Text-Image Matching for Cross-Modal Remote Sensing Image Retrieval via Graph Neural Network

Hongfeng Yu, Fanglong Yao, Wanxuan Lu, Nayu Liu, Peiguang Li, Hongjian You, and Xian Sun, *Senior Member, IEEE*

***Abstract*—The rapid development of remote sensing (RS) technology has produced massive images, which makes it difficult to obtain interpretation results by manual screening. Therefore, researchers began to develop automatic retrieval method of RS images. In recent years, cross-modal RS image retrieval based on query text has attracted many researchers because of its flexible and has become a new research trend. However, the primary problem faced is that the information of query text and RS image is not aligned. For example, RS images often have the attributes of multiscale and multiobjective, and the amount of information is rich, while the query text contains only a few words, and the information is scarce. Recently, graph neural network (GNN) has shown its potential in many tasks with its powerful feature representation ability. Therefore, based on GNN, this article proposes a new cross-modal RS feature matching network, which can avoid the degradation of retrieval performance caused by information misalignment by learning the feature interaction in query text and RS image, respectively, and modeling the feature association between the two modes. Specifically, to fuse the within-modal features, the text and RS image graph modules are designed based on GNN. In addition, in order to effectively match the query text and RS image, combined with the multihead attention mechanism, an image-text association module is constructed to focus on the parts related to RS image in the text. The experiments on two public standard datasets verify the competitive performance of the proposed model.**

***Index Terms*—Cross-modal feature fusion, cross-modal remote sensing (RS) image retrieval, graph neural network (GNN).**

I. Introduction

The interpretation of remote sensing (RS) images provides a powerful support for the monitoring of natural resources, natural disaster, and urban development, etc. Recently, due to the increasing maturity of RS image acquisition technology, the number of RS image acquisition is also increasing rapidly. How to quickly and accurately find the required data from the massive RS images is the key to improve the efficiency of data application. The traditional RS image retrieval methods are mainly through manual screening, but when facing a huge number of RS images, these methods falls into a bottleneck. Therefore, in order to efficiently organize and manage the massive RS images, researchers began to explore automatic RS image retrieval methods, and gradually formed a research trend in the field of RS.

The query and retrieval data belong to the same mode is called single-modal RS image retrieval. Chen et al. [1] proposed a hash algorithm to improve the efficiency of retrieval. Based on the abovementioned method, Demir et al. [2] proposed a dual-core nonlinear hash algorithm with higher accuracy and faster efficiency. RS image retrieval based on single-modal has shown a good performance in coarse-grained RS image retrieval, but it is still difficult to cope with object-level RS image retrieval.

In recent years, cross-modal RS image retrieval has slowly become a research hotspot due to the advantages of form flexibility. Guo et al. [3] are committed to the research of audio-based RS image retrieval, using convolutional neural networks (CNNs) and AudioNet to extract the features of audio and image patterns, respectively, and calculate the matching degree of cross-modalities. Afterwards, some RS image retrieval methods based on text have become mainstream [4], [5], [6]. These methods basically follow a two-stage retrieval process, i.e., first generate a text semantic description using the RS image caption method, and then retrieve the RS image by measuring how well the query text matches the caption. However, the two-stage retrieval method is essentially a text-to-text retrieval, which not only ignores the direct matching between the RS image and the text, but also easily affects the quality of the semantic description. In order to avoid the disadvantages of two-stage retrieval, Yuan et al. [7] proposed a one-stage retrieval method that directly learns how well the query text matches the RS image. More importantly, the method also provides keywords to complement the query text, improving the retrieval capacity of remotely sensed images.

Although the abovementioned cross-modal RS image retrieval methods have promoted the development of this field, it still encounters difficulties in solving the problem of cross-modal information misalignment, resulting in the decline of the RS image retrieval performance. In addition to having multiscale and multitarget properties, RS images have a lot of background

Manuscript received 23 July 2022; revised 20 September 2022 and 30 November 2022; accepted 19 December 2022. Date of publication 26 December 2022; date of current version 2 January 2023. *(Corresponding author: Fanglong Yao.)*

Hongfeng Yu, Fanglong Yao, Nayu Liu, Peiguang Li, Hongjian You, and Xian Sun are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Network Information System Technology, (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yuhf@aircas.ac.cn; yaofanglong17@mails.ucas.ac.cn; 695704204@qq.com; lipeiguang17@mails.ucas.ac.cn; hjyou@mail.ie.ac.cn; sunxian@aircas.ac.cn).

Wanxuan Lu is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: luwx@aircas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3231851

noise. However, the query text often contains only a few words and is poorly informative. Furthermore, when extracting modal features of text and images, previous methods ignore the interaction between features within the modality, such as information fusion between words in text, information interaction between target features in images, etc.

Recently, the graph neural network (GNN) has shown its potential in a variety of tasks with its powerful feature representation ability. Therefore, based on GNN, this article proposes a new cross-modal RS feature matching network, named as CMFM-Net, which can avoid the degradation of the retrieval performance caused by information misalignment by learning the feature interaction in query text and RS image, respectively, and modeling the feature association between the two modes. Specifically, in order to fuse the within-modal features, the text and RS image graph modules are designed based on GNN. In addition, in order to effectively match the query text and RS image, combined with the multihead attention mechanism, an image-text association module is constructed to focus on the parts related to RS image in the text.

In summary, the main contributions of this article are as follows.

1) In this article, a cross-modal RS feature matching network is proposed, which solves the problem of information misalignment of cross-modal in RS image retrieval through feature interaction within modality and feature association between modalities.
2) Based on the GNN, the text and RS image graph modules are constructed to ensure the interactive fusion of features in the modality. In addition, the image-text association module is designed to learn the associations between modalities and highlights some words in the text associated with the RS image.
3) Qualitative and quantitative experimental results on the two public datasets of RSICD and RSITMD verify the effectiveness of CMFM-Net in cross-modal RS image retrieval.

The rest of this article is organized as follows. In Section II, the related works of cross-modal RS image retrieval are briefly introduced. In Section III, we illustrate the implementation details of the proposed CMFM-Net model. In Sections IV and V, we give a quantitative comparison between the proposed model and other baselines. Finally, Section VI concludes this article.

II. Related Work

In this section, we review the previous research on cross-modal retrieval of nature scenes and RS images. Then, we turn to graph-based approach that is more relevant to our work.

A. Cross-Modal Retrieval

Cross-modal retrieval focuses on enabling flexible retrieval of different modalities (e.g., text-image, speech-image) [8]. It performs retrieval of related data of one type by referring to the data of another type as a query. It aims to learn a common global representation space in which similarity between data of different modalities can be directly measured by commonly used distance measures, such as Euclidean distance and cosine distance. According to the difference of basic methods, the existing methods can be divided into traditional methods and deep learning-based methods.

The early traditional approaches consider the relationship between different modal data to be linear and represent the data in jointly by a linear projection. One method is canonical correlation analysis (CCA) [9] that learns the linear projection matrix by maximizing the overall correlation between cross-modal data. and there is extensive work here based on CCA [10], [11], [12]. And the other traditional method is crossmodal factor analysis [13], which minimizes the Frobenius norm between common representations of differnet modalities by directly. In addition, there are the local group based prior knowledge [14], [15], [16], semantic hierarchy [17], [18], [19], etc. Recently, deep learning-based methods have brought great improvements in cross-modal retrieval. These methods can better learn the nonlinear relationships between different modal data. Deep canonical correlation analysis [20] is a nonlinear extension of CCA that learns complex nonlinear transformations of two modalities by jointly learning the parameters of both transformations to maximize the total correlation. Sharama et al. [21] propose generalized multiview analysis by supervised use of semantic category labels to solve a joint, relaxed quadratic constrained quadratic program over different feature spaces. It is a supervised extension of CCA, and the method is extremely robust and has the ability to replace CCA whenever classification or retrieval purposes and labeling information are available.

