Hypergraph-Enhanced Textual-Visual Matching Network for Cross-Modal Remote Sensing Image Retrieval via Dynamic Hypergraph Learning

Fanglong Yao, Xian Sun, *Senior Member, IEEE*, Nayu Liu, Changyuan Tian, Liangyu Xu, Leiyi Hu, and Chibiao Ding, *Senior Member, IEEE*

***Abstract*—Cross-modal remote sensing (RS) image retrieval aims to retrieve RS images using other modalities (e.g., text) and vice versa. The relationship between objects in the RS image is complex, i.e., the distribution of multiple types of objects is uneven, which makes the matching with query text inaccurate, and then restricts the performance of remote sensing image retrieval. Previous methods generally focus on the feature matching between RS image and text and rarely model the relationships between features of RS image. Hypergraph (hyperedge connecting multiple vertices) is an extended structure of a regular graph and has attracted extensive attention for its superiority in representing high-order relationships. Inspired by the advantages of the hypergraph, in this work, a hypergraph-enhanced textual-visual matching network (HyperMatch) is proposed to circumvent the inaccurate matching between the RS image and query text. Specifically, a multiscale RS image hypergraph network is designed to model the complex relationships between features of the RS image for forming the valuable and redundant features into different hyperedges. In addition, a hypergraph construction and update method for an RS image is designed. For constructing a hypergraph, the features of an RS image running as vertices and cosine similarity is the metric to measure the correlation between them. Vertex and hyperedge attention mechanisms are introduced for the dynamic update of a hypergraph to realize the alternating update of vertices and hyperedges. Quantitative and qualitative experiments on the RSICD and RSITMD datasets verify the effectiveness of the proposed method in cross-modal remote sensing image retrieval.**

***Index Terms*—Cross-modal remote sensing (RS) image retrieval, dynamic hypergraph learning, hypergraph-enhanced textual-visual matching network (HyperMatch), multiscale RS image hypergraph.**

Manuscript received 12 September 2022; revised 28 October 2022; accepted 28 November 2022. Date of publication 6 December 2022; date of current version 23 December 2022. This work was supported by the National Natural Science Foundation of China under Grant #62171436. *(Corresponding author: Chibiao Ding.)*

The authors are wih 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, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yaofanglong17@mails.ucas.ac.cn; sunxian@aircas.ac.cn; 695704204@qq.com; tianchangyuan21@mails.ucas.edu.cn; xuliangyu21@mails.ucas.ac.cn; huleiyi21@mails.ucas.ac.cn; cbding@mail.ie.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3226325

I. Introduction

With the development of remote sensing (RS) information acquisition technology, the number of remote sensing images has increased exponentially. The collected remote-sensing images have diverse scenes and different types of objects. In addition, the resolution of remote sensing images varies significantly due to the different standards of sensors. Therefore, managing massive remote sensing images is complex, and it is urgent to update remote sensing retrieval technology. Cross-modal remote sensing image aims to use other modes such as text as queries to retrieve remote sensing images. With its flexible form, it has become a research hotspot in the field in recent years.

Previous methods generally generate textual descriptions and then retrieve remote sensing images by measuring the matching degrees between query text and the textual descriptions [1], [2], [3]. These methods are essentially text-to-text retrieval, which ignores the direct matching between remote sensing image and query text and are susceptible to the quality of generated textual description. To avoid the disadvantages of two-stage retrieval, Yuan et al. [4] propose an end-to-end retrieval method to directly learn the matching degree between query text and remote sensing image.

Although the above methods have promoted the development of cross-modal remote sensing image retrieval and aroused widespread concern in the industry, they still face the following three challenges.

1) As shown in Fig. 1, many objects are in the remote sensing image, including planes, cars, buildings, and other objects. In addition, the distribution of similar objects is uneven, and the scale of objects is inconsistent, i.e., different objects with various pixels. How to reasonably model the relationship between complex objects in remote sensing images and deal with the multiscale problem of remote sensing images has become the first challenge.
2) In addition, the relationship between words in the query text also needs to be quantified. Different terms have different contributions to other words. The second challenge is how to accurately quantify the contribution relationship between words in the query text.
3) There is a corresponding relationship between the object in the remote sensing image and the entity in the query text.

