Semantics-Consistent Representation Learning for Remote Sensing Image-Voice Retrieval

Hailong Ning, Bin Zhao, and Yuan Yuan, *Senior Member, IEEE*

Abstract—With the development of earth observation technology, massive amounts of remote sensing (RS) images are acquired. To find useful information from these images, cross-modal RS image-voice retrieval provides a new insight. This paper aims to study the task of RS image-voice retrieval so as to search effective information from massive amounts of RS data. Existing methods for RS image-voice retrieval rely primarily on the pairwise relationship to narrow the heterogeneous semantic gap between images and voices. However, apart from the pairwise relationship included in the datasets, the intra-modality and non-paired inter-modality relationships should also be taken into account simultaneously, since the semantic consistency among non-paired representations plays an important role in the RS image-voice retrieval task. Inspired by this, a semantics-consistent representation learning (SCRL) method is proposed for RS image-voice retrieval. The main novelty is that the proposed method takes the pairwise, intra-modality, and non-paired inter-modality relationships into account simultaneously, thereby improving the semantic consistency of the learned representations for the RS image-voice retrieval. The proposed SCRL method consists of two main steps: 1) semantics encoding and 2) semantics-consistent representation learning. Firstly, an image encoding network is adopted to extract high-level image features with a transfer learning strategy, and a voice encoding network with dilated convolution is devised to obtain high-level voice features. Secondly, a consistent representation space is conducted by modeling the three kinds of relationships to narrow the heterogeneous semantic gap and learn semantics-consistent representations across two modalities. Extensive experimental results on three challenging RS image-voice datasets, including Sydney, UCM and RSICD image-voice datasets, show the effectiveness of the proposed method.

Index Terms—Remote Sensing Image-Voice Retrieval, Heterogeneous Semantic Gap, Semantics-Consistent Representation

I. INTRODUCTION

WITH the rapid development of remote sensing (RS) technology, tons of remote sensing images have been produced [1]–[3]. There is no doubt that mining useful information from these RS images [4] is very important. However,

H. Ning is with Shaanxi Key Laboratory of Ocean Optics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, P.R. China, the University of Chinese Academy of Sciences, Beijing 100049, P. R. China, and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi¡‾an 710072, P.R. China.

B. Zhao and Y. Yuan are with School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi¡‾an 710072, P.R. China.
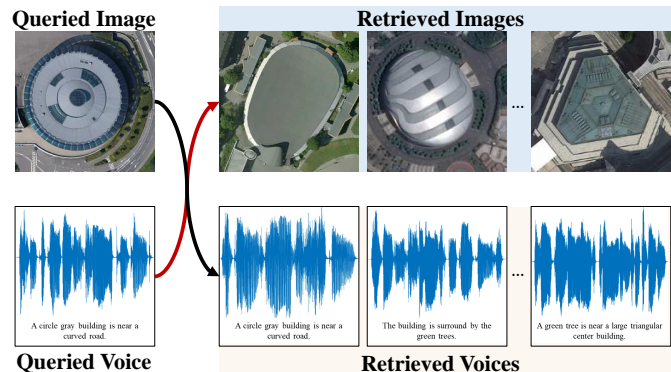

Fig. 1. The schematic diagram of the RS image-voice retrieval task. The goal of the task is to leverage the queried RS image (voice) to retrieve the relevant voices (images).

considering that the RS data is too large, it is unpractical to find the useful information manually due to the time-consuming workload. Driven by the practical demand, many researchers pay attention to the large-scale RS data retrieval task [2], [4]–[6]. It can automatically refine precise semantic information contained in the RS data, and has wide application prospects in military targets detection, military audio intelligence generation, and disaster monitoring scenarios [7]–[9].

Generally, the RS data retrieval methods can be roughly divided into two categories: uni-modal retrieval and cross-modal retrieval. Specifically, the uni-modal RS data retrieval methods conduct the similarity search in the same modality [7], [10]. For instance, Li *et al.* [7] present a large-scale RS image retrieval method to retrieve the RS images with similar semantics. By comparison, the source and target data are from different modalities in cross-modal RS data retrieval methods [8], [11]. For instance, Chaudhuri *et al.* [8] propose a deep neural network to learn a discriminative shared feature space for the input RS images and the corresponding voice based annotations. This paper focuses on the cross-modal RS data retrieval, which is more challenging than uni-modal RS data retrieval due to the heterogeneous semantic gap of cross-modal data.

As shown in Fig. 1, the task of RS image-voice retrieval is to find the relevant RS image (voice) given a queried voice (image). Notice that the task is more convenient for humans than the way by keyboard inputting text to retrieve similar RS images, because it is based on human speeches to retrieve target RS images [11], [12]. As a consequence, the task is more practical in some scenarios that are emergent and not convenient for keyboard input, such as military targets

detection, military audio intelligence generation, and disaster monitoring [9], [11].

Due to the practicability of the RS image-voice retrieval task, several related works have been developed [8], [11]–[14]. These works can be classified into two categories: CNN feature based methods and hash based methods. The former methods conduct the retrieval task by learning CNN features and their consistency relationship [15]. As an example, Guo *et al.* [11] propose a deep visual-audio network to learn the correspondence between RS images and voices. The latter methods encode high-dimensional data points into compact binary code for retrieval [16]. For instance, Chen *et al.* [13], [14] present to integrate the feature learning and hash code learning into a unified framework to achieve fast retrieval property between RS images and voices.

Practically, these works commonly narrow the distance between representations from the paired multi-modalities directly, resulting the similar paired representations for cross-modal retrieval. However, the obtained similar representations by these works are difficult to effectively alleviate the pervasive heterogeneous semantic gap, because the cross-modal retrieval aims to retrieve all the data with the same semantic concept rather than only one data coming from the same pair. As a consequence, it is necessary to fully mine the relationships of both paired and non-paired data within and across modalities. Naturally, we propose to simultaneously model the semantic relationships, including pairwise, intra-modality, and non-paired inter-modality relationships. Specifically, the pairwise relationship models the consistency between the paired image and voice. The intra-modality relationship is to enhance the consistency between two representations from the same modality and belonging to the same semantic concept, thereby avoiding the adverse impact caused by the problems of inter-class similarity and intra-class diversity. The non-paired inter-modality relationship is to enhance the consistency between two non-paired representations from different modalities but belonging to the same semantic concept. In order to further improve the retrieval performance, the pairwise, intra-modality, and non-paired inter-modality relationships should be considered comprehensively.

Inspired by the discussion above, a semantics-consistent representation learning (SCRL) method is proposed to make full use of these three relationships for the RS image-voice retrieval. As depicted in Fig. 2, the proposed SCRL method can be divided into two steps: 1) semantics encoding and 2) semantics-consistent representation learning. Firstly, the VGG16 network [17] is adopted to automatically extract high-level RS image features with a transfer learning strategy. Meanwhile, the input voices are processed as Mel-Frequency Cepstral Coefficients (MFCC) features [18] and input into a 1-D dilated convolutional network to automatically extract high-level voice features. Secondly, to explicitly compute the similarity between representations from different modalities, a semantics-consistent representation space is explored by considering three kinds of relationships simultaneously. The relationships include pairwise, intra-modality, and non-paired inter-modality relationships. They are measured by the consistency loss to narrow the heterogeneous semantic gap

across two modalities. In addition, to further learn significantly discriminative and more compact semantic representations, a classification loss is adopted for each branch. Afterwards, the consistency loss and the classification loss are combined jointly to learn ultimate semantics-consistent representations for the RS image-voice retrieval.

To sum up, the main contributions can be summarized into threefolds:

- A novel RS image-voice retrieval method, SCRL, is proposed to fully explore the pairwise, intra-modality, and non-paired inter-modality relationships simultaneously. Then, semantics-consistent representations across the two modalities can be learned to effectively alleviate the pervasive heterogeneous semantic gap.
- An effective voice encoding network is proposed to learn high-level voice features. The network can capture the long range correlation of voice signal by introducing the 1-D dilated convolutional kernel.
- Experimental results on three challenging RS image-voice datasets show that the proposed SCRL method can significantly improve the performance of the RS image-voice retrieval. Especially, the mAP value can be improved by nearly 9% compared with the state-of-the-arts.

The remainder of this paper is arranged as follows: Section II introduces the related works about the task of RS image-voice retrieval. Section III elaborates the proposed SCRL method. Section IV shows and analyzes the experimental results. Finally, Section V presents the conclusion.

II. RELATED WORKS

In this section, the related works about this paper are primarily restricted on uni-modal and cross-modal RS data retrieval methods.