B. Cross-Modal Retrieval of RS

Because of large scale, complex background, and large aspect ratios, the cross-modal retrieval of RS scenes is more complicated compared with natural scenes. Li et al. [22] design source-invariant deep hashing convolutional neural networks for image–image cross-modal RS image retrieval. In [23], a novel deep cross-modality hashing network is proposed for cross-modality retrieval between synthetic aperture radar (SAR) and optical images. As the image caption has developed, caption-based retrieval methods have become the mainstream of RS image retrieval [24], [25], [26]. Li et al. [27] introduced a multilevel attention model that combined with beam search to generate multiple captions and then select the best caption by using the best prior knowledge. In [28], the authors propose a fine-grained and structured attention-based method, which can utilize the structural characteristics of semantic contents in RS images. For audio-image cross-modal RS image retrieval, Mao et al. [29] design a deep visual-audio network to learn the correspondence of RS image and audio. Chen et al. [5] produce a deep triplet-based hashing to integrate relative semantic similarity relation learning and hash code learning into an end-to-end network. In [5], the authors design a cross-modal RS image-voice retrieval approach

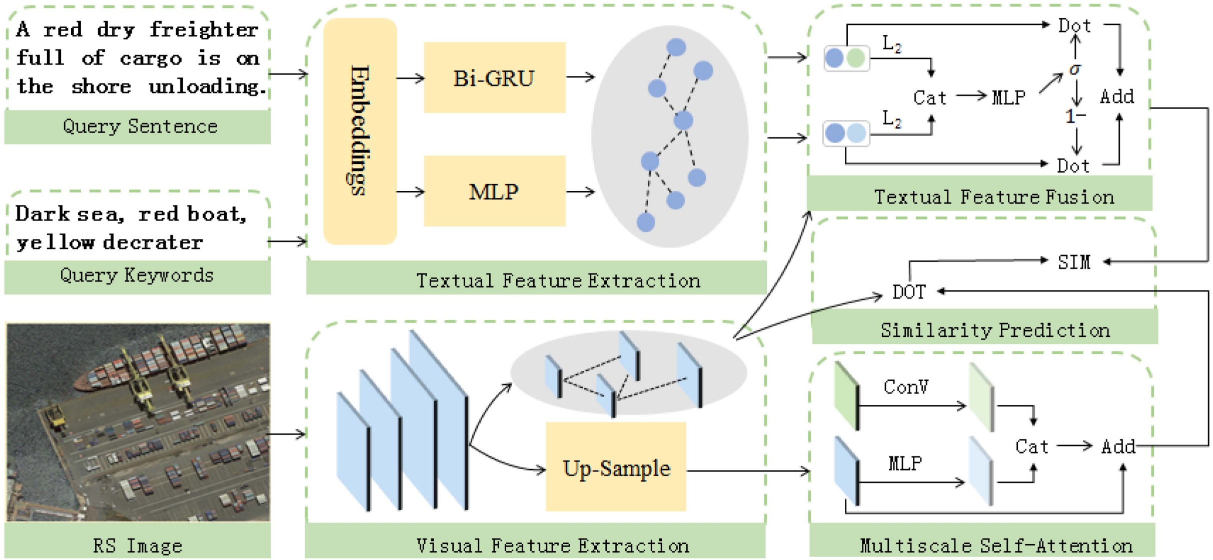

A red dry freighter full of cargo is on the shore unloading.
Query Sentence
Dark sea, red boat, yellow decrater
Query Keywords
RS Image
Embeddings
Bi-GRU
MLP
Textual Feature Extraction
Up-Sample
Visual Feature Extraction
L2
L2
Cat
MLP
σ
1-
Dot
Dot
Add
Textual Feature Fusion
SIM
DOT
Similarity Prediction
ConV
MLP
Cat
Add
Multiscale Self-Attention

Fig. 1. Overview of the proposed CMFM-Net. Given the visual input (RS image) and textual input (query sentence and keywords), we first perform the feature extraction for each modal, then we integrate the multiscale visual features and fuse the textual features under the guidance of visual information, in the last, we calculate the similarity between final textual and visual features.

called DIVR, to capture more feature about the RS data and generate hash codes with fast retrieval characteristics and low memory.

C. Graph-Based Learning

Graph is widely used for various tasks, such as machine translation [31], [32], [33], semantic segmentation [34], [35], [36], classification [37], [38], [39], object detection [40], [41], and image retrieval [42], [43], [44]. Graph-based learning is powerful tools for graph data representation, and it has a wide range of applications as an emerging approach in RS. For instance, Qin et al. [38] proposed a novel semisupervised learning framework that is based on spectral–spatial graph convolutional networks (GCNs) to alleviate deficient labeled pixels problem for hyperspectral image classification task. Ren et al. [39] exploit the rich spatial and spectral information contained in superpixel-based weighted graph structure and propose a semisupervised approach to achieve robust classification of PolSAR images on irregular domains. Saha et al. [37] propose a GNN-based multitarget domain adaptation method for RS field, which incrementally adapt a source-trained classifier for multiple targets. GNN is also widely adopted in cross-modal retrieval tasks. Dong et al. [45] exploit a GCN to reconstruct a sample representation based on the neighborhood relationship between the sample itself and its neighbors, construct a local graph for each instance to reconstruct node features based on the local graph, and project the features of both modalities to a common space. In [46], the authors design an end-to-end adaptive label-aware graph convolutional network, which obtains modality-invariant and discriminative representations for efficiently explores and cross-modal retrieval. In addition, it efficiently explores and preserves the semantic structure of labels in a data-driven manner. To solve suffering from the lack of reliable learning guidance and cross-modal discrepancy, Zhang et al. [47] introduced an aggregation-based graph convolutional hashing to exploit the structural feature of multiple modalities from different perspectives. In [48], as for multimodal image retrieval task, the authors build a GNN for similar pictures and their tags to learn visual-semantic representations. The difference between the method in [48] and our proposed method is that this method is oriented to CBIR task, that is, content-based image retrieval in natural scenes. While our method is oriented to TBRSIR task, that is, text-based RS image retrieval. In addition, the node on the graph in this method represents a complete image or image tag, while the node in our method refers to the features map.

III. Method

In this section, we introduce the details about our proposed CMFM-Net. As depicted in Fig. 1, the CMFM-Net consists of visual and textual feature extraction modules, a multiscale self-attention module, a textual feature fusion module, and the final prediction module. In the following, we first define the task in the formulation and provide an overview of the CMFM-Net. Then, we introduce the core modules in detail.

A. Problem Description

Given the RS image I, query sentence T, and query keywords K, we first extract their features and map them into a common d-dimensional space, $\widehat{v} \in \mathbb{R}^d$, $\widehat{t} \in \mathbb{R}^d$, and $\widehat{k} \in \mathbb{R}^d$, respectively. Then, after a series of feature fusion, especially $\widehat{t}$ and $\widehat{k}$, we obtain a score s that measures the cross-modal similarity

$$s = \sigma\left(\mathcal{F}_{\widehat{v}}\widehat{v} \odot \mathcal{N}\left(\mathcal{F}_{\widehat{t}}\widehat{t}, \mathcal{F}_{\widehat{k}}\widehat{k}\right)\right) \tag{1}$$

where $\mathcal{F}_{\widehat{v}}$, $\mathcal{F}_{\widehat{t}}$, and $\mathcal{F}_{\widehat{k}}$ indicate the feature processor for visual image, textual sentence, and textual keywords, respectively, and

$\mathcal{N}$ denotes the fusion of keywords and sentence, and σ indicate Sigmoid function. Notably, since keywords are utilized to locate the expected objects in images, it supplement the sentence retrieval that aims to capture the associated target. Thus, we adjust the model to take the sentence and keywords as textual inputs by weighted sum them

$$\mathcal{N}\left(\mathcal{F}_{\widehat{t}}\widehat{t}, \mathcal{F}_{\widehat{k}}\widehat{k}\right) = \lambda_1 \cdot \mathcal{F}_{\widehat{t}}\widehat{t} + \lambda_2 \cdot \mathcal{F}_{\widehat{k}}\widehat{k}. \tag{2}$$

Furthermore, to promote the feature fusion between modalities, we introduce a cross-modal interaction strategy. Specifically, considering that there are a large number of targets in the RS image and scale differences between various targets, thus, we adopt to use visual features to guide the textual features extraction, which is formulated as

$$\phi\left(\mathcal{F}_{\widehat{v}}\widehat{v}, \mathcal{N}\left(\mathcal{F}_{\widehat{t}}\widehat{t}, \mathcal{F}_{\widehat{k}}\widehat{k}\right)\right) = \mathcal{F}_{\widehat{v}}\widehat{v} \to \mathcal{N}\left(\mathcal{F}_{\widehat{t}}\widehat{t}, \mathcal{F}_{\widehat{k}}\widehat{k}\right). \tag{3}$$

In summary of above, formula 1 eventually evolve into the following formulation:

$$s = \sigma\left(\mathcal{F}_{\widehat{v}}\widehat{v} \odot \phi\left(\mathcal{F}_{\widehat{v}}\widehat{v}, \mathcal{N}\left(\mathcal{F}_{\widehat{t}}\widehat{t}, \mathcal{F}_{\widehat{k}}\widehat{k}\right)\right)\right). \tag{4}$$

B. Method Overview

As mentioned above, our method takes RS image, query sentence, and query keywords as inputs, thus, we perform feature extraction for each type of information.