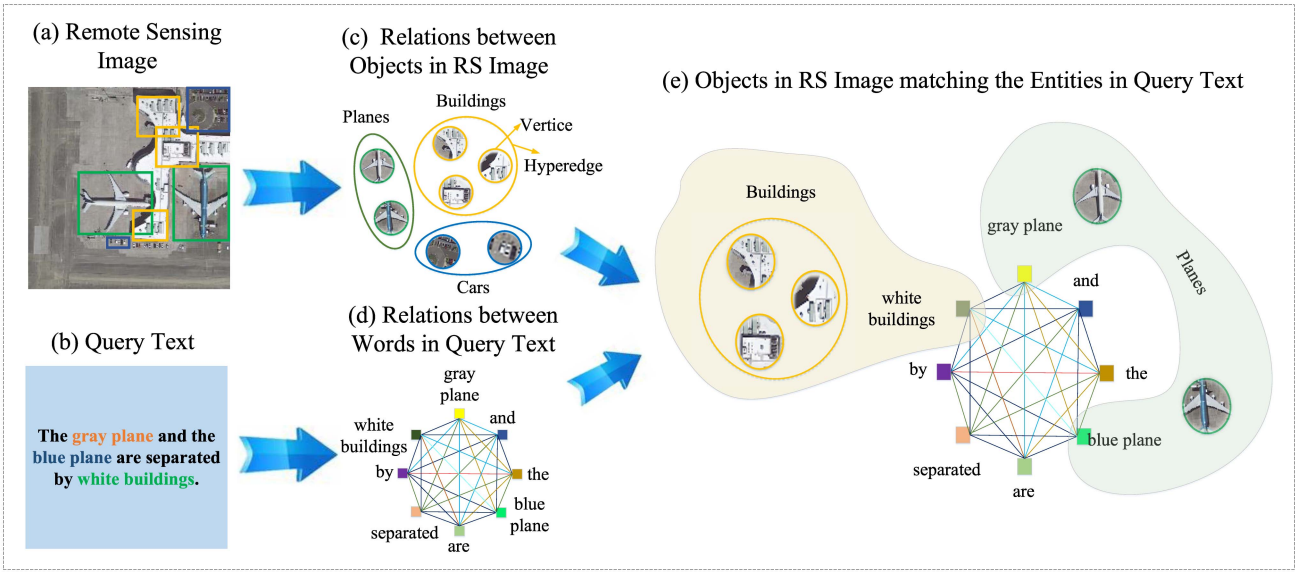


Fig. 1. Matching process between RS image and query text in cross-modal RS image retrieval. (a) Objects in an remote sensing image are characterized by many categories, uneven distribution, and different scales. (b) Query text containing various entities. (c) Hypergraph models the relationships between objects in RS image, and the objects of the same category are clustered into a hyperedge. Note that objects in an RS image are regarded as vertices, and objects belonging to the same category are connected by a hyperedge. (d) Undirected fully connected graph quantifies the mutual contribution of words in the query text. (e) Matching objects in the RS image and entities in the query text.

The third challenge is measuring the correlation between the query text and the remote sensing image and making the entity in the query text accurately match the object in the remote sensing image.

In recent years, graph neural network has developed rapidly to model the relationship between vertices. Inspired by the graph neural network, we use the undirected fully connected graph structure to model the relationship between words in the query text. Words are regarded as the vertices, and edges are used to quantify the contributions of words to each other.

The relationship between data is not only a simple pairwise relationship but also a more complex relationship between multiple vertices. Unlike the ordinary graph structure in which one edge can only connect two vertices, the hyperedge in the hypergraph can connect any number of vertices, which makes the hypergraph naturally suitable for modeling multivertex relationships. Inspired by the superiority of hypergraph, we choose to use hypergraph to model the complex relationship between objects in a remote-sensing image and use the same hyperedge to connect the objects belonging to the same category, as shown in Fig. 1(c). To solve the multiscale problem of remote sensing image, we design high-level and low-level RS image hypergraph networks to learn the correlation between multiobjects at different scales. Specifically, for the high-level RS image hypergraph network, the high-level RS image features are used as the vertices of the hypergraph, and the related RS image features are formed as the hyperedges. The vertices and hyperedges of low-level RS image are similar to those of high-level RS image hypergraph network.