A. Uni-Modal RS Data Retrieval Methods

Uni-modal RS data retrieval methods aim to search similar RS data in the same modality. Some early works are based on the hand-crafted feature [19]–[21]. For instance, Luo *et al.* [19] present a method by comparing multiple-resolution wavelet features for satellite images retrieval. Yang and Newsam [20] leverage local invariant features for RS image retrieval. Aptoula [21] proposes to apply global morphological texture descriptors for RS image retrieval. However, these methods are not suitable for large-scale RS image retrieval since they are based on low-level hand-crafted features. With the rapid development of artificial intelligent technology [22]–[26], a large number of uni-modal RS data retrieval methods based on deep learning have emerged. For instance, Tang *et al.* [27] develop a two-stage re-ranking method to improve the retrieval performance. Shao *et al.* [10] design a fully convolutional network to solve the problem of poor retrieval performance due to multiple labels for single images. Ye *et al.* [28] utilize weighted distance as similarity criteria to learn determinative representations for the RS image retrieval. To address the RS image retrieval problem due to multiple land-cover classes,
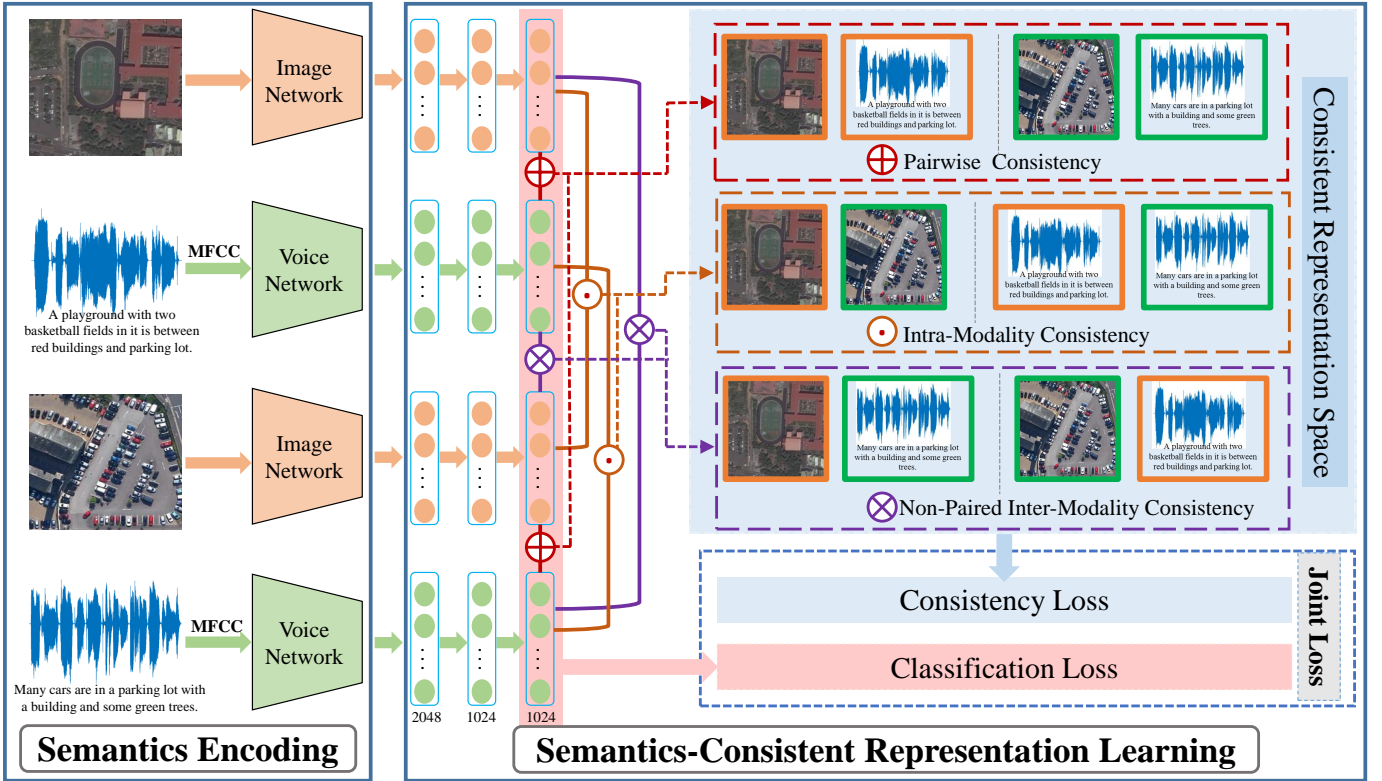
Fig. 2. The overall framework of the proposed SCRL method. Firstly, the image encoding network is adopted to extract high-level image features, and the voice encoding network is devised to obtain high-level voice features. Secondly, the extracted image features and voice features are fed into the representation step to explore the consistent representation space. By exploring the space, the heterogeneous semantic gap is narrowed and the semantics-consistent representations across two modalities are learned by comprehensively considering the pairwise consistency ($\oplus$), intra-modality consistency ($\odot$), and non-paired inter-modality consistency ($\otimes$). MFCC represents the operation to extract the Mel-Frequency Cepstral Coefficients feature of the input voice.

Chaudhuri *et al.* [29] introduce a semi-supervised graph-theoretic method with only a small number of training images characterized by multi-labels. Kang *et al.* [30] propose a graph relation network to preserve the complex semantic relations pervading RS scenes for multi-label RS image retrieval. In response to the scarcity of labeled images, Liu *et al.* [31] present an unsupervised deep transfer learning method based on the similarity learning for the RS image retrieval. Ye *et al.* [32] propose to extract the domain-invariant feature for the RS image retrieval using an unsupervised domain adaptation model. To obtain the quicker response during retrieval, several techniques consider binary features [4], [5], [33]. Song *et al.* [33] develop an image retrieval and compression method based on binary generative adversarial network. Demir and Bruzzone [5] introduce two kernel-based nonlinear hashing methods for the RS image retrieval to reduce the amount of memory required. Li *et al.* [4] introduce a RS image retrieval method, which integrates the feature learning network and hashing network into a unified framework.

B. Cross-Modal RS Data Retrieval Methods

The objective of cross-modal RS data retrieval methods is to search similar RS data across different modalities. Due to the urgent demand for RS data analysis, cross-modal RS data

retrieval attracts increasing attention from researchers. Li *et al.* [2] propose an end-to-end training method to learn source-invariant features for the cross-source RS image retrieval. Xiong *et al.* [34] present a discriminative distillation network to eliminate the inconsistency across different modalities. To develop an effective retrieval system in harsh environments, Xiong *et al.* [35] design another deep cross-modality hashing network for the Optical and SAR RS images retrieval. Recently, a more natural cross-modal RS data retrieval method has emerged, which is based on RS images and human voices. As the first attempt for the task of RS image-voice retrieval task, Guo *et al.* [11] devise a deep visual-audio network to learn the correspondence between RS images and voices. Subsequently, they develop another cross-modal RS image-voice retrieval method to further improve the retrieval performance [12]. Based on the works of Guo *et al.*, Chen and Lu [13] present to leverage the triplet similarity of deep features to solve the insufficient utilization problem for the relative semantic similarity relationships. In addition, Chen *et al.* [14] integrate the feature learning and hash code learning into a unified framework to achieve fast retrieval property. As another implementation, Chaudhuri *et al.* [8] propose a deep neural network to learn a common embedding of RS images and spoken words rather than complete spoken sentences.

Song *et al.* [36] adopt an inter-media hashing (IMH) model for exploring the data correlations from heterogeneous data sources.

This paper is dedicated to conduct the cross-modal RS image-voice retrieval by fully considering the semantic relationships among representations across different modalities. The proposed method comprehensively improves the pairwise, intra-modality, and non-paired inter-modality consistency so as to narrow the heterogeneous semantic gap and learn semantics-consistent representations for RS image-voice retrieval.

III. THE PROPOSED METHOD

The proposed SCRL method is composed of two steps: 1) semantics encoding and 2) semantics-consistent representation learning. In this section, the overall framework of the proposed method is described firstly. Secondly, the semantics encoding and semantics-consistent representation learning steps are elaborated, respectively. Thirdly, the loss functions are presented. Finally, the optimizing strategy is introduced.

A. Overall Framework

As shown in Fig. 2, the proposed SCRL method consists of four parallel branches and can be divided into two steps, including semantics encoding and semantics-consistent representation learning steps. Two of the branches are related to the image modality, and the rest are related to the voice modality. The branches, which are responsible for the same modality, share the same parameters. As a consequence, the features from the same modality own the same transforms. After each iteration, four features from the two modalities are obtained by the semantics encoding step. Then, the obtained features are used as the input of the semantics-consistent representation learning step to learn four representations. During this step, the relationships of different representations are explored in a defined consistent representation space. The relationships include pairwise, intra-modality and non-paired inter-modality relationships. All the relationships are sought by calculating the distance between two representations to preserve the semantic consistency across different modalities. Specifically, the pairwise relationship is built to make the semantic information be consistent for the paired RS image-voice. The intra-modality relationship is considered to keep the semantic consistency within each modality. The non-paired inter-modality relationship is developed to enforce the semantic information to be consistent across different modalities for unpaired samples. Afterwards, the three kinds of relationships are measured by a consistency loss. Finally, we adopt a joint loss function to simultaneously minimize the consistency loss of the consistent representation space and the classification loss of each branch.

The formal definitions in this paper are given as follows. Let $\mathcal{D} = \left\{ \left(\mathbf{X}_m^I, \mathbf{X}_m^V, \mathbf{y}_m \right) \right\}_{m=1}^N$ be the collection of N instances of RS image-voice pairs, where $\mathbf{X}_m^I$ is the m-th input image sample, $\mathbf{X}_m^V$ is the m-th input voice sample, and $\mathbf{y}_m$ is the m-th semantic label. The goal of the semantics encoding step is to learn two mapping functions $\mathcal{F}_I$ and $\mathcal{F}_V$ for extracting high-level image features $\left\{ \mathbf{s}_m^I \right\}_{m=1}^N$ and

voice features $\left\{ \mathbf{s}_m^V \right\}_{m=1}^N$. Afterwards, the semantics-consistent representation learning step aims to learn two mapping functions $\mathcal{R}_I$ and $\mathcal{R}_V$ for obtaining the semantics-consistent image representations $\left\{ \varphi_m^I \right\}_{m=1}^N$ and voice representations $\left\{ \varphi_m^V \right\}_{m=1}^N$. The obtained semantics-consistent representations are exploited for RS image-voice retrieval.

B. Semantics Encoding

As shown in Fig. 2, the semantics encoding step consists of four parallel networks, in which the two image encoding networks share the same parameters and the two voice encoding networks share the same parameters. The details about image encoding networks and voice encoding networks are described as follows.

1) Image Semantics Encoding: According to the previous works [37], [38], the high-quality feature is quite important for the retrieval task. CNN features have been proved to be high-efficiency for the retrieval task because of the success in expressing the high-level semantic information. To learn high-quality CNN features, massive amounts of data with manual annotations, such as ImageNet, are necessary. Unfortunately, large-scale dataset with manual annotation in RS domain is unavailable. In order to solve this problem, a transfer learning strategy has emerged. The transfer learning strategy applies knowledge learned from one domain to other domains [39]. In this paper, the transfer learning strategy is leveraged to extract high-quality image features. To be more specific, the VGG16 network [17] pre-trained on ImageNet is adopted to acquire the RS image features since some useful texture features may share in natural images and RS images [40].