1) RS Image: Following the method in [49], we utilize the ResNet-18 model pretrained on the ImageNet dataset [50], [51], [52], to extract the visual features from RS image. Specifically, we first select the last convolution layer's feature map with size $512 \times 8 \times 8$ as the features of the whole image and transform it into a feature sequence $V = \{v_1, v_2, \ldots, v_{64}\}$, each element $v_i \in \mathcal{R}^{512}$ refers to a feature map. To prompt the information extrange among features maps, we introduce the image graph module.

Concretely, we take the feature maps as nodes and construct a graph, then we adopt the multihead attention [53] mechanism to perform features transfer. The multihead attention is designed to capture the correlation among items from diverse H subspaces, and then the feed-forward module supplements the nonlinearity to features in the attention results. Finally, layer normalization [54] with residual connection [55] is applied to normalize representation.

By stacking L layers of image graph module, we obtain the global information feature vector of the RS image image by average pooling

$$\text{image} = \frac{1}{64}\sum_{i=1}^{64} V_i^L. \tag{5}$$

Besides, although the features image contains important information in the RS image, there are apparent drawbacks for only using the global features. For one thing, the RS image contains a large number of targets compared with the natural image. Thus, using global features image merely for subsequent retrieval will cause information redundancy, which is unable to obtain favorable representations for the image. For another thing, RS images have multiscale characteristics, and single-scale features cannot represent the image well. As the increase of convolution layer, the small targets will be filtered out by the pooling layer, which means that these small targets will not appear in the global features. The deeper features map with a larger scale can capture high-level semantic information of salient objects, whereas the shallower feature map with a smaller scale can extract fine features information. Considering the abovementioned factors, we further extract features from each layer of the convolution network.

After the extraction, we upsample the feature maps of the first three layers and then concatenate them together as the low-level image features. Subsequently, the feature maps of the last two layers are sampled and connected as the high-level image features. The abovementioned process can be represented as

$$\left\{v^{(m)}\right\}_{m=1}^{5}, v^{(g)} = \text{CNN}\left(I, \theta_I\right) \tag{6}$$

$$\left\{F_m\right\}_{m=1}^{5} = \text{Upsample}\left(\left\{v^{(m)}\right\}_{m=1}^{5}\right) \tag{7}$$

$$v^{(l)} = \text{Cat}\left(F_1, F_2, F_3\right) \tag{8}$$

$$v^{(h)} = \text{Cat}\left(F_4, F_5\right) \tag{9}$$

where $v^{(m)}$ is each layer's feature map output by convolution network; the upsample operation denotes upsampled layer, which is used to match different feature maps to the same size. F_m are the feature maps obtained after $v^{(m)}$ upsampled, where all feature maps of the low-level features have the same size, and the same for the high-level features. Cat(x, y) is used to represent the channelwise concatenation of the feature vectors x and y with consistent dimensions. $v^{(l)}$ and $v^{(h)}$ represent the low-level features and high-level features extracted from RS image. Finally, we use the low-level features $v^{(l)}$, high-level features $v^{(h)}$, and global features image as the final image features.

2) Query Sentence: The query sentence $s = \{w_1, w_2, \ldots, w_n\}$ is denoted as a sequence of words, where n indicates its length and w_i refers to the index of the corresponding word in the vocabulary $\mathcal{W}$. Given the word embedding matrix $\mathbf{M}_w \in R^{|\mathcal{W}| \times d}$, we map each word w_i into a d-dimensional vector e_i and obtain the vector sequence $\mathbf{E}_s \in \mathcal{R}^{n \times d}$. Then, a bidirectional GRU (Bi-GRU) [56] is employed as sentence encoder to learn the representation of word $h_i \in \mathcal{R}^d$ in its context

$$h_i = \text{BiGRU}\left(e_i, h_{i-1}\right). \tag{10}$$

Based on the encoded representation of sequence $H = \{h_1, h_2, \ldots, h_n\}$, we introduce the sentence graph module to perform further enhancement.

Specifically, we construct a fully connected graph where each node indicates a word. We use h_i^l represent the states of ith node at layer l, and initialize the state of each node h_i^0 by the abovementioned encoded representation: $h_i^0 = h_i$. Since there is lack of prior connection between words, we adopt a multilayer perceptron (MLP) to compute the edge weight $c_{i,j}$ between node h_i and its neighbor $h_j \in H$, which is formulated as

$$c_{i,j} = \mathbf{W}_1^{l-1}\left(\text{ReLU}\left(\mathbf{W}_0^{l-1}\left(h_i^{l-1} \| h_j^{l-1}\right)\right)\right) \tag{11}$$

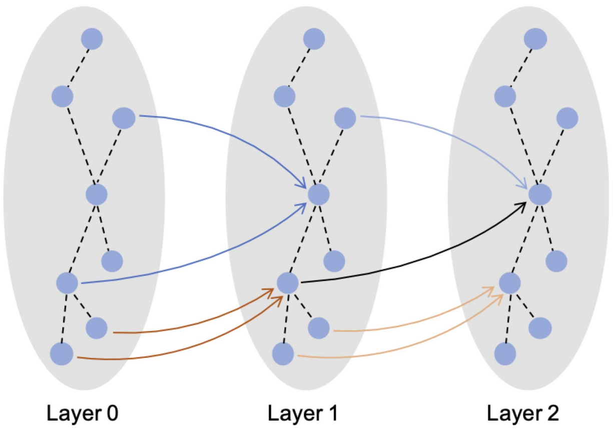


Fig. 2. Sentence graph updating.

where $\mathbf{W}_0^{l-1} \in \mathcal{R}^{2*d}$ and $\mathbf{W}_1^{l-1} \in \mathcal{R}^{d}$ are learnable matrices, and $\cdot \| \cdot$ denotes the concatenation operation. Then, the edge weights that normalized by softmax function are used to aggregate features from the connected nodes, thus, we obtain the representation for node h_i^l at layer l

$$h_i^l = \sum_j^n \frac{\exp\left(c_{i,j}\right)}{\sum_k^n \exp\left(c_{i,k}\right)} \cdot h_j^{l-1}. \tag{12}$$

As shown in Fig. 2, by stacking L layers of sentence graph module, we assume that each node could grasp enough information by communicating with others. As a result, the global information feature vector of the sentence can be represented by sent

$$\text{sent} = \frac{1}{n} \sum_{i=1}^{n} h_i^L. \tag{13}$$

3) Qeury Keywords: The features of keywords are extracted in a similar way with query sentence. Given the set of keywords with m words $k = \{w_1, w_2, \ldots, w_m\}$, we utilize the embedding matrix $\mathbf{M}_w$ shared with query sentence to map each word into a d-dimensional vector e_i. Considering that keywords are usually short and there is no explicit sequential relationship between words, hence, we employ an MLP and obtain the features h_i

$$h_i = \text{MLP}\left(e_i\right). \tag{14}$$

Then, we introduce keywords graph module to perform further enhancement. Similarly, we construct a fully connected graph, initialize the state of each node h_i^0 with h_i, compute the edge weights referring to (11), aggregate features (13), stack L layers of keywords graph module, and obtain the global information feature vector of the keywords kw.

