology preserving hashing for similarity search," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 123–132.

[40] J. T. Hoe, K. W. Ng, T. Zhang, C. S. Chan, Y.-Z. Song, and T. Xiang, "One loss for all: Deep hashing with a single cosine similarity based learning objective," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 24286–24298, 2021.

[41] 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.

[42] X. Zhao, Z. Li, and J. Yi, "Sift feature-based second-order image hash retrieval approach," *J. Softw.*, vol. 13, no. 2, pp. 103–116, 2018.

[43] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5427–5440, Nov. 2016.

[44] D. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1858–1866.

[45] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 609–617.

[46] 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.

[47] Z. Lin, G. Ding, J. Han, and J. Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4342–4355, Dec. 2017.

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

[49] Z. Wang et al., "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5764–5773.

Yichao Zhang is currently working toward the Ph.D. degree with the Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

He is currently with the University of Chinese Academy of Sciences, Beijing, China. His current research interests include pattern recognition, computer vision, and machine learning.

Xiangtao Zheng (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2014 and 2017, respectively.

He is currently an Associate Professor with the Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His main research interests include computer vision and pattern recognition.

Xiaoqiang Lu (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from Dalian University of Technology, Dalian, China, in 2010.

He is currently a Full Professor with the Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His current research interests include pattern recognition, machine learning, hyperspectral image analysis, cellular automata, and medical imaging.