When using the pre-trained VGG16 network for image semantics encoding, the fully connected layers are removed to obtain high-level image features. Following the previous work [41], the network is normalized by scaling weights to ensure the mean activation of each convolutional filter over images and positions to be one. Specifically, the mean activation s_i^l of each convolutional filter in the l-th layer over the whole training image set $\mathcal{X}$ and all M_l spatial locations can be calculated as follows:

$$s_i^l = \frac{1}{NM_l} \sum_{\mathcal{X}} \sum_{j=1}^{M_l} \text{relu}(w_i^l * x_j^{l-1} + b_i^l), \quad (1)$$

where w_i^l represents the weight of i-th convolutional filter in the l-th layer, x_j^{l-1} stands for the j-th patch in the $(l-1)$-th layer, $*$ indicates the convolutional operation, b_i^l denotes the bias term in the l-th layer, and N is the total number of the training images. Then, the parameters w_i^l and b_i^l are scaled by $\frac{1}{s_i^l}$ as follows:

$$\mathbb{E}_{\mathcal{X},j}[\text{relu}(\frac{w_l^i}{s_i^l} * x_j^{l-1} + \frac{b_l^i}{s_i^l})] = 1. \quad (2)$$

According to the above process, the normalization is conducted from bottom layers to top layers sequentially. The process of image semantics encoding can be written as:

$$\mathbf{s}_m^I = \mathcal{F}_I(\mathbf{X}_m^I; \theta_I), \quad (3)$$

where $\mathbf{s}_m^I \in \mathbb{R}^{h_1 \times w_1 \times c_1}$ represents the extracted semantics feature of the m-th image $\mathbf{X}_m^I$, $\mathcal{F}_I$ indicates the mapping function from image to the corresponding feature, and θ_I is the parameter.

2) Voice Semantics Encoding: Voice belongs to a continuous one-dimensional signal and the word in it cannot be distinguished easily by computer. In order to quantify the continuous voice, Mel-Frequency Cepstral Coefficients (MFCC) feature is first extracted to representant the voice signal according to previous works [18], [42]. Here, the MFCC feature is expressed as $\mathbf{X}_m = \Upsilon(\mathbf{X}_m^V)$.

The obtained MFCC feature is reshaped, and then fed into the voice encoding network to perform high-level voice features. To capture long range correlations, 1-D dilated convolutional kernels are adopted to construct the voice network. Specifically, the voice network is composed of 5 cascaded 1-D convolutional layers and pooling layers, where each convolutional layer adopts dilated convolutional kernel. The flattened MFCC feature is fed into the voice network and the learned high-level voice feature can be denoted as:

$$\mathbf{s}_m^V = \mathcal{F}_V(\mathbf{X}_m; \theta_V), \tag{4}$$

where $\mathbf{s}_m^V \in \mathbb{R}^{h_2 \times w_2 \times c_2}$ represents the extracted high-level feature of the m-th MFCC feature $\mathbf{X}_m$, $\mathcal{F}_V$ indicates the mapping function from MFCC feature to the corresponding high-level voice feature, and θ_V is the parameter.

C. Semantics-Consistent Representation Learning

This step aims to learn semantics-consistent representations by comprehensively considering the pairwise, intra-modality, and non-paired inter-modality relationships across different modalities. As shown in Fig. 2, these three relationships are measured by a consistency loss to model the similarity of the learned representations. In addition, a classification loss is adopted for each branch to enhance the semantic discrimination ability of representations, resulting in more compact representations. The consistency loss and classification loss are combined jointly to learn semantics-consistent representations for the RS image-voice retrieval. Hereinafter, the details about the semantics-consistent representation learning are dwelled on.

1) The Architecture of Semantics-Consistent Representation Learning: As shown in Fig. 2, the semantics-consistent representation learning step contains 4 parallel branches. Similar to the semantics encoding step, two of the branches are related to the image modality, and the rest are related to the voice modality. Actually, the branches responsible for the same modality shares the same weights. Each branch is constructed by three cascaded fully connected layers. The specific parameters of each branch are shown in Fig. 2. The features from the semantics encoding step are taken as inputs for the semantics-consistent representation learning step. Then, the representations are output from each branch in the representation learning step.

As for the image branch, the learned image representation can be expressed as:

$$\varphi_m^I = \mathcal{R}_I(\mathbf{s}_m^I; \Theta_I), \tag{5}$$

where $\varphi_m^I \in \mathbb{R}^{d_1}$ represents the learned semantics-consistent representation for the m-th image, $\mathcal{R}_I$ indicates the mapping function from the image feature to the corresponding semantics-consistent representation, and Θ_I is the parameter.

As for the voice branch, the learned voice representation can be expressed as:

$$\varphi_m^V = \mathcal{R}_V(\mathbf{s}_m^V; \Theta_V), \tag{6}$$

where $\varphi_m^V \in \mathbb{R}^{d_2}$ represents the learned semantics-consistent representation for the m-th voice, $\mathcal{R}_V$ indicates the mapping function from the voice feature to the corresponding semantics-consistent representation, and Θ_V is the parameter.

2) The Pairwise Consistency: The pairwise consistency represents the semantic information of representations from the paired RS image-voice should be consistent since they describe the same semantic concept in the applied RS image-voice datasets. To measure the similarity of representations, the cosine distance is adopted as it is commonly used in cross-modal retrieval [43], [44]. The cosine distance can be defined as:

$$D(\mathbf{x}, \mathbf{y}) = \mathbf{1} - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}, \tag{7}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent two vectors with the same length, and $\cdot$ indicates the operation of dot product between the two vectors.

Based on the cosine distance, the pairwise consistency can be maintained by the pairwise consistency loss, which is defined as:

$$\mathcal{L}_{pair} = D(\varphi_i^I, \varphi_i^V), \tag{8}$$

where φ_i^I and φ_i^V denote representations of the i-th image and voice, respectively. Minimizing the pairwise consistency loss leads to a common space, where representations describing the same semantic concept are clustered together and the pairwise consistency is preserved.

3) The Intra-Modality Consistency: The intra-modality consistency measures the relationships between two representations from the same modality and belonging to the same semantic concept. In order to preserve the intra-modality consistency, an intra-modality consistency loss is introduced, which is defined as:

$$\begin{aligned} \mathcal{L}_{intra} = &\hbar \left(1 - \ell_{ij} \left(\xi - D\left(\varphi_i^I, \varphi_j^I\right)\right)\right) \\ &+ \hbar \left(1 - \ell_{ij} \left(\xi - D\left(\varphi_i^V, \varphi_j^V\right)\right)\right), \end{aligned} \tag{9}$$

where $\hbar(x) = max(0, x)$ is the hinge function, ξ is the pre-defined margin threshold. ℓ_{ij} denotes the label indicator. If φ_i^I and φ_j^I, or φ_i^V and φ_j^V describe the same semantic concept, $\ell_{ij} = +1$; otherwise, $\ell_{ij} = -1$. It is to noted that $i \neq j$. Minimizing the intra-modality consistency loss results in a common space, where representations describing the same semantic concept within the same modality are enforced to be close, while representations describing different semantic concepts within the same modality are enforced to be far.

4) The Non-Paired Inter-Modality Consistency: The non-paired inter-modality consistency measures the relationships between two representations from different modalities but belonging to the same semantic concept. Since the pairwise loss only model the relationships of paired image-voice pairs

with the same semantic concept in the applied RS image-voice datasets, the relationships between other non-paired data cannot been explored effectively, and the compact representations for different semantic concepts are unable to be learned. To this end, the inter-modality consistency loss is used to make up for the lack of the pairwise loss. To maintain the non-paired inter-modality consistency, the inter-modality consistency loss is defined as:

$$\mathcal{L}_{inter} = \hbar \left(1 - \ell_{ij} \left(\zeta - D \left(\varphi_i^V, \varphi_j^I \right) \right) \right) \\ + \hbar \left(1 - \ell_{ij} \left(\zeta - D \left(\varphi_i^I, \varphi_j^V \right) \right) \right), \quad (10)$$

where ζ is the pre-defined margin threshold. If φ_i^V and φ_j^I, or φ_i^I and φ_j^V describe the same semantic concept, $\ell_{ij} = +1$; otherwise, $\ell_{ij} = -1$. Minimizing the inter-modality consistency loss leads to a common space, where representations describing the same semantic concept across the two modalities are enforced to be close, while representations describing different semantic concepts across the two modalities are enforced to be far.

D. Loss Function

1) The Consistency Loss: The consistency loss is defined in the consistent representation space to model the similarity of the learned representations, which combines the pairwise, intra-modality and non-paired inter-modality consistency mentioned above. As a consequence, the consistent representation space possesses all the advantages which are contained in all three common spaces obtained by pairwise, intra-modality and inter-modality consistency loss, respectively. Considering that the intra-modality and inter-modality loss are calculated from the same two pairs of samples, the significance of the intra-modality loss and the inter-modality consistency loss is set equal. Therefore, the consistency loss in the consistent representation space is defined as:

$$\mathcal{L}_{consi} = \mathcal{L}_{pair} + \eta_1(\mathcal{L}_{intra} + \mathcal{L}_{inter}), \quad (11)$$

where η_1 is the trade-off coefficient controlling the contribution of the last two terms.