C. Multiscale Self-Attention

Following the work [7], we adopt the multiscale self-attention module to filter the background noise in RS image and solve the problem of inaccurate image-text matching caused by the multiscale and multicategory characteristics of objects in RS image. Specifically, we first fuse the low-level and high-level features that obtained in (8) and (9) as follows:

$$\widehat{v}^{(l)} = \text{PReLU}\left(\text{conv}_{3\times 3}\left(v^{(l)}\right)\right) \tag{15}$$

$$\widehat{v}^{(h)} = \text{PReLU}\left(\text{MLP}\left(v^{(h)}\right)\right) \tag{16}$$

where a $\text{conv}_{3\times 3}$ and MLP is applied to low-level and high-level features, respectively, and a parametric rectified linear unit (PReLU) is adopted to supplement nonlinear information. Afterward, we fuse the mean-pooling result of $v^{(h)}$ and fuse it with the concatenated features of $\widehat{v}^{(l)}$ and $\widehat{v}^{(h)}$, obtaining the visual representation $v^{(lh)}$

$$v^{(lh)} = \text{Mean}\left(v^{(h)}\right) \oplus \text{Cat}\left(\widehat{v}^{(l)}, \widehat{v}^{(h)}\right). \tag{17}$$

Then, considering RS image contains various targets, thus, inherent the redundant information. To alleviate the affect of redundant features in $v^{(l,h)}$, we design a redundant feature filter. We first perform the L_2 regularization for $v^{(l,h)}$ and adopt an MLP to conduct the feature transform. And then, we employ a sigmoid function to calculate activation values, which are used to select the features from the transformed representation, and obtain the final visual features $\widehat{v^{(i)}}$

$$v^{(i)} = \text{MLP}\left(L_2\left(v^{(l,h)}\right)\right) \tag{18}$$

$$\text{act} = \sigma\left(v^{(i)}\right) \tag{19}$$

$$\widehat{v^{(i)}} = \text{act} \cdot v^{(i)}. \tag{20}$$

D. Textual Feature Fusion

Despite that query sentence describe the connection among multiple objects, it lacks the target information. Query keywords list the interested targets in the RS image, which can be used to supplement the sentence. Therefore, we design textual feature fusion module to fuse sentence and keywords information. In addition, while fusing the features of query sentence and query keywords, we deliberately introduce RS image features to avoid the problem of image and text misalignment.

Specifically, for the sentence features sent, we first concatenate them with visual features $\widehat{v^{(i)}}$ and adopt an MLP to perform transformation, then we calculate the sigmod value to reassign the weights for each feature and obtain the visual-aware sentence representation $\text{sent}^{(v)}$ by elementwise multiplication

$$\widehat{\text{sent}} = \text{MLP}\left(\text{Cat}\left(\text{sent}, \widehat{v^{(i)}}\right)\right) \tag{21}$$

$$\text{act} = \sigma\left(\widehat{\text{sent}}\right) \tag{22}$$

$$\text{sent}^{(v)} = \text{act} \cdot \widehat{\text{sent}}. \tag{23}$$

In a similarly manner, we obtain the visual-aware keywords representation $\text{kw}^{(v)}$. Afterward, we propose a dynamic fusion module to fuse the features of sentence and keywords. In particular, we perform the L_2 regularization for $\text{sent}^{(v)}$ and $\text{kw}^{(v)}$, respectively, and adopt an MLP to transform their concatenated results. Next, we calculate activation values via sigmod function, which act as the gate to control the feature fusion, and obtain

TABLE I
STATISTICS OF TWO DATASETS USED IN OUR EXPERIMENTS

Dataset	Images	Classes	Size	Keyword
RSICD	10921	31	224	No
RSITMD	4743	32	256	Yes

the final textual features F_t

$$\text{sent}^t = \text{MLP}\left(\text{Cat}\left(L_2\left(\text{sent}^{(v)}\right), L_2\left(\text{kw}^{(v)}\right)\right)\right) \tag{24}$$

$$F_t = \text{sent}^{(v)} \cdot \sigma\left(\text{sent}^t\right) \oplus \text{kw}^{(v)} \cdot \left(1 - \sigma\left(\text{sent}^t\right)\right). \tag{25}$$

E. Loss Function

Given a sample pair (T, I), we take the triplet loss as our optimization objective, which is formulated as

$$L_{\text{ct}}(I,T) = \sum_{\widehat{T}} \left[\alpha_{\text{ct}} - S(I,T) + S\left(I, \widehat{T}_h\right)\right]_+ + \sum_{\widehat{I}} \left[\alpha_{\text{ct}} - S(I,T) + S\left(\widehat{I}_h, T\right)\right]_+ \tag{26}$$

where $\widehat{T}$ and $\widehat{I}$ denote the negative sentences and images, $[x]_+$ equals $\max(x, 0)$, $S(I,T)$ represents the similarity of image and text, and α_{ct} is the margin that has considered the text similarity, which is defined as

$$\alpha_{\text{ct}} = \gamma \frac{-e^{\beta S(T,T_I)} + e^{\beta}}{-1 + e^{\beta}}. \tag{27}$$

$S(T, T_I) \in (0,1)$ refers to BLEU and METEOR scores that measures the similarity between the text T and sentences corresponding to image T_I, e is the natural index, γ is the maximum margin, and β is the decay coefficient.

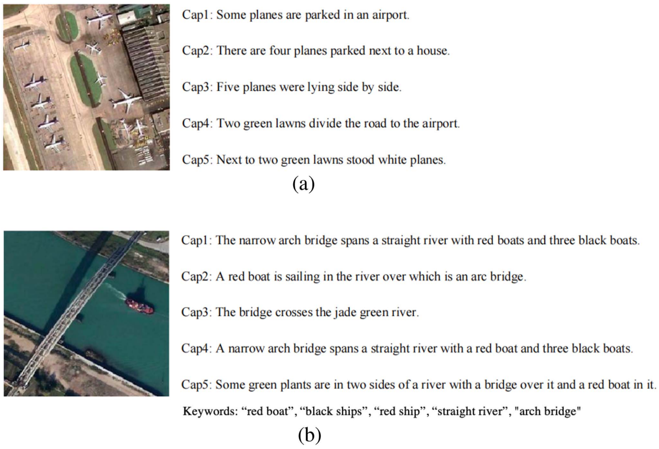
IV. DATASET DESCRIPTION AND EXPERIMENTAL SETUP

To verify the effectiveness of our proposed method, we conduct a series of experiments on two public datasets. In this section, we first introduce the characteristics of these two datasets, then describe the evaluation metrics and implementation details. Finally, we provide the description of four state-of-the-art methods.

A. Dataset

Our proposed method is trained and evaluated on two datasets: RSICD [57] and RSITMD [7] for the cross-modal image-text retrieval. Here, we give the statistics of these two RS datasets in Table I.

1) *RSICD [57]:* This dataset is constructed by Lu et al. The dataset is composed of 10 921 images of pixel size 224 × 224, and it is the largest dataset used for RS image captioning. All images are captured from the airplane or satellite. Five different sentences are exploited to describe every image. Examples of images along with their descriptions are shown in Fig. 3(a). The total number of sentences in RSICD is 54 605, and the total words of these sentences are 3323. This dataset contains land use images in 28 classes, including airport, bare land, forest, pond, baseball field, industrial, viaduct, beach, meadow, port, bridge, medium residential, railway station, center, mountain, resort, church, park, river, commercial, school, sparse residential, dense residential, square, storage tanks, desert, parking, stadium, with 240 to 1031 images per category.



Cap1: Some planes are parked in an airport.

Cap2: There are four planes parked next to a house.

Cap3: Five planes were lying side by side.

Cap4: Two green lawns divide the road to the airport.

Cap5: Next to two green lawns stood white planes.

(a)

Cap1: The narrow arch bridge spans a straight river with red boats and three black boats.

Cap2: A red boat is sailing in the river over which is an arc bridge.

Cap3: The bridge crosses the jade green river.

Cap4: A narrow arch bridge spans a straight river with a red boat and three black boats.

Cap5: Some green plants are in two sides of a river with a bridge over it and a red boat in it.

Keywords: "red boat", "black ships", "red ship", "straight river", "arch bridge"

(b)

Fig. 3. Data samples of the two datasets used in the experiment. Data format of the two datasets are the same, and each image is accompanied by five language descriptions. The above and below are examples of RSICD and RSITMD, respectively. (a) RSICD sample. (b) RSITMD sample.