In this article, we introduce dynamic hypergraph learning into cross-modal remote sensing image retrieval. A hypergraph-enhanced textual-visual matching network (HyperMatch) is proposed to circumvent the problem of inaccurate RS image and query text matching. To model the relationships between multiple objects in RS image at different scales, the high-level hypergraph network and low-level hypergraph network are designed, respectively. For the construction of a hypergraph, cosine similarity is employed to measure the correlation between objects. Hypergraph attention is elaborated for the dynamic alternating update of vertices and hyperedges for hypergraph evolution. In addition, an undirected fully connected graph network is applied to quantify the mutual contribution of words in the query text. Furthermore, the multiscale feature fusion and the image-guided multimodal fusion are designed to fuse the RS image features at different scales and extract the valuable text features for accurate matching with the RS image, respectively.

In summary, the contributions are as follows.

1) This article introduces hypergraph learning into cross-modal RS image retrieval and correspondingly propose a HyperMatch to avoid inaccurate matching between RS image and query text.
2) Aiming at the issue of multiple types, uneven distribution, and multiscale objects in RS images, the high-level and low-level RS image hypergraph networks are designed to model the relationship between objects at different scales, respectively, to cluster the similar object features into a hyperedge. Besides, an undirected fully connected graph network is conceived to quantify the contribution of words to each other in the query text.
3) A dynamic hypergraph learning algorithm for RS image is proposed to measure the correlation between objects and realize the alternated updating of vertices and hyperedges.

4) Quantitative and qualitative experiments on the published RSICD and RSITMD datasets verify the effectiveness of the proposed method in cross-modal remote sensing image retrieval.

II. Related Work

In this section, we mainly review the previous work that is most relevant to our proposed method, including cross-modal remote sensing image retrieval, text-image matching in natural scenes, and hypergraph learning.

A. Cross-Modal Remote Sensing Image Retrieval

To address the modality discrepancy caused by imaging mechanisms of synthetic aperture radar (SAR) and optical images, Xiong et al. [5] propose a cross-modality hashing network to extract the contour and texture shared features from across modalities. A CNN-RNN framework accompanied by beam search is exploited in [6] to generate multiple captions for retrieving RS images. Mao et al. [7] design a deep visual-audio network to directly capture the correspondence of image and audio for speech-to-image retrieval. Demir et al. [8] introduce hashing-based approximate nearest neighbor search to project high-dimensional image feature vectors into compact binary hash codes for content-based image retrieval. Hang et al. [9] propose an unsupervised feature learning model using multimodal data, hyperspectral, and light detection and ranging (LiDAR). A multiscale progressive segmentation network is proposed in [10] to address the issue of simultaneously segmenting objects with large-scale variations in high-resolution remote sensing imageries. Hang et al. [11] propose a spectral super-resolution network guided by the spectral correlation and the projection properties of hyperspectral imagery. To cope with cross-source RS image retrieval, Li et al. [12] introduce a source-invariant hashing convolutional neural network which can be optimized in an end-to-end manner. To reduce the memory and improve the retrieval efficiency, Chen et al. [13] propose an image–voice retrieval network to capture more information on RS data for generating hash codes with low memory. A cross-source distillation network with a well-designed joint optimization configuration is proposed in [14] to solve the data drift in cross-source content-based RS image retrieval (CS-CBSIR). Lv et al. [15] explore an image translation-based framework to address the data drift in CS-CBSIR by mapping the source domain to the object domain and keeping the generated images' content similar to the origin. To reduce the occupancy and overhead of cross-modal RS image retrieval algorithm, Yuan et al. [16] come up with a concise but effective cross-modal retrieval method via contrast learning and knowledge distillation.

B. Textual-Visual Matching

Wang et al. [17] present a fusion layer-based approach to extract the relationship between crossmodal features and a straightforward gradient-updating method to reduce the computational complexity for textual-visual matching. Li et al. [18] devise an identity-aware two-stage deep learning framework to scan incorrect matchings and refine the matching results with a latent coattention mechanism. Lee et al. [19] present stacked cross-attention to discover the latent alignments between image regions and words in the text for inferring image-text similarity. To learn modality-invariant feature representations, a text-image modality adversarial matching method incorporating adversarial learning is introduced in [20]. Liu et al. [21] propose a graph-structured matching network to construct graph structure for image and text and exploit graph convolution to propagate node correspondence for inferring fine-grained phrase correspondence. To learn the matching relations between image and text, Ma et al. [22] employ convolutional architecture to encode the image and compose semantic words. Messina et al. [23] introduce a transformer-based relationship-aware network to map visual and textual modalities into a common abstract concept space by sharing the weights of self-attentive layers. To capture the interrelationship of cross-modalities, Nguyen et al. [24] introduce a local and global scene graph matching model to extract and learn insightful features of nodes and edges from image and text graphs. Gu et al. [25] incorporate image-to-text and text-to-image generative models into cross-modal feature embedding for learning high-level and local-grounded representation.