2) The Classification Loss: As for the representation on the top of each branch, a classification process is added to mine the intrinsic semantic information in each image and voice and further to model the discrimination and compactness of the learned representations. Specifically, a softmax layer is adopted for each image branch as:

$$p_m^I = \text{softmax}(\mathbf{W}^I \varphi_m^I + \mathbf{b}^I), \quad (12)$$

where p_m^I is the probability of belonging to the corresponding semantic concept for the m-th image. softmax represents the softmax activation function. $\mathbf{W}^I$ and $\mathbf{b}^I$ are the parameters in the softmax layer. Similarly, a softmax layer is adopted for each voice branch as:

$$p_m^V = \text{softmax}(\mathbf{W}^V \varphi_m^V + \mathbf{b}^V), \quad (13)$$

where p_m^V is the probability of belonging to the corresponding semantic concept for the m-th voice. $\mathbf{W}^V$ and $\mathbf{b}^V$ are the parameters in the softmax layer.

Then, the classification loss can be defined as:

$$\mathcal{L}_{class} = -y_{it}\log(p_{it}^I + \epsilon) - y_{it}\log(p_{it}^V + \epsilon), \quad (14)$$

where y_{it} is the true semantic label of the i-th sample, where t indexes the t-th class. p_{it}^I is the predicted probability distribution of the i-th images. p_{it}^V is the predicted probability distribution of the i-th voices, where t indexes the t-th class. ϵ represents a regularization constant to avoid the NaN value in the loss. By minimizing the classification loss function, the semantic discrimination ability of the common representations in the consistent representation space can be greatly enhanced.

Algorithm 1 The proposed SCRL method

Input:
  Training collection $\mathcal{D} = \left\{ \left(\mathbf{X}_m^I, \mathbf{X}_m^V, \mathbf{y}_m \right) \right\}_{m=1}^N$ of RS image-voice pairs;
  Learning rate lr, trade-off coefficients η_1 and η_2, iterative epoch K.

Output:
  The parameters θ_I and θ_V in the feature extraction step;
  The parameters Θ_I and Θ_V in the representation learning step.

Initialization:
  The parameter θ_I is initialized by the pre-trained VGG16 network;
  The parameters θ_V, Θ_I and Θ_V are randomly initialized by truncated_normal distribution.

Repeat:
  1: Sample the input RS image-voice pairs $\left(\mathbf{X}_i^I, \mathbf{X}_i^V, \mathbf{y}_i \right)$ and $\left(\mathbf{X}_j^I, \mathbf{X}_j^V, \mathbf{y}_j \right)$;
  2: Calculate the image features $\mathbf{s}_i^I$ and $\mathbf{s}_j^I$ via Eq. 3, and voice features $\mathbf{s}_i^V$ and $\mathbf{s}_j^V$ according to Eq. 4;
  3: Learn the semantics-consistent representations φ_i^I and φ_j^I according to Eq. 5, as well as φ_i^V and φ_j^V via Eq. 6;
  4: Calculate the joint loss $\mathcal{L}_{joint}$ according to Eq. 16;
  5: Update the parameters θ_V, Θ_I and Θ_V by utilizing RMSProp.

Until: A fixed iterative epoch K or convergence.
Return: θ_I, θ_V, Θ_I and Θ_V

---

3) The Joint Loss: With the above definition, a joint loss function is raised to simultaneously calculate the consistency loss of the consistent representation space and the classification loss of each branch. The joint loss function can be written as:

$$\mathcal{L}_{joint} = \mathcal{L}_{consi} + \eta_2 \mathcal{L}_{class}, \quad (15)$$

where η_2 is the trade-off coefficient controlling the contribution of the second term. By combining Eq. 11 and Eq. 15, the joint loss can be rewritten as:

$$\mathcal{L}_{joint} = \mathcal{L}_{pair} + \eta_1(\mathcal{L}_{intra} + \mathcal{L}_{inter}) + \eta_2 \mathcal{L}_{class}. \quad (16)$$

By minimizing the joint loss, the semantics-consistent representations are learned for the RS image-voice retrieval.

E. Optimizing Strategy

Based on the joint loss, the proposed SCRL method is optimized as follows. The parameter θ_I in the image encoding

network is initialized by VGG16 network pre-trained on the ImageNet [45]. The parameters θ_V in the voice encoding network, Θ_I and Θ_V in the semantics-consistent representation step are randomly initialized by truncated_normal distribution [46]. In each training iteration, the optimizing process can be divided into five main steps. Firstly, we sample two pairs of RS image-voice from the training collection of RS image-voice pairs. Secondly, the sampled RS image-voice pairs are input into the semantics encoding step to obtain image and voice features according to Eq. 3 and Eq. 4, respectively. Thirdly, the extracted image and voice features are taken as input of the semantics-consistent representation step for learning semantics-consistent representations via Eq. 5 and Eq. 6. Fourthly, the joint loss is calculated according to Eq. 16. Finally, the parameters θ_V, Θ_I and Θ_V are updated by minimizing the joint loss with the RMSprop optimization algorithm [47]. When a fixed iterative epoch, which is set as 50, is reached or the model is convergent, the optimizing process is terminated. Afterwards, the parameters θ_I, θ_V, Θ_I and Θ_V are leveraged to compute ultima semantics-consistent representations for the RS image-voice retrieval. Noted that the proposed SCRL method is trained end-to-end. The details about the optimization process of the proposed SCRL method are shown in Algorithm 1.

IV. EXPERIMENT AND RESULTS

In this section, experimental datasets, implementation details, evaluation metrics and parameter analysis are introduced. In addition, the experimental results are compared with some state-of-the-art methods and the ablation analysis is presented to prove the effectiveness of the proposed SCRL method.

A. Datasets

In order to verify the proposed SCRL method, three challenging RS image-voice datasets are applied, including Sydney, UCM and RSICD image-voice datasets [11]. A brief introduction about the applied RS image-voice datasets is given as follows.

1) Sydney Image-Voice Dataset [11]: The Sydney image-voice (Sydney IV) dataset contains 613 RS images and 3065 voices of 7 classes, where each image corresponds to five different voices. In this paper, we follow the previous works [12]–[14], and randomly sample a voice from the five voices for each image to construct RS image-voice pairs. As for the data partitioning, 80% RS image-voice pairs are randomly selected for training, and the rest 20% are selected for testing.

2) UCM Image-Voice Dataset [11]: The UCM image-voice (UCM IV) dataset includes 2100 RS images and 10500 voices, where each image corresponds to five different voices. Note that the dataset can be divided into 21 classes, where each class includes 100 images and 500 voices. In this paper, we follow the previous works [12]–[14], and randomly sample a voice from the five voices for each image to construct RS image-voice pairs. As for the data partitioning, 80% RS image-voice pairs are randomly selected as training set, and the rest 20% are selected as testing set.

TABLE I
THE DETAILED ARCHITECTURE OF THE VOICE NETWORK.

Layer	Output-Size	Kernel	Stride
Conv1	$24000\times1\times1$	$7\times1\times1$	$1\times1\times1$
Conv2	$12000\times1\times6$	$3\times1\times6$	$1\times2\times1$
Conv3	$6000\times1\times12$	$3\times1\times12$	$1\times2\times1$
Pool1	$3000\times1\times12$	$1\times2\times1$	$1\times2\times1$
Conv4	$1500\times1\times24$	$3\times1\times24$	$1\times2\times1$
Conv5	$750\times1\times48$	$3\times1\times48$	$1\times2\times1$
Pool2	$375\times1\times48$	$1\times2\times1$	$1\times2\times1$

3) RSICD Image-Voice Dataset [11]: The RSICD image-voice (RSICD IV) dataset involves 10921 RS images and 54605 voices of 30 classes, where each image corresponds to five different voices. In this paper, we follow the previous works [12]–[14], and randomly sample a voice from the five voices for each image to construct RS image-voice pairs. As for the data partitioning, 80% RS image-voice pairs are randomly selected for training, and the rest 20% are selected for testing.

B. Implementation Details

In this work, the proposed SCRL method contains four branches, in which two branches are responsible for the image modality, and the others are responsible for the voice modality. As for branches, responsible for the same modality, the parameters are sharing. The specific architecture of the voice encoding network is reported in TABLE I. The dilation rate in the first convolutional layer is set as 3 and the others are set as 2. The input images are uniformly adjusted to 224×224. Before input into the voice network, the voices are resampled at 22050Hz and preprocessed as MFCC features by a window with size of 16 millisecond and shift of 5 millisecond. From the output of the semantics encoding step, the image features are with size of $7\times7\times512$ (namely $h_1 = w_1 = 7$ and $c_1 = 512$), and the voice features are with size of $375\times1\times48$ (namely $h_2 = 375$, $w_2 = 1$ and $c_2 = 48$). Before input into the semantics-consistent representation learning step, the image features are adjusted as size of 512 by the operation of global average pooling [48], and the voice features are adjusted as size of 18000 by the operation of flattening. In the semantics-consistent representation step, $ReLU$ and $Tanh$ are applied as activation functions of image branches and voice branches, respectively [49], [50]. The ultima semantics-consistent representations are with size of 1024 (namely $d_1 = d_2 = 1024$).

The proposed method is optimized utilizing a RMSProp optimizer [46], in which the weight decay is set as 0.0005 and momentum is set to 0.9. The learning rate is set as 0.0004. The batch size is set to 32. The trade-off coefficients η_1 and η_2 are set as 1 and 0.1, respectively, which are determined by a grid search strategy [14]. The experiment is conducted on the PC with a TITAN X (Pascal) GPU and 64G RAM.

C. Evaluation Metrics and Comparison Methods

In this work, two protocols are adopted, including using images as queried samples to retrieve the corresponding voices (I→V) and using voices as queried samples to retrieve the
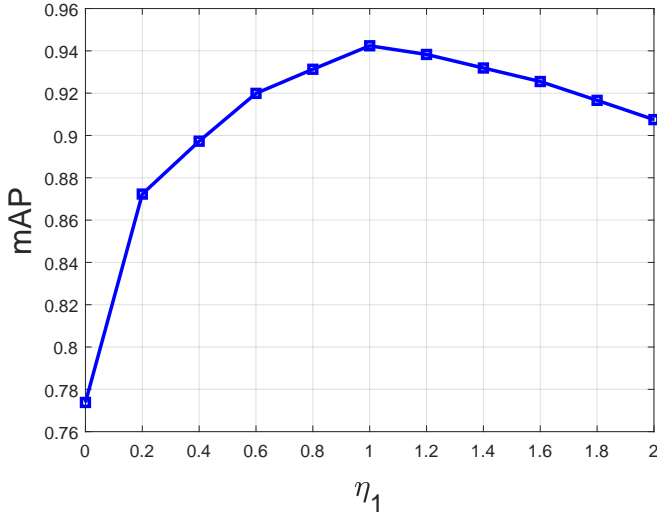
Fig. 3. The retrieval performance with different values for η_1 by the I→V protocol on the Sydney IV dataset.