2) *RSITMD [7]:* This dataset is proposed by Yuan et al. It is a fine-grained RS image-text dataset, including 4743 images and 23 715 sentence descriptions. Some images in RSITMD are selected from the RSICD dataset, and others are from Google Earth. RSITMD includes 32 categories, including industrial, stadium, storage tanks, square, playground, river, viaduct, pond, port, farmland, resort, school, park, dense residential, sparse residential, bridge, beach, commercial, center, parking, airport, church, medium residential, meadow, desert, forest, railway station, mountain, baseball field, intersection, bare land, boat, with 55 to 207 images per category. Furthermore, RSITMD provides 1 to 5 fine-grained keywords for each sample image to further reduce the consistency of retrieved information.

For each dataset, we use 80% of the samples as the training set, 10% of the sample for validation, and the remaining 10% as the test set.

B. Evaluation Metrics

We choose two evaluation indicators Recall at K ($R@K$) and mR as our metrics, which are the most frequently used indicator in cross-modal image-text retrieval tasks.

$R@K$ (K = 1, 5, and 10) indicates the percentage of queries in which the ground-truth matchings are contained in the top K retrieved results. The higher value of $R@K$ means a better performance. In the experiment, we choose the value of K as 1, 5, and 10 for analysis, i.e., "$R@1$", "$R@5$," and "$R@10$."

TABLE II
HYPERPARAMETERS OF OUR METHOD

Hyperparameters	Value
Image feature size	512
Word embedding dimension	300
GRU hidden state size	512

To evaluate the overall performance of the model more reasonably, we obtained mR by calculating the average of all six recall rates of $R@K$ as proposed by Huang et al. [58]. It is the sum of all $R@1$, $R@5$, and $R@10$ scores of image-to-text and text-to-image retrieval.

C. Implementation Details

All our experiments are performed on a single Nvidia Titan RTX GPU. For the image modality, we first resize it to 278 × 278, and then perform random scaling (from 0.5 to 2.0), random horizontal flipping and random cropping to enhance the training set images. The final size of the images is 256 × 256 pixels. To ensure the consistency of the experiments and to guarantee that the model does not suffer from overfitting caused by the excessive depth of the network, we use ResNet-18 [55] as the visual feature extractor for experiments.

During the training, we achieve the best performance during the training by using Adam optimizer with a learning rate initialization of 0.003, batch size and dropout rate of 16 and 0.5. We apply early stopping with the patience of 20 to avoid the model from overfitting. Furthermore, we summarize the hyperparameters in Table II.

D. State-of-the-Art Approaches

In this part, we contrast our proposed method with the following four methods on two RS image-text datasets.

1) *VSE++ [59]:* VSE++ is one of the enlightening works of image retrieval in the field of natural images. This work is based on the visual-semantic embedding (VSE) model [60]. The VSE model uses a long short-term memory to encode sentences to obtain textual feature representations, and a CNN is used to extract image feature. Two mapping matrices are learned under the supervision of a bidirectional hierarchical loss function so that two cross-modal features can be mapped to the same embedding space for alignment. VSE++ improves the bidirectional loss function in the VSE model. After that, triplet loss is proposed to train the image-text matching model. VSE++ introduces the idea of the most difficult negative sample, which makes the final image retrieval accuracy much higher.
2) *SCAN [61]:* The SCAN model applies an attention mechanism to text and images, respectively, and learns better representations of text and images, then uses hard triplet loss in a shared subspace to measure the similarity between text and images. It uses Faster RCNN [62] to extract image features on the basis of VSE++ to generate k target regions for each image, and transform the embedding matrix into an h-dimensional vector. Each word of the text gets a one-hot vector, which is a 300-dimensional vector after embedding, and then uses a Bi-GRU to get an h-dimensional vector. Finally, it aligns the image and the text with the same semantic target. This work is compared with the text-to-image method and image-to-text method proposed in this article in the experimental part.
3) *CAMP [63]:* This work proposes an adaptive message passing method to adaptively control the flow of cross-modal information transmission. The fusion features are used to calculate the matching scores of images and texts while using CAMP's triple loss method and BCE loss method as a control. CAMP believes that salient regions in images and salient words in sentences should be paid attention. It takes into account the interactions between regions and words and finds fine-grained cues for cross-modal matching by filtering out irrelevant information. That is, for CAMP to find the gist of images and texts, interacting the two is beneficial to capture fine-grained cross-modal cues for text-image matching.
4) *MTFN [49]:* Based on the idea of rank decomposition, the MTFN model designs a multimodal fusion network to calculate the distance of embedded features. MTFN is a novel image-text retrieval network. Instead of learning a latent common space for image-sentence pair, MTFN design a similarity function to accurately measure the distance between the input image and the sentence, and train this network using a ranking loss function.

V. EXPERIMENTAL RESULT

In order to evaluate the effectiveness of the proposed method CMFM-Net, we conduct a comprehensive experimental analysis from the following three aspects: overall comparison, ablation experiments, and visual analysis. The ResNet-18 network is used here instead of the target detection network. Considering that the input of our proposed method is keywords and sentences, we filter the keywords in the sentence and input them into the keyword branch when only sentences are input.

A. Experimental Results on RSICD Dataset

Table III shows the test results of all methods on RSICD dataset. Compared with the state-of-the-art models in recent years, our method's overall effect has reached the best. In other words, our CMFM-Net model is the best of all models. Compared with the previous best model MTFN [49], our experimental results increase mR by 2.94%, which shows that our method is effective. Our proposed method learns the feature fusion intramodalities and the feature association intermodalities to avoid the poor retrieval performance caused by the information asymmetry, which improves the overall retrieval accuracy.

In sentence retrieval, it shows that our model can surpass other methods on metrics including $R@5$ and $R@10$. It can be seen that

TABLE III
COMPARISONS OF SENTENCE-IMAGE RETRIEVAL EXPERIMENTS ON RSICD TEST SET

Approach	Sentence retrieval			Image retrieval			mR
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
VSE++ [59]	3.38	9.51	17.46	2.82	11.32	18.10	10.43
SCAN t2i [61]	4.39	10.90	17.64	3.91	16.20	26.49	13.25
SCAN i2t [61]	**5.85**	12.89	19.84	3.71	16.40	26.73	14.23
CAMP-Triplet [63]	5.12	12.89	21.12	4.15	15.23	27.81	14.39
CAMP-BCE [63]	4.20	10.24	15.45	2.72	12.76	22.89	11.38
MTFN [49]	5.02	12.52	19.74	4.90	17.17	29.49	14.81
CMFM-Net (ours)	5.40	**18.66**	**28.55**	**5.31**	**18.57**	**30.03**	**17.75**

All values in the table default to percent signs. The bold entities represent the best results.

TABLE IV
COMPARISONS OF SENTENCE-IMAGE RETRIEVAL EXPERIMENTS ON RSITMD TEST SET

Approach	Sentence retrieval			Image retrieval			mR
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
VSE++ [59]	10.38	27.65	39.60	7.79	24.87	38.67	24.83
SCAN t2i [61]	10.18	28.53	38.49	**10.10**	28.98	43.53	26.64
SCAN i2t [61]	11.06	25.88	39.38	9.82	29.38	42.12	26.28
CAMP-Triplet [63]	**11.73**	26.99	38.05	8.27	27.79	44.34	26.20
CAMP-BCE [63]	9.07	23.01	33.19	5.22	23.32	38.36	22.03
MTFN [49]	10.40	27.65	36.28	9.96	31.37	45.84	26.92
CMFM-Net (ours)	10.84	**28.76**	**40.04**	10.00	**32.83**	**47.21**	**28.28**

All values in the table default to percent signs. The bold entities represent the best results.

SCAN i2t [61] performs the best among all these state-of-the-art methods on $R@1$. Although our proposed method CMFM-Net is only second to SCAN i2 [61] in terms of $R@1$, it displays a significantly better performance than SCAN i2t on other metrics. Specifically, compared with SCAN i2t [61], our model can exhibit an increase of 5.77% ($R@5$) and 8.71% ($R@10$), respectively.

In image retrieval, it can be seen that our model can achieve outstanding result on all metrics including $R@1$, $R@5$, and $R@10$. We can see that other methods perform better than sentence retrieval in this part, but our model can even surpass the best method by 0.41%($R@1$), 1.40% ($R@5$), and 0.54% ($R@10$), respectively.

B. Experimental Results on RSITMD Dataset

The results on the RSITMD dataset are shown in the Table IV. Our method outperforms all state-of-the-art methods in overall performance. It can be seen that MTFN [49] performs the best among all other methods, but our method can still outperforms MTFN. Specifically, compared with MTFN [49], our model can obtain an increase of 1.36% on mR metric, which shows the superiority of our model.