C. Hypergraph Learning

To uncover complex higher-order interactions in different applications, Zhang et al. [26] develop a new self-attention-based graph neural network for handling homogeneous and heterogeneous hypergraphs with variable hyperedge size. For the adapting of hypergraph topology, Zhang et al. [27] devise a hypergraph Laplacian adaptor which adopts a self-attention mechanism to capture global information and trainable distance matrix to empower the updating of the topology in an end-to-end manner. To explore the data distribution's local structure, Ma et al. [28] present an approximation algorithm of hypergraph p-Laplacian regularization to preserve the geometry of the probability distribution. Duan et al. [29] present a local constraint-based sparse manifold hypergraph learning algorithm to discover the manifold-based light structure and the multivariate discriminant sparse relationship of hyperspectral images. Wei et al. [30] introduce an information-sharing mechanism to share the same structural distribution while preserving the specificity of each low-dimensional representation via adjusting the view-dependent hyperedge weights. To reduce the dimension of the hyperspectral image, Luo et al. [31] propose a sparse-adaptive hypergraph discriminant analysis method for adaptively revealing the intrinsic structure relationships with sparse representation.

III. Problem

In this work, we focus on text-based cross-modal RS image retrieval. Therefore, establishing a text-image matching model is the primary problem to be solved. RS image possesses the attributes of multiscale and multiobjective, so how to reasonably model the relationship between complex objects and deal with the multiscale issue becomes the first challenge. In addition,

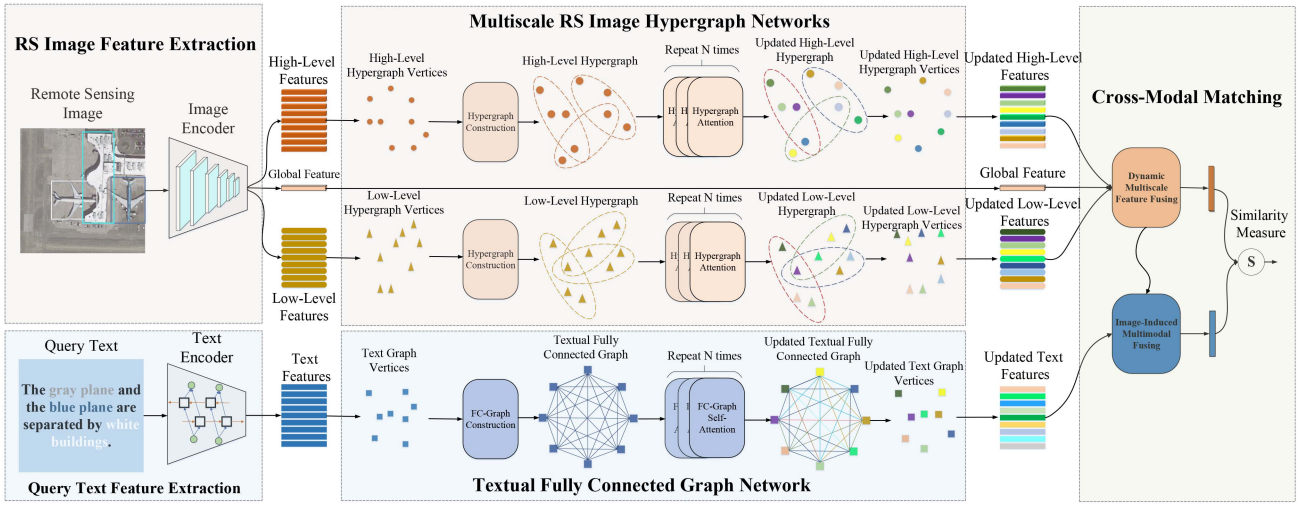


Fig. 2. Overview of HyperMatch Aiming at the issue of multiple types, uneven distribution, and multiscale objects in an RS image, the multiscale RS image hypergraph networks are designed to model the relationship between objects at different scales by clustering the similar object features into a hyperedge. Besides, a textual fully connected graph network is conceived to quantify the contribution of words to each other in the query text. In addition, we develop a cross-modal matching module to grasp the coreference relationship and improve the retrieval accuracy.

the query text is composed of multiple words, and different terms have different contributions to others. How to accurately quantify the contributing relationships among words is the second challenge. Corresponding relationships exist between the objects in the RS image and the entities in the query text. The third challenge is ensuring that the objects match the related entities.