Fig. 4. The retrieval performance with different values for η_2 by the I→V protocol on the Sydney IV dataset.

corresponding images (V→I). To evaluate the performance of the proposed SCRL method, two commonly used evaluation metrics are adopted, including mean average precision (mAP) [51] and the precision of the top-k ranking result (**P@k**) [52]. The mAP score measures the mean value of average precision, which considers the precision and the returned ranking results at the same time. Precision is the ratio of the returned relevant samples to the queried samples. The **P@k** score measures the precision of the top-k retrieved samples. In this paper, the **P@k** score is reported when k equals to 1, 5 and 10, which is denoted as **P@1**, **P@5** and **P@10**, respectively. In addition, the precision curve is shown when the number of retrieved samples changes to further evaluate the proposed SCRL method.

In order to assess the effectiveness of the proposed SCRL method, 7 comparison methods are adopted, including SIFT+M [53], DBLP [54], CNN+SPEC [55], DVAN [11], CMIR-NET [8], DIVR [14], and DTBH [13] methods. The SIFT+M method [53] leverages SIFT features of images and MFCC features of voices to perform the RS image-voice retrieval. The DBLP method [54] adopts an unsupervised manner to learn the coherence between audio and visual modalities. The CNN+SPEC method [55] seamlessly unifies the learning of different modalities in an unsupervised manner. The DVAN method [11] presents a novel image-voice learning framework to learn the cross-modal similarity. The CMIR-NET method [8] proposes to learn the discriminative shared feature space of the input data for RS image-voice retrieval. The DIVR method [14] leverages multi-scale context information guiding the low-dimensional hash code for RS image-voice retrieval. The DTBH method [13] establishes a deep triplet-based hashing method, which integrates the hash code learning and the representation learning into a unified framework. These methods are implemented in this work. It is to note that the hash-based retrieval methods, including DTBH and DIVR methods, use a 64-bit hash code for comparison. The experimental results and the corresponding analysis are given as follows.
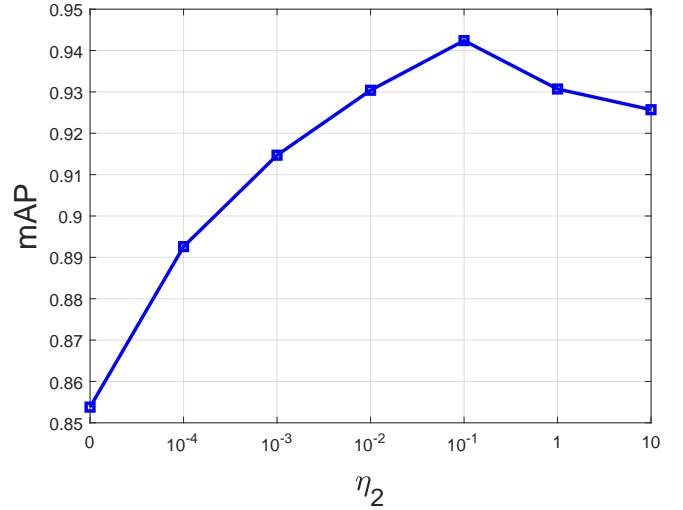
D. Parameter Analysis

The hyperparameter η_1 controls the contributions of the intra-modality and non-paired inter-modality consistency, and η_2 controls the contribution of modeling the semantic classification. To study the performance impacts of them, we conduct parameter experiments about the two parameters with the I→V protocol on the Sydney IV dataset. To this end, a grid search strategy is adopted to tune the two parameters following the previous work [56]. Specially, the range of η_1 and η_2 are set as $\{0, 0.2, 0.4, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$ and $\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10\}$, respectively. Note that $\eta_1 = 0$ represents the SCRL with $\mathcal{L}_{pair}$ and $\mathcal{L}_{class}$ only, and $\eta_2 = 0$ represents the SCRL with $\mathcal{L}_{consi}$ only. The parameter experiments are conducted by changing one hyperparameter (*e.g.*, η_1) while fixing the other (*e.g.*, η_2). The detailed implementation and analysis are as follows.

Firstly, the value of η_2 is fixed at 1 empirically, η_1 is changed from 0 to 2 with increment 0.2 per step. The corresponding mAP scores about the parameter η_1 are recorded and shown in Fig. 3. From the figure, it can be seen that both lower and higher values for η_1 result in lower mAP value, and the best mAP is obtained when η_1 equals to 1. Notice that when $\eta_1 = 0$, the retrieval performance of the proposed method is significantly reduced, because the neglect of the intra-modality consistency and non-paired inter-modality consistency leads to the same semantic concept across the two modalities not being matched correctly.

Secondly, the value of η_1 is fixed at 1, η_2 is changed from 0 to 10 with a ten times increment each step. Fig. 4 shows the performance for different η_2 values. From the figure, we can observe the best performance is obtained when $\eta_2 = 0.1$. It is noted that when η_2 equals to 0, the retrieval performance gets worse slightly, which indicates the classification loss contributes to enhancing the semantic discrimination ability and compactness of the semantics-consistent representations.

As a result, the trade-off coefficients η_1 and η_2 in the proposed SCRL method are set to 1 and 0.1, respectively.

TABLE II
THE COMPARISON RESULTS BETWEEN THE PROPOSED SCRL
METHOD AND OTHER METHIDS ON THE SYDNEY IV DATASET.

Protocols	Methods	mAP	P@1	P@5	P@10
I→V	SIFT+M [53]	31.67	11.21	35.00	37.59
	DBLP [54]	44.38	56.51	52.65	49.68
	CNN+SPEC [55]	46.67	58.62	55.00	51.64
	DVAN [11]	71.77	75.86	73.62	72.93
	CMIR-NET [8]	78.44	84.68	82.54	81.04
	DIVR [14]	81.35	88.26	86.35	84.47
	DTBH [13]	92.45	97.41	95.63	93.78
	SCRL-Pair	75.80	80.11	75.07	71.90
	SCRL-Intra	79.79	85.43	81.20	74.78
	SCRL-Inter	80.19	86.24	82.50	76.57
	SCRL-Class	85.38	90.87	88.11	85.99
	SCRL-Dilation	92.08	93.68	93.03	91.65
	SCRL	**94.24**	**95.50**	**95.17**	**93.95**
V→I	SIFT+M [53]	26.50	34.48	24.48	23.28
	DBLP [54]	34.87	21.63	26.78	30.94
	CNN+SPEC [55]	35.72	17.24	27.76	31.21
	DVAN [11]	63.88	67.24	63.34	67.07
	CMIR-NET [8]	71.28	76.69	74.52	71.60
	DIVR [14]	75.97	80.44	78.05	76.27
	DTBH [13]	87.49	92.18	90.36	88.82
	SCRL-Pair	75.19	78.67	76.75	75.12
	SCRL-Intra	79.96	85.37	84.55	83.17
	SCRL-Inter	80.52	85.99	85.15	83.69
	SCRL-Class	84.87	88.62	87.64	86.75
	SCRL-Dilation	91.14	92.68	89.59	89.11
	SCRL	**93.53**	**94.31**	**91.38**	**90.00**

E. Ablation Analysis

In this subsection, five variations of the proposed SCRL method are conducted on three challenging RS image-voice datasets to examine: 1) the importance of the pairwise consistency loss; 2) the impact of the intra-modality consistency loss; 3) the importance of the inter-modality consistency loss; 4) the importance of the classification loss; 5) the effect of the dilated convolutional kernel. The details about the implementation of different variations are elaborated as follows. Firstly, the variation of the proposed SCRL method without the pairwise consistency loss (SCRL-Pair) is implemented to verify the effect of the pairwise consistency loss. Secondly, the intra-modality consistency loss is discarded evolving into a new variation (SCRL-Intra). Thirdly, we omit the inter-modality consistency loss obtaining another new variation (SCRL-Inter) to determine the importance of the non-paired inter-modality relationship. Fourthly, the classification loss is abandoned evolving into the variation SCRL-Class. Finally, the dilation rate in the voice network is set as 1 generating the new variation SCRL-Dilation. The results of these variations are reported in TABLE II-IV. According to the results, we give the analysis as the following aspects.

1) Pairwise Consistency: The pairwise consistency loss aims to narrow the distance between the representations from the RS image-voice pair. From the results in TABLE II-IV, we can intuitively find that the retrieval performance is greatly improved when the pairwise consistency is taken into account. Concretely, compared with the method without the pairwise
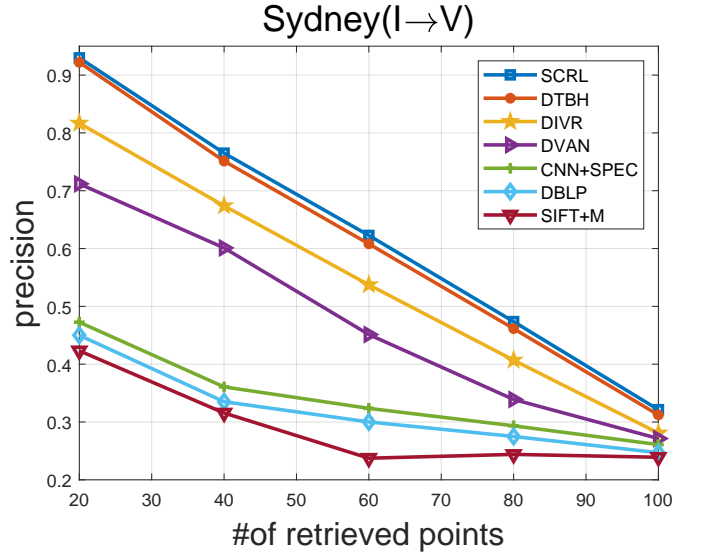

Fig. 5. The precision curves with different retrieved samples by the I→V protocol on the Sydney IV dataset.

consistency loss (SCRL-Pair), the proposed SCRL method can improve the mAP value from 75.80% to 94.24%, from 35.25% to 67.97%, and from 15.80% to 28.84% for the I→V protocol on Sydney, UCM and RSICD IV datasets, respectively. For the V→I protocol, the mAP value is also improved greatly by the proposed SCRL method compared with the SCRL-Pair method. The improvements demonstrate that the pairwise consistency loss is effective for learning semantics-consistent representations.