In sentence retrieval, our model has achieved the best results on other indicators except on $R@1$. CAMP-Triplet [63] performs best $R@1$. Our proposed method yields 28.76%($R@5$) and 40.04%($R@10$), and it outperforms CAMP-Triplet [63] by over 1.77%($R@5$) and 1.99%($R@10$).

In image retrieval, our method also achieves the best results, except on $R@1$. SCAN t2i [61] performs best on this metric, achieving a score of 10.10%. However, our method outperforms SCAN t2i [61] in other metrics. Our proposed method is better than SCAN t2i [61] by 3.85%($R@5$) and 3.68% ($R@10$), respectively.

C. Ablation Studies

In this section, we conduct ablation experiments on the CMFM-Net model to explore the contribution of each module to the model. To better analyze the performance of the model, we list a series of configurations in Tables V and VI. We use m_1, m_2, and m_3 to represent different modules.

1) m_1: text graph module.
2) m_2: RS image graph module.
3) m_3: image-text association module.

Four variant models are shown in both tables. The results of the ablation experiment in Tables V and VI together show that these three modules m_1–m_3 all contribute to this task in varying degrees. Specifically, the experimental results with the three modules separately are much better than those without any

TABLE V
COMPARISONS OF SENTENCE-IMAGE RETRIEVAL EXPERIMENTS ON RSICD TEST SET

Configuration			RSICD dataset						
			Sentence retrieval			Image retrieval			mR
m_1	m_2	m_3	R@1	R@5	R@10	R@1	R@5	R@10	
✓	✓		**5.95**	17.02	27.26	4.70	18.06	30.17	17.19
	✓	✓	5.92	16.96	27.05	4.43	18.28	**32.05**	17.45
✓		✓	5.40	17.66	27.81	4.30	17.31	31.02	17.25
✓	✓	✓	5.40	**18.66**	**28.55**	**5.31**	**18.57**	30.03	**17.75**

All values in the table default to percent signs. The bold entities represent the best results.

TABLE VI
COMPARISONS OF SENTENCE-IMAGE RETRIEVAL EXPERIMENTS ON RSITMD TEST SET

Configuration			RSITMD dataset						
			Sentence retrieval			Image retrieval			mR
m_1	m_2	m_3	R@1	R@5	R@10	R@1	R@5	R@10	
✓	✓		9.73	26.55	**40.04**	8.63	31.59	50.18	27.79
	✓	✓	9.29	23.45	37.61	7.72	32.30	**50.53**	26.82
✓		✓	9.29	24.56	35.84	9.47	30.35	48.01	26.25
✓	✓	✓	**10.84**	**28.76**	**40.04**	**10.00**	**32.83**	47.21	**28.28**

All values in the table default to percent signs. The bold entities represent the best results.

one module, indicating that each module plays an active role in the model's understanding of image and textual information. It is shown that the text graph module improves the retrieval accuracy by promoting the interactive fusion of features in the text modality. The RS image graph module helps to improve retrieval accuracy, because it strengthens the relationship between different objects in image modalities. The image-text association module strengthens associations between modalities by highlighting components in text associated with RS images.

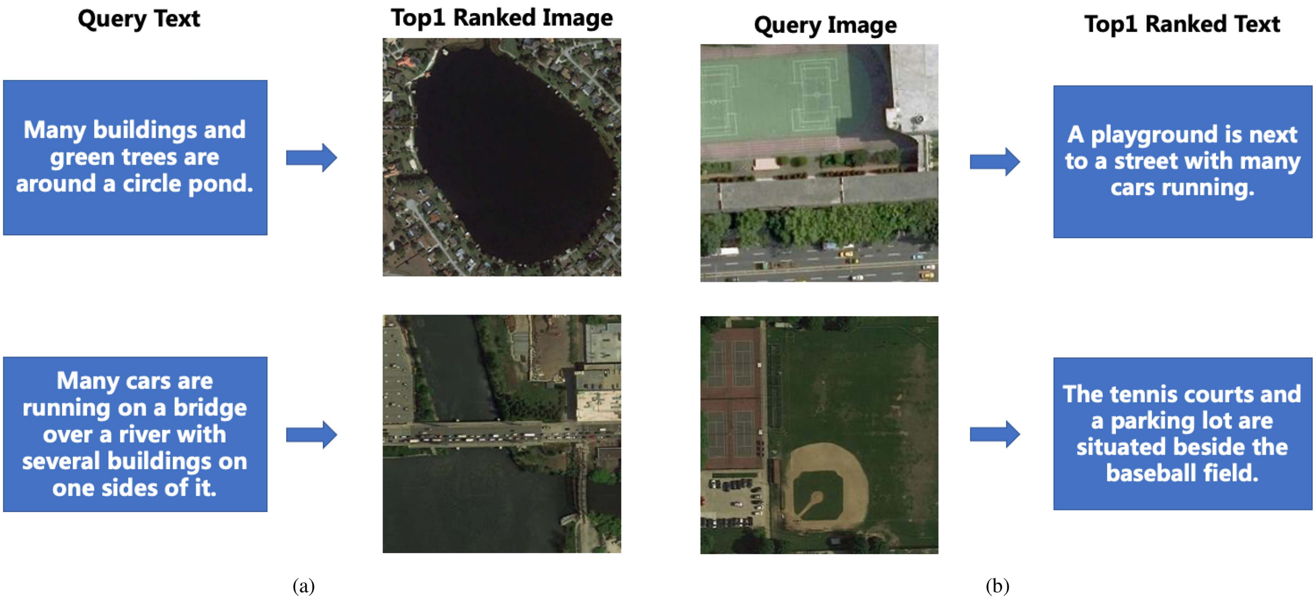
Table V shows the performance of each model variant on the RSICD dataset. The results in Table V tell us that the performance drop from removing each modalities is 0.3%, 0.50%, and 0.56%, respectively. It can be seen that the image-text association module has the greatest impact on the model performance in this dataset, while the RS image graph module has a greater impact on the model than the text graph module. Therefore, for the RSICD dataset, interactions between image and text modalities are the most important.

Table VI shows the performance of each model variant on the RSITMD dataset. It shows that the removal of modules m_1, m_2, and m_3 separately brings about 1.46%, 2.03%, and 0.49% performance losses to the model, which shows that the addition of text graph module and RS image graph module improves the model more than image-text association module. This result still cannot deny the contribution of the image-text association module to the model. From the ablation results in the RSITMD dataset, we can infer that the model's better understanding of the modal interior will help the model's overall understanding of multiple modalities. In other words, the attention within the modalities may be more important than the attention between the modalities in the RSITMD dataset.

Figs. 4 and 5 show the visualization results of our model on different datasets. Figs. 4(a) and 5(a) show the qualitative image retrieval results for a given query text on the RSICD and RSITMD dataset. For each query sentence, we display the top-one retrieved image ranked. Figs. 4(b) and 5(b) show the qualitative results of text retrieval with a given query image on the RSICD and RSITMD dataset. For each query image, we show the top one retrieved text ranked according to the similarity scores predicted by our model. From these results, we find that our method can retrieve correct results even in complex scenes with RS. As can be seen from the overall visualization results, our model is able to discover comprehensive and fine-grained correspondences between images and sentences by enhancing intramodality associations as well as interactions between different modalities.

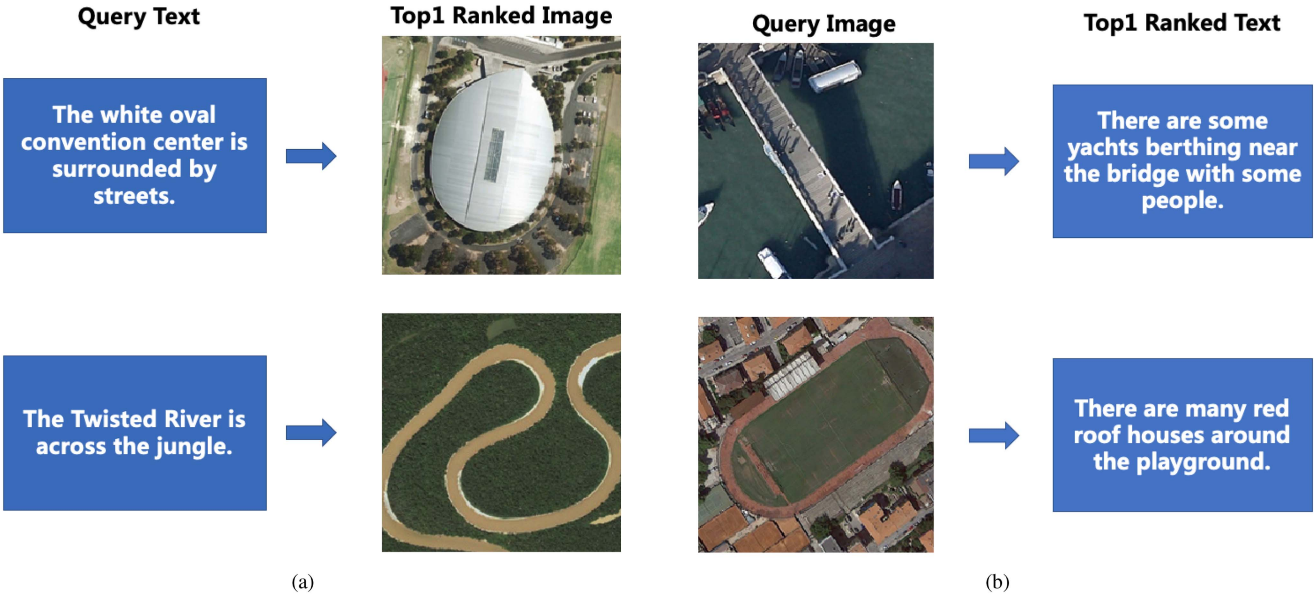
D. Explore Fast Locate Using Text

We follow the method of using text to locate RS images in large scenes proposed by Yuan et al. [7], which verifies the ability of our proposed method to locate target objects. We first cut the RS images in various ways to obtain patches with different sizes. We obtain a probability map for each patch by computing the similarity between each patch and the query text. Then, we stitch the obtained probability maps in place. By performing median


Query Text
Top1 Ranked Image
Query Image
Top1 Ranked Text
Many buildings and green trees are around a circle pond.
A playground is next to a street with many cars running.
Many cars are running on a bridge over a river with several buildings on one sides of it.
The tennis courts and a parking lot are situated beside the baseball field.

(a) (b)

Fig. 4. Visual results of the text-to-image retrieval and the image-to-text retrieval on RSICD dataset. For each query text/image, we show the top-one ranked image/text retrieved by our proposed method.


Query Text
Top1 Ranked Image
Query Image
Top1 Ranked Text
The white oval convention center is surrounded by streets.
There are some yachts berthing near the bridge with some people.
The Twisted River is across the jungle.
There are many red roof houses around the playground.

(a) (b)

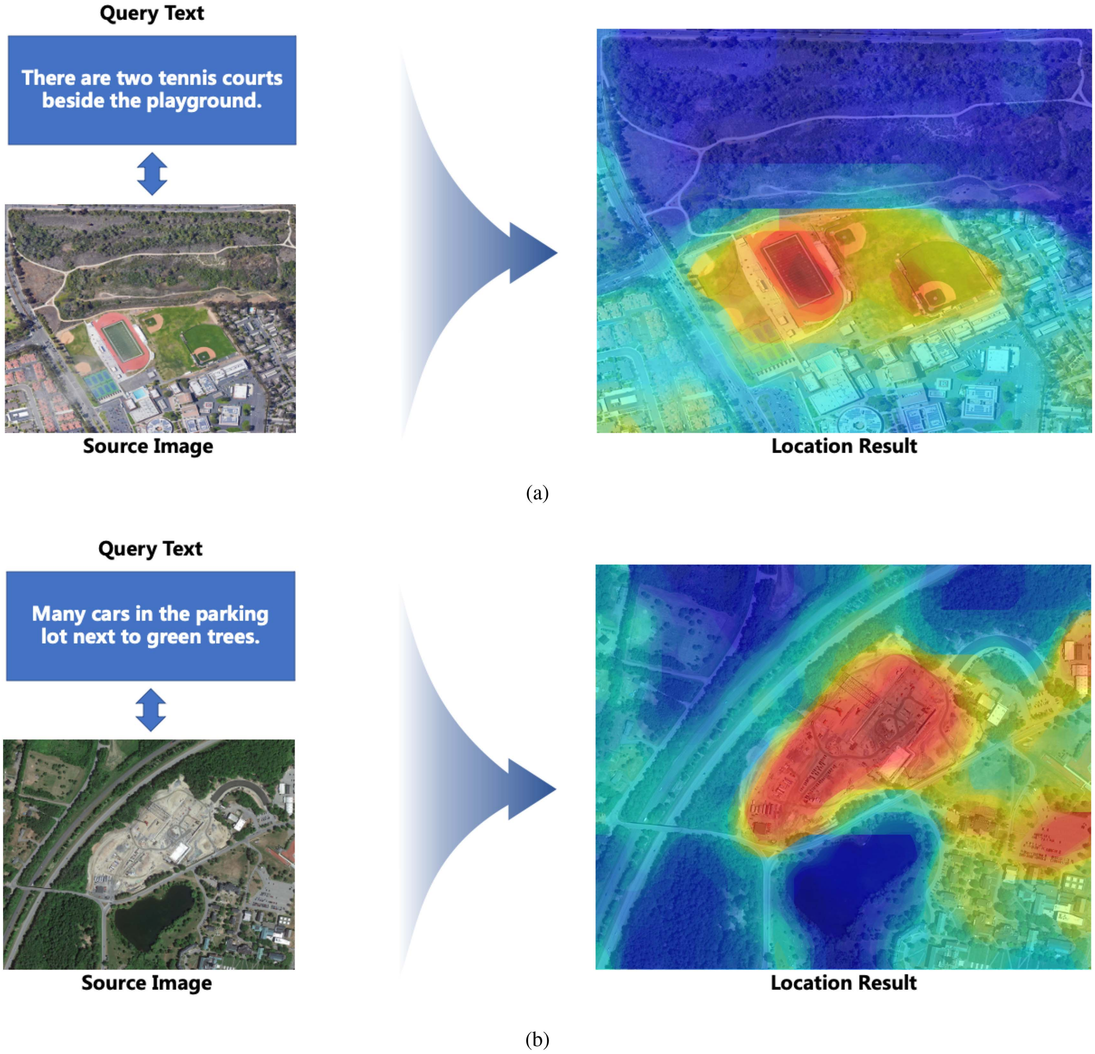
Fig. 5. Visual results of the text-to-image retrieval and the image-to-text retrieval on RSITMD dataset. For each query text/image, we show the top-one ranked image/text retrieved by our proposed method.

filtering on the generated probability map to filter the impulse noise in the probability map, the final location result can be obtained.

Fig. 6 shows some localization results. In Fig. 6(a), we try to find two tennis courts adjacent to the playground from the RS image of the large scene. It can be seen that two tennis courts and one playground in the location result have large probability values, but part of probability still falls on some lawns, which reflects that the model still has room for improvement.

In Fig. 6(b), we attempt to find parking lot around green trees. It can be seen from the results that our model not only locates a large parking lot, but also has a relatively large response value for several other small parking lots, indicating that our model has a relatively comprehensive ability to locate targets.

The abovementioned experiments show that although our proposed method CMFM-Net is not trained with the location ground truth, CMFM-Net can effectively locate the target object in the RS image with the large scene. The accuracy of localization is not as good as the object detection method with bounding box ground truth training, but the existing localization ability is enough to make the retrieval method more stable.


Query Text
There are two tennis courts beside the playground.
Source Image
Location Result
Query Text
Many cars in the parking lot next to green trees.
Source Image
Location Result

(a)

(b)

Fig. 6. Exploration results of location by text.

VI. Conclusion

The purpose of cross-modal RS image retrieval is to retrieve the most relevant RS images by using other modes (such as text) as queries. In recent years, with its flexible form, it has become a new research hotspot. Aiming at the problem that the information of query text and RS image is not aligned, this article proposes a new cross-modal RS feature matching network, i.e., CMFM-Net, based on GNN with strong representation ability. By learning the feature interaction in query text and RS image, respectively, and modeling the feature association between the two modes, we can avoid the information misalignment and improve the retrieval performance. Specifically, in order to fuse the intramodal features, the text and RS image graph module is designed. In addition, combined with the multihead attention mechanism, an image text association module is constructed to focus on the parts related to RS images in the text. Quantitative and qualitative experiments on two public standard datasets, RSICD and RSITMD, verify the good performance of the model.