2) Intra-Modality Consistency: The intra-modality consistency loss aims to narrow the distance between two representations in the same modality to keep the semantic consistency of the same semantic concept. From the results in TABLE II-IV, we can intuitively find that the retrieval performance is greatly improved when the intra-modality consistency is taken into account. The specific comparison situation corresponds to the results of SCRL-Intra method and the proposed SCRL method in TABLE II-IV. The improvement is because the intra-modality consistency loss can constrain the model to shorten the distance between two representations from the same semantic concept in each modality.

3) Non-paired Inter-Modality Consistency: The non-paired inter-modality consistency aims to learn the relationship between two representations in different modalities to keep the semantic consistency across the two modalities. The specific comparison situation corresponds to the results of SCRL-Inter method and the proposed SCRL method in TABLE II-IV. From the results, we can notice the retrieval performance is significantly improved when the intra-modality consistency is considered. This is because the intra-modality consistency loss contributes to narrowing the distance between two representations from different modalities when they describe the same semantic concept.

4) Semantic Discrimination: The classification loss aims to enhance the semantic discrimination ability of representations so as to retrieve the relevant samples more easily. The specific comparison situation corresponds to the results
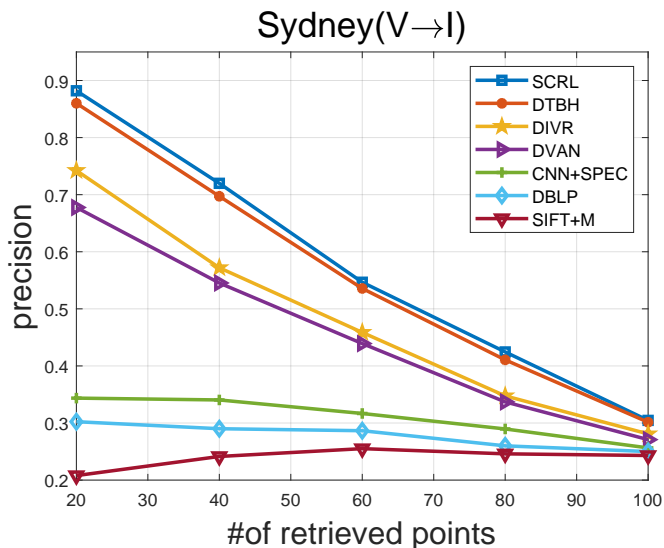
Fig. 6. The precision curves with different retrieved samples by the V→I protocol on the Sydney IV dataset.

Protocols	Methods	mAP	P@1	P@5	P@10
	SIFT+M [53]	8.55	4.56	4.65	4.56
	DBLP [54]	25.48	24.18	23.87	23.24
	CNN+SPEC [55]	26.25	29.50	25.52	23.65
	DVAN [11]	36.79	32.37	33.29	33.74
	CMIR-NET [8]	45.82	52.92	49.74	43.38
	DIVR [14]	50.94	59.34	54.17	50.12
I→V	DTBH [13]	64.24	73.10	69.69	65.63
	SCRL-Pair	35.25	35.24	35.48	35.50
	SCRL-Intra	52.64	59.90	57.10	54.19
	SCRL-Inter	53.62	61.95	56.33	52.17
	SCRL-Class	60.54	68.81	64.86	59.62
	SCRL-Dilation	65.46	72.57	70.43	67.86
	SCRL	**67.97**	**77.38**	**75.29**	**72.26**
	SIFT+M [53]	6.66	3.58	4.41	4.68
	DBLP [54]	19.33	17.12	17.62	16.31
	CNN+SPEC [55]	21.79	19.42	19.86	19.23
	DVAN [11]	32.28	32.37	33.91	34.34
	CMIR-NET [8]	40.37	46.74	43.75	39.62
	DIVR [14]	45.34	52.23	48.76	44.98
V→I	DTBH [13]	60.13	70.26	66.63	61.73
	SCRL-Pair	34.69	39.76	37.38	36.76
	SCRL-Intra	50.74	56.67	52.86	48.40
	SCRL-Inter	49.31	56.86	53.86	51.90
	SCRL-Class	59.82	66.90	59.05	55.76
	SCRL-Dilation	66.59	72.62	68.52	64.31
	SCRL	**66.83**	**84.05**	**77.14**	**74.07**

of SCRL-Class method and the proposed SCRL method in TABLE II-IV. From the comparison results, we can see the scores of all evaluation metrics are lifted when the classification loss is added. This proves the classification loss contributes to enhancing the semantic discrimination ability of representations.

5) Long Range Correlation: The dilation convolution is adopted to capture the long range correlation of each voice sample. The concrete comparison situation corresponds to the results of SCRL-Dilation method and the proposed SCRL method in TABLE II-IV. The results show the retrieval performance is slightly improved, which indicates the dilation convolution can indeed capture the long range correlation relationship within each voice sample by increasing the receptive field.

F. Results and Analysis

The experimental results and the corresponding analysis on three challenging RS image-voice datasets are given as follows.

1) Results on Sydney IV Dataset: TABLE II shows the comparison results between the proposed SCRL method and other compared methods on the Sydney IV dataset. Fig 5 and Fig 6 show the precision curves with different retrieved samples by the I→V and V→I protocol, respectively. By observing the results in TABLE II, Fig 5 and Fig 6, we can find: 1) the proposed SCRL method achieves the highest value in terms of most evaluation metrics. 2) Fig 5 and Fig 6 shows the proposed SCRL method outperforms other comparison methods at all returned neighbors. 3) As for the I→V protocol, the proposed SCRL method improves the mAP value from SIFT+M (31.67%), DBLP (44.38%), CNN+SPEC (46.67%), DVAN (71.77%), CMIR-NET (78.44%), DIVR (81.35%), DTBH (92.45%) to 94.24%. Meanwhile, for the V→I protocol, the proposed SCRL method improves the mAP value from SIFT+M (26.50%), DBLP (34.87%), CNN+SPEC (35.72%), DVAN (63.88%), CMIR-NET (71.28%), DIVR (75.97%),

DTBH (87.49%) to 93.53%. The improvements demonstrate that the modeling for the pairwise, intra-modality, and non-paired inter-modality relationships contributes to learning semantics-consistency representations across the two modalities, thereby facilitating cross-modal retrieval. In addition, the first column of Fig. 7 shows an example of the top five retrieved results by the proposed SCRL method with the I→V protocol. The first row of Fig. 8 shows an example of the top five retrieved results by the proposed SCRL method with the V→I protocol. As shown in the two figures, the proposed SCRL method can retrieve the relevant samples effectively, which further proves the effectiveness for modeling the semantics-consistent relationships comprehensively.