References

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

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

[3] 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.

[4] C. R. Shyu, M. Klaric, G. J. Scott, A. S. Barb, and K. Palaniappan, "GeoIRIS: Geospatial information retrieval and indexing system–content mining, semantics modeling, and complex queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 839–852, Apr. 2007.

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

[6] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[7] Z. Yuan, W. Zhang, K. Fu, X. Li, and X. Sun, "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 1, May 2021, Art. no. 4404119.

[8] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.

[9] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics*. Berlin, Germany: Springer, 1992, pp. 162–190.

[10] N. Rasiwasia et al., "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[11] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4094–4102.

[12] D. Zeng, Y. Yu, and K. Oyama, "Deep triplet neural networks with cluster-CCA for audio-visual cross-modal retrieval," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 3, pp. 1–23, 2020.

[13] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 604–611.

[14] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[15] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.

[16] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[17] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1201–1216, Jun. 2016.

[18] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2017.

[19] J. Dong, X. Li, and C. G. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3377–3388, Dec. 2018.

[20] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[21] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2160–2167.

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

[23] 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, no. 1, pp. 5284–5296, Sep. 2020.

[24] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. IEEE Int. Conf. Comput. Inf. Telecommun. Syst.*, 2016, pp. 1–5.

[25] 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. 2017.

[26] X. Li, X. Zhang, W. Huang, and Q. Wang, "Truncation cross entropy loss for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5246–5257, Jun. 2020.

[27] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, "A multi-level attention model for remote sensing image captions," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 939.

[28] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 1, Apr. 2021, Art. no. 5603814.

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

[30] 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.

[31] R. Wang, H. Zhao, S. Ploux, B.-L. Lu, and M. Utiyama, "A bilingual graph-based semantic model for statistical machine translation," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 2950–2956.

[32] M. Xia, G. Huang, L. Liu, and S. Shi, "Graph based translation memory for neural machine translation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7297–7304.

[33] D. Cai and W. Lam, "Graph transformer for graph-to-sequence learning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 7464–7471.

[34] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8950–8959.

[35] H. Hu, D. Ji, W. Gan, S. Bai, W. Wu, and J. Yan, "Class-wise dynamic graph convolution for semantic segmentation," in *Proc. Comput. Vis. 16th Eur. Conf.*, 2020, pp. 1–17.

[36] G.-S. Xie, J. Liu, H. Xiong, and L. Shao, "Scale-aware graph neural network for few-shot semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5475–5484.

[37] S. Saha, S. Zhao, and X. X. Zhu, "Multitarget domain adaptation for remote sensing classification using graph neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, no. 1, Feb. 2022, Art. no. 6506505.

[38] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 241–245, Feb. 2018.

[39] S. Ren and F. Zhou, "Semi-supervised classification for PoLSAR data with multi-scale evolving weighted graph convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, no. 1, pp. 2911–2927, Feb. 2021.

[40] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu, "Cascade graph neural networks for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 346–364.

[41] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12997–13007.

[42] X. Jia, H. Zhao, Z. Lin, A. Kale, and V. Kumar, "Personalized image retrieval with sparse graph representation learning," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 2735–2743.

[43] Y. Gu, K. Vyas, M. Shen, J. Yang, and G.-Z. Yang, "Deep graph-based multimodal feature embedding for endomicroscopy image retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 481–492, Feb. 2020.

[44] M. Zhao, J. Liu, Z. Zhang, and J. Fan, "A scalable sub-graph regularization for efficient content based image retrieval with long-term relevance feedback enhancement," *Knowl. Based Syst.*, vol. 212, 2021, Art. no. 106505.

[45] X. Dong, L. Liu, L. Zhu, L. Nie, and H. Zhang, "Adversarial graph convolutional network for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1634–1645, Mar. 2021.

[46] S. Qian, D. Xue, Q. Fang, and C. Xu, "Adaptive label-aware graph convolutional networks for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 24, no. 1, pp. 3520–3532, Aug. 2021.

[47] P.-F. Zhang, Y. Li, Z. Huang, and X.-S. Xu, "Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 24, no. 1, pp. 466–479, Jan. 2021.

[48] A. K. Misraa, A. Kale, P. Aggarwal, and A. Aminian, "Multi-modal retrieval using graph neural networks," 2020. [Online]. Available: https://arxiv.org/abs/2010.01666

[49] T. Wang, X. Xu, Y. Yang, A. Hanjalic, and J. Song, "Matching images and text with multi-modal tensor fusion and re-ranking," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 12–20.

[50] A. Radoi and M. Datcu, "Multilabel annotation of multispectral remote sensing images using error-correcting output codes and most ambiguous examples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2121–2134, Jul. 2019.

[51] G. Wang, B. Fan, S. Xiang, and C. Pan, "Aggregating rich hierarchical features for scene classification in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4104–4115, Sep. 2017.

[52] Z. Zhang, W. Zhang, W. Diao, M. Yan, and X. Sun, "VAA: Visual aligning attention model for remote sensing image captioning," *IEEE Access*, vol. 7, pp. 137355–137364, 2019.

[53] A. Vaswani et al., "Attention is all you need," 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[54] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," Jul. 2016, *arXiv:1607.06450*.

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

[56] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," Sep. 2014, *arXiv:1409.1259*.

[57] 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. 2017.

[58] H. Yan, W. Qi, and W. Liang, “Learning semantic concepts and order for image and sentence matching,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 636–650, Mar. 2020.

[59] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “VSE++: Improving visual-semantic embeddings with hard negatives,” Jul. 2017, *arXiv:1707.05612*.

[60] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” *Comput. Sci.*, vol. 1411, 2014, Art. no. 2539.

[61] K. H. Lee, C. Xi, H. Gang, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.

[62] P. Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6077–6086.

[63] 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.

Hongfeng Yu received the B.Sc. degree in cartography and geographic information system and the M.Sc. degree in photogrammetry and remote sensing from Peking University, Beijing, China, in 2013 and 2016, respectively. He is currently working toward the Ph.D. degree in signal and information processing with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include deep learning and multimodal remote sensing interpretation.

Fanglong Yao received the B.Sc. degree in electronic information science and technology from Inner Mongolia University, Hohhot, China, in 2017 and the Ph.D. degree in signal and information processing from Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2022.

He is currently a Postdoctor with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include brain-inspired intelligence, cognition of complex system evolution, spatiotemporal data analysis, deep learning, multimodal learning, multitask learning, hypergraph learning, and causal learning.

Wanxuan Lu received the B.Sc. degree in detection, guidance and control technology from the Beijing Institute of Technology, Beijing, China, in 2016 and the Ph.D. degree in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2021.

She is currently an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. Her research interests include computer vision and remote sensing image processing.

Nayu Liu received the B.Sc. degree in electronic information science and technology from Xidian University, Xian, China, in 2018. He is currently working toward the Ph.D. degree in signal and information processing with the Aerospace Information Research Institute, University of Chinese Academic of Sciences, Beijing, China.

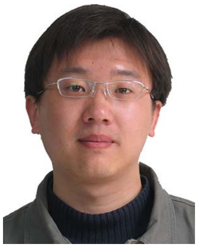
His research interests include deep learning, natural language processing, and multimodal learning.

Peiguang Li received the B.Sc. degree in electronic information science and technology from Central South University, Changsha, China, in 2017 and the Ph.D. degree in signal and information processing from the Aerospace Information Research Institute, University of Chinese Academic of Sciences, Beijing, China, in 2022.

His research interests include natural language inference, fake news detection, and multimodal information processing.

Hongjian You received the B.S. degree in engineering from Wuhan University, Wuhan, China, in 1992, the M.S. degree in engineering from Tsinghua University, Beijing, China, in 1995, and the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2001.

He is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His main research interests include remote sensing image processing and analysis, and SAR image applications.



Xian Sun (Senior Member, IEEE) received the B.Sc. degree in electronic information science and technology from Beihang University, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2006 and 2009, respectively.

He is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image understanding.