2) Results on UCM IV Dataset: TABLE III shows the comparison results between the proposed SCRL method and other compared methods on the UCM IV Dataset. Fig 9 and Fig 10 show the precision curves with different retrieved samples by the I→V and V→I protocol, respectively. Similar experimental results can be seen on the Sydney IV dataset. For instance, for the I→V protocol, the proposed SCRL method improves the mAP value from SIFT+M (31.67%), DBLP (44.38%), CNN+SPEC (46.67%), DVAN (71.77%), CMIR-NET (78.44%), DIVR (81.35%), DTBH (92.45%) to 67.97%. Meanwhile, for the V→I protocol, the proposed SCRL method improves the mAP value from SIFT+M (26.50%), DBLP (34.87%), CNN+SPEC (35.72%), DVAN (63.88%), CMIR-NET (71.28%), DIVR (75.97%), DTBH (87.49%) to 66.83%. This is because the exploration for the semantics-consistent
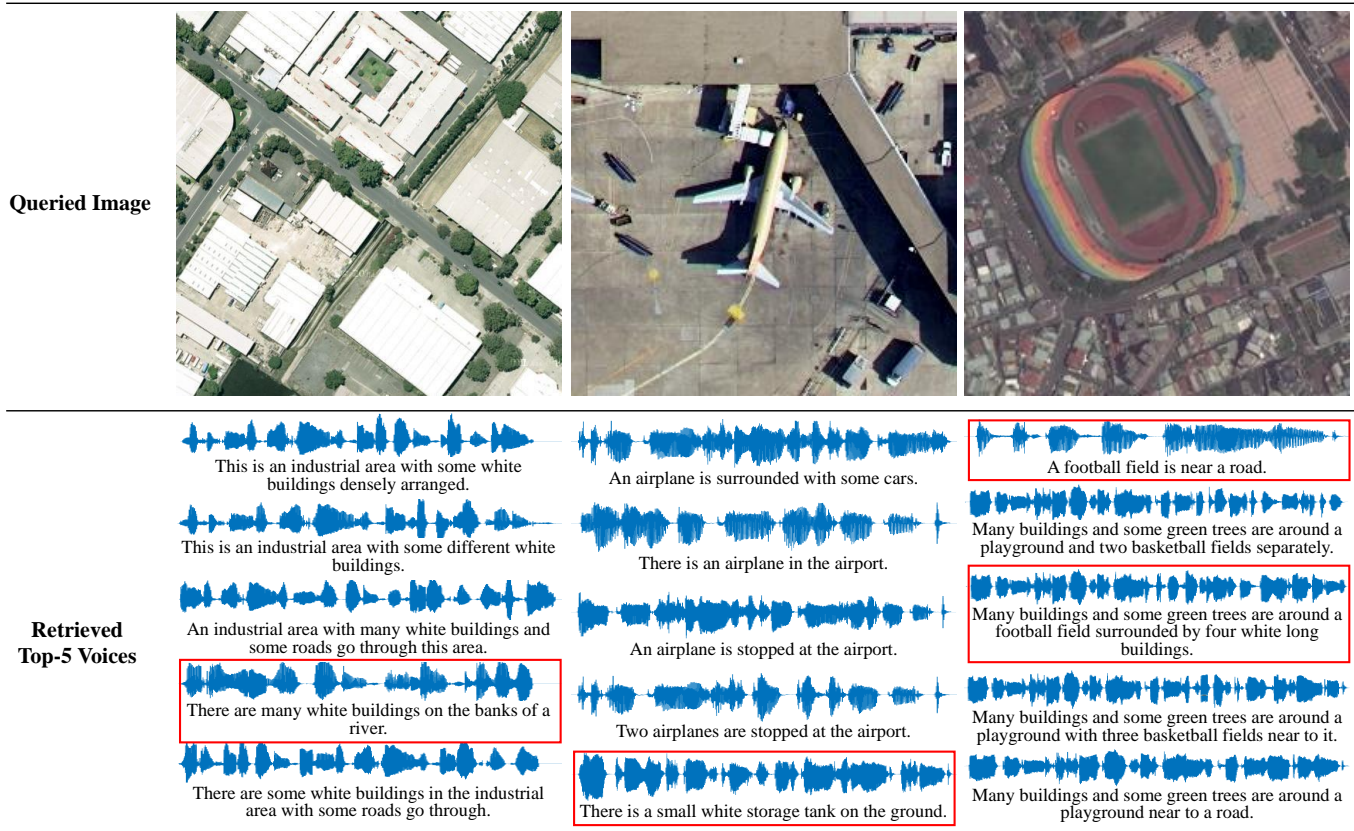
Fig. 7. Some examples of the I→V retrieval results by the proposed SCRL method on Sydney, UCM and RSICD IV datasets. The result examples on the three datasets correspond to the first column, second column and third column, respectively. The wrong retrieval results are marked with red boxes.
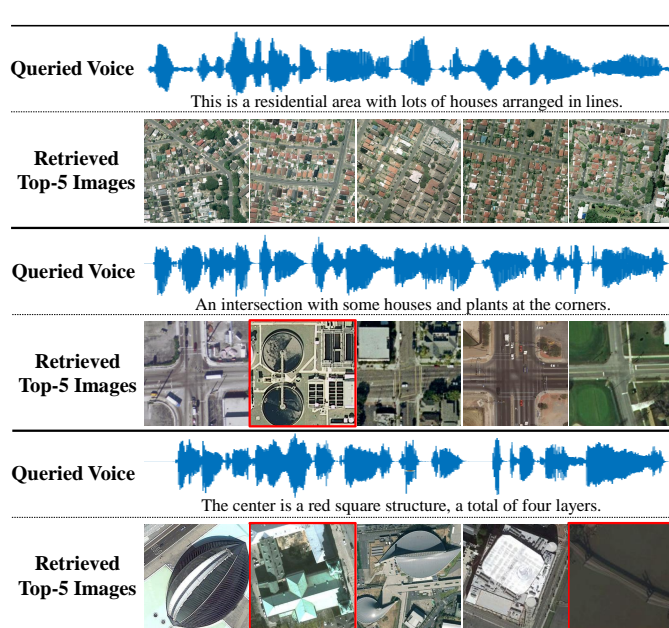

Fig. 8. Some examples of the V→I retrieval results by the proposed SCRL method on Sydney, UCM and RSICD IV datasets. The result examples on the three datasets correspond to the first row, second row and third row, respectively. The wrong retrieval results are marked with red boxes.
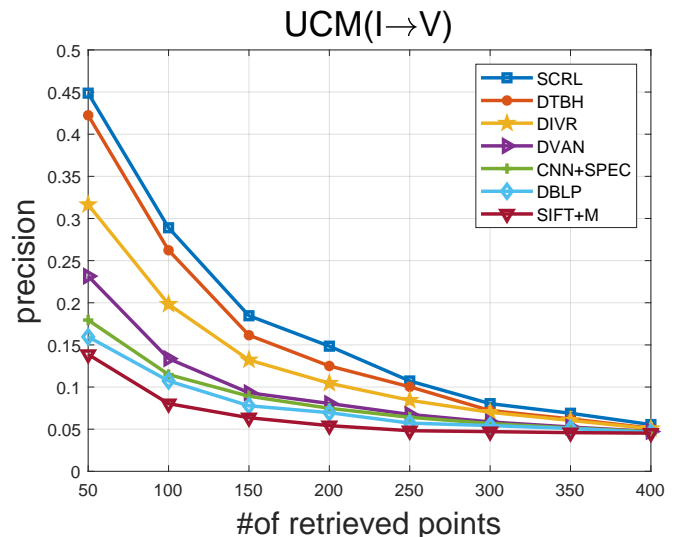

Fig. 9. The precision curves with different retrieved samples by the I→V protocol on the UCM IV dataset.

representation space can effectively promote the model to narrow the heterogeneous semantic gap across the two modalities. In addition, the second column of Fig. 7 shows an example of the top five retrieved results by the proposed SCRL method with the I→V protocol. The second row of Fig. 8 shows an example of the top five retrieved results by the proposed
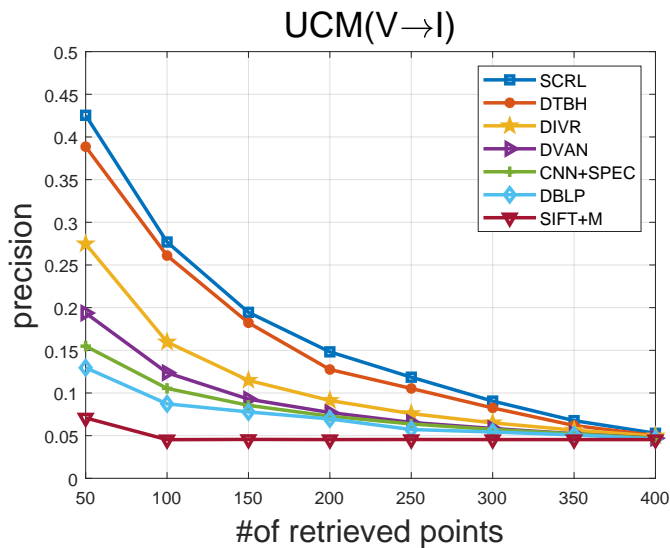
UCM(V→I)

Fig. 10. The precision curves with different retrieved samples by the V→I protocol on the UCM IV dataset.

SCRL method with the V→I protocol. The results show that the proposed SCRL method can effectively retrieve relevant samples, which further proves the effectiveness of considering the pairwise, intra-modality, and non-paired inter-modality relationships comprehensively.

3) Results on RSICD IV Dataset: TABLE IV shows the comparison results between the proposed SCRL method and other compared methods on the most challenging RSICD IV Dataset. Fig 11 and Fig 12 show the precision curves with different retrieved samples by the I→V and V→I protocol, respectively. Although the RSICD IV dataset is more challenging [14], the proposed SCRL method can improve the retrieval performance to a great extent. For example, the SCRL method lifts the mAP value from 23.46% to 28.84% for the I→V protocol compared with the state-of-the-art DTBH method. Meanwhile, the SCRL method improves the mAP value from 23.46% (DTBH) to 28.84% for the V→I protocol. This demonstrates the high efficiency of the proposed SCRL method for learning semantics-consistent representations across the two modalities, because the semantics-consistent relationships, including pairwise, intra-modality, and non-paired inter-modality relationships, are modeled comprehensively. In addition, some examples of the top five retrieved results by the proposed SCRL method are shown in Fig. 7 (3rd column) and Fig. 8 (3rd row). It can be seen that the proposed SCRL method can retrieve diverse but relevant samples effectively, even on very challenging image-voice dataset, which further verifies the effectiveness of simultaneously modeling various semantics-consistent relationships.

V. Conclusions

In this paper, a semantics-consistent representation learning (SCRL) method is proposed for the task of remote sensing (RS) image-voice retrieval. The main novelty is that the proposed method takes the pairwise, intra-modality, and non-paired inter-modality relationships into account simultaneously, thereby improving the semantic consistency of

TABLE IV
THE COMPARISON RESULTS BETWEEN THE PROPOSED SCRL METHOD AND OTHER METHIDS ON THE RSICD IV DATASET (%).

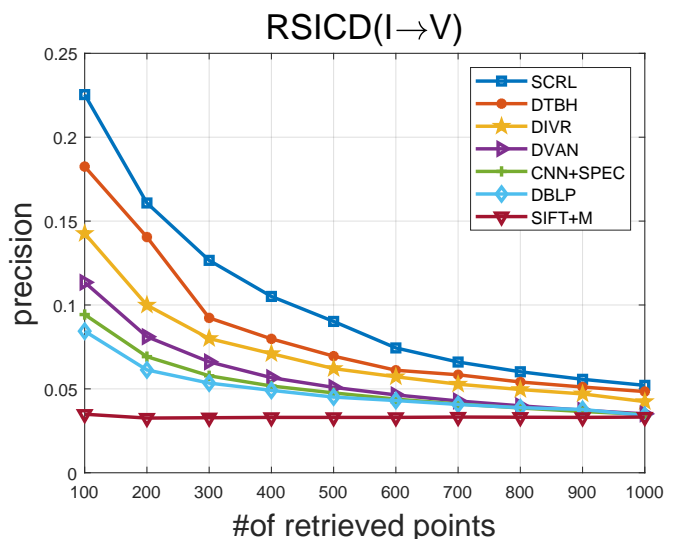| Protocols	Methods	mAP	P@1	P@5	P@10
	SIFT+M [53]	5.04	6.22	5.34	4.50
	DBLP [54]	12.70	15.32	15.21	14.22
	CNN+SPEC [55]	13.24	16.82	16.62	15.69
	DVAN [11]	16.29	22.49	22.56	21.70
	CMIR-NET [8]	17.78	24.11	23.52	22.54
	DIVR [14]	19.62	25.43	24.84	24.20
I→V	DTBH [13]	23.46	27.58	26.84	25.49
	SCRL-Pair	15.80	17.09	16.84	16.05
	SCRL-Intra	20.06	21.78	21.34	20.75
	SCRL-Inter	20.02	22.46	22.44	21.73
	SCRL-Class	25.92	29.54	28.35	27.00
	SCRL-Dilation	27.54	30.88	29.96	29.81
	SCRL	**28.84**	**31.10**	**29.81**	**28.34**
	SIFT+M [53]	4.85	3.66	3.60	3.541
	DBLP [54]	8.14	6.21	6.08	6.76
	CNN+SPEC [55]	9.96	7.13	7.00	7.44
	DVAN [11]	15.71	16.18	15.10	14.76
	CMIR-NET [8]	17.25	17.94	16.58	15.36
	DIVR [14]	18.58	19.76	18.31	17.59
V→I	DTBH [13]	22.72	23.30	22.48	21.17
	SCRL-Pair	15.35	18.56	17.31	16.35
	SCRL-Intra	24.52	30.11	29.08	28.08
	SCRL-Inter	25.21	30.60	30.72	29.64
	SCRL-Class	28.33	31.73	31.57	30.83
	SCRL-Dilation	29.55	32.59	31.25	30.08
	SCRL	**31.26**	**33.76**	**33.01**	**32.01**

RSICD(I→V)

Fig. 11. The precision curves with different retrieved samples by the I→V protocol on the RSICD IV dataset.

RSICD(V→I)

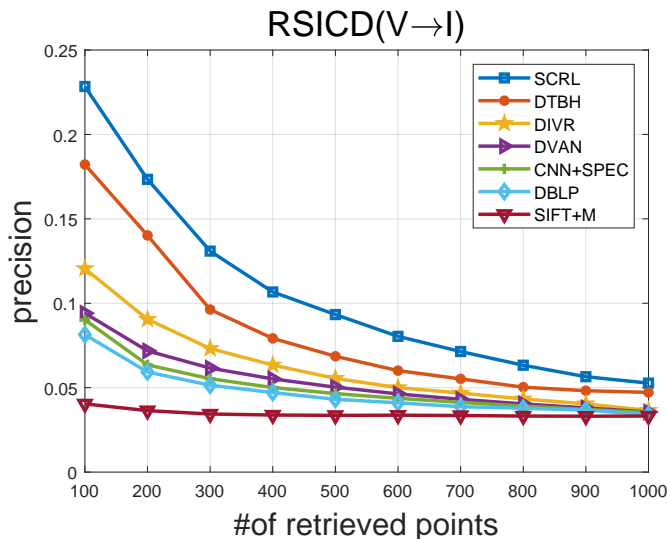
Fig. 12. The precision curves with different retrieved samples by the V→I protocol on the RSICD IV dataset.

the learned representations for the RS image-voice retrieval. Specifically, a consistent representation space is explored with a joint loss function comprehensively considering the three kinds of relationships for the extracted representations. By exploring the consistent representation space, the heterogeneous semantic gap can be mitigated to a great extent and semantics-consistent representations can be learned for the RS image-voice retrieval. Experimental results on Sydney, UCM and RSICD image-voice datasets demonstrate that the proposed joint loss is effective to mine the pairwise, intra-modality, and non-paired inter-modality relationships for better RS image-voice retrieval. In addition, the proposed SCRL method achieves better performance compared with other state-of-the-art methods, manifesting its superiority. In the future, we will explore how to extend the proposed SCRL method to address the large-scale image-audio retrieval cases such as embedding the hashing code learning to our work.

REFERENCES

[1] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4558–4572, 2020.

[2] 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 Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6521–6536, 2018.

[3] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5535–5548, 2019.

[4] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 950–965, 2018.

[5] B. Demir and L. Bruzzone, "Hashing-based scalable remote sensing image search and retrieval in large archives," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 2, pp. 892–904, 2016.

[6] Y. Chen and X. Lu, "Supervised deep hashing with a joint deep network," *Pattern Recognition*, vol. 105, p. 107368, 2020.

[7] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 950–965, 2018.

[8] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "Cmir-net : A deep learning based model for cross-modal retrieval in remote sensing," *Pattern Recognition Letters*, vol. 131, pp. 456–462, 2020.

[9] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, 2017.

[10] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 318–328, 2020.

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

[12] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 11, pp. 4644–4654, 2019.

[13] Y. Chen and X. Lu, "A deep hashing technique for remote sensing image-sound retrieval," *Remote Sensing*, vol. 12, no. 1, p. 84, 2019.

[14] Y. Chen, X. Lu, and S. Wang, "Deep cross-modal image-voice retrieval in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7049–7061, 2020.

[15] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 318–328, 2020.

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

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[18] R. Hidayat, A. Bejo, S. Sumaryono, and A. Winursito, "Denoising speech for mfcc feature extraction using wavelet transformation in speech recognition system," in *2018 10th International Conference on Information Technology and Electrical Engineering*, 2018, pp. 280–284.

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

[20] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 818–832, 2013.

[21] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 3023–3034, 2014.

[22] H. Ning, X. Zheng, Y. Yuan, and X. Lu, "Audio description from image by modal translation network," *Neurocomputing*, 2020.

[23] B. Zhao, L. Hua, X. Li, X. Lu, and Z. Wang, "Weather recognition via classification labels and weather-cue maps," *Pattern Recognition*, vol. 95, pp. 272–284, 2019.

[24] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Remote sensing image scene classification using rearranged local features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1779–1792, 2018.

[25] B. Zhao, X. Li, and X. Lu, "Cam-rnn: Co-attention model based rnn for video captioning," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5552–5565, 2019.

[26] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Spatial structure preserving feature pyramid network for semantic image segmentation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 3, pp. 73:1–73:19, 2019.

[27] X. Tang, L. Jiao, W. J. Emery, F. Liu, and D. Zhang, "Two-stage reranking for remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5798–5817, 2017.

[28] 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 Geoscience and Remote Sensing Letters*, vol. 15, no. 10, pp. 1535–1539, 2018.

[29] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, 2017.

[30] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multi-label remote-sensing image classification and retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[31] Y. Liu, L. Ding, C. Chen, and Y. Liu, "Similarity-based unsupervised deep transfer learning for remote sensing image retrieval," *IEEE Trans-

actions on Geoscience and Remote Sensing, vol. 58, no. 11, pp. 7872–7889, 2020.

[32] F. Ye, W. Luo, M. Dong, H. He, and W. Min, "Sar image retrieval based on unsupervised domain adaptation and clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 9, pp. 1482–1486, 2019.

[33] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen, "Unified binary generative adversarial network for image retrieval and compression," *International Journal of Computer Vision*, vol. 128, pp. 2243—22 646, 2020.

[34] W. Xiong, Z. Xiong, Y. Cui, and Y. Lv, "A discriminative distillation network for cross-source remote sensing image retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1234–1247, 2020.

[35] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5284–5296, 2020.

[36] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, pp. 785–796.

[37] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *2017 IEEE International Conference on Computer Vision*, 2017, pp. 3456–3465.

[38] X. Lu, Y. Chen, and X. Li, "Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 106–120, 2018.

[39] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4109–4118.

[40] Z. Chen, T. Zhang, and C. Ouyang, "End-to-end airplane detection using transfer learning in remote sensing images," *Remote Sensing*, vol. 10, no. 1, p. 139, 2018.

[41] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[42] A. Chowdhury and A. Ross, "Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely degraded audio signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616–1629, 2020.

[43] Y. Chen and X. Lu, "Deep category-level and regularized hashing with global semantic similarity learning," *IEEE Transactions on Cybernetics*, 2020.

[44] W. Xiong, Y. Lv, X. Zhang, and Y. Cui, "Learning to translate for cross-source remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4860–4874, 2020.

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

[46] Y. Yuan, H. Ning, and X. Lu, "Bio-inspired representation learning for visual attention prediction," *IEEE Transactions on Cybernetics*, pp. 1–14, 2019.

[47] A. Yuan, X. Li, and X. Lu, "3g structure for image caption generation," *Neurocomputing*, vol. 330, pp. 17–28, 2019.

[48] T. Zhi, L. Duan, Y. Wang, and T. Huang, "Two-stage pooling of deep convolutional features for image retrieval," in *2016 IEEE International Conference on Image Processing*, 2016, pp. 2465–2469.

[49] D. Hu, F. Nie, and X. Li, "Deep binary reconstruction for cross-modal hashing," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 973–985, 2019.

[50] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 4062–4076, 2019.

[51] N. Khurshid, M. Tharani, M. Taj, and F. Z. Qureshi, "A residual-dyad encoder discriminator network for remote sensing image matching," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2001–2014, 2020.

[52] P. Li and P. Ren, "Partial randomness hashing for large-scale remote sensing image retrieval," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 3, pp. 464–468, 2017.

[53] X. Zhao, Z. Li, and J. Yi, "Sift feature-based second-order image hash retrieval approach," *Journal of Software*, vol. 13, no. 1, pp. 103–116, 2018.

[54] D. F. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems*, 2016, pp. 1858–1866.

[55] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *2017 IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[56] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 154–162.