Main problem (Cross-modal matching): Given an RS image $\mathcal{I}$ and a query text $\mathcal{T}$, the goal of cross-modal RS image retrieval is to build a model $\mathcal{F}$ to measure the matching degree $\mathcal{S}$ between them, i.e.,

$$\mathcal{S} = \mathcal{F}(\mathcal{I}, \mathcal{T}) \tag{1}$$

where $\mathcal{S}$ denotes the cross-modal similarity for measuring the matching degree.

Challenge 1 (Relationships between objects): For an RS image $\mathcal{I}$, it owns multiple objects at different scales. Thus, it is necessary to devise a module not only to solve the multiscale of objects but also to cluster the objects that belong to the same categories, i.e.,

$$\begin{cases} \mathcal{I}_{updated}^{high} = \mathcal{F}_1^{high}(\mathcal{I}; \vartheta_1^{high}) \\ \mathcal{I}_{updated}^{low} = \mathcal{F}_1^{low}(\mathcal{I}; \vartheta_1^{low}) \end{cases} \tag{2}$$

where $\mathcal{I}_{updated}^{high}$ is the RS image features encoded high-scale object information, $\mathcal{F}_1^{high}(\cdot)$ represents the module to model the relationships between high-scale objects, and ϑ_1^{high} stands for the learnable parameters. The meanings of $\mathcal{I}_{updated}^{low}$, $\mathcal{F}_1^{low}(\cdot)$, and ϑ_1^{low} are similar to $\mathcal{I}_{updated}^{high}$, $\mathcal{F}_1^{high}(\cdot)$, and ϑ_1^{high}.

Challenge 2 (Relationships between entities): The contribution of entities to each other in the query text is different, so we need to build a module to measure the contribution relationships, i.e.,

$$\mathcal{T}_{updated} = \mathcal{F}_2(\mathcal{T}; \vartheta_2) \tag{3}$$

where $\mathcal{T}_{updated}$ is the updated query text features, $\mathcal{F}_2(\cdot)$ represents modules for learning the contribution relationships between entities, and ϑ_2 stands for the learnable parameters.

Challenge 3 (Objects matching entities): RS images retrieved through query text usually have a high degree of compatibility, which is mainly reflected in the correspondence between the objects of the RS image and the entities in the query text. Therefore, it is essential to construct a matching method to learn the correspondence relationships, as follows:

$$\mathcal{S} = \mathcal{F}_3(\mathcal{I}_{updated}^{high}, \mathcal{I}_{updated}^{low}, \mathcal{T}_{updated}; \vartheta_3) \tag{4}$$

where $\mathcal{S}$ denotes the cross-modal similarity for measuring the matching degree, $\mathcal{F}_3(\cdot)$ refers to the matching method for objects in RS image and entities in query text.

IV. Methodology

To accurately retrieve the RS images according to the query text or find the appropriate descriptions through the RS image, we construct a hypergraph-enhanced textual-visual matching network named HyperMatch. As illustrated in Fig. 2, HyperMatch contains RS image feature extraction, query text feature extraction, multiscale RS image hypergraph network, textual fully connected graph network, and cross-modal matching. In the following, we will introduce the components in detail.

A. Preliminaries

Hypergraph definition: A hypergraph with n vertices and m hyperedges can be defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathcal{V} = \{\mathcal{V}_i\}_{i=1}^n$, $\mathcal{E} = \{\mathcal{E}_j\}_{j=1}^m$ represents the set of vertices and hyperedges, respectively. $\mathbf{W} = diag(w_1, w_2, \ldots, w_m)$ is a diagonal matrix of hyperedge weights. The structure of a hypergraph can also be formulated by an incidence matrix $\mathbf{H} \in \mathbb{R}^{n \times m}$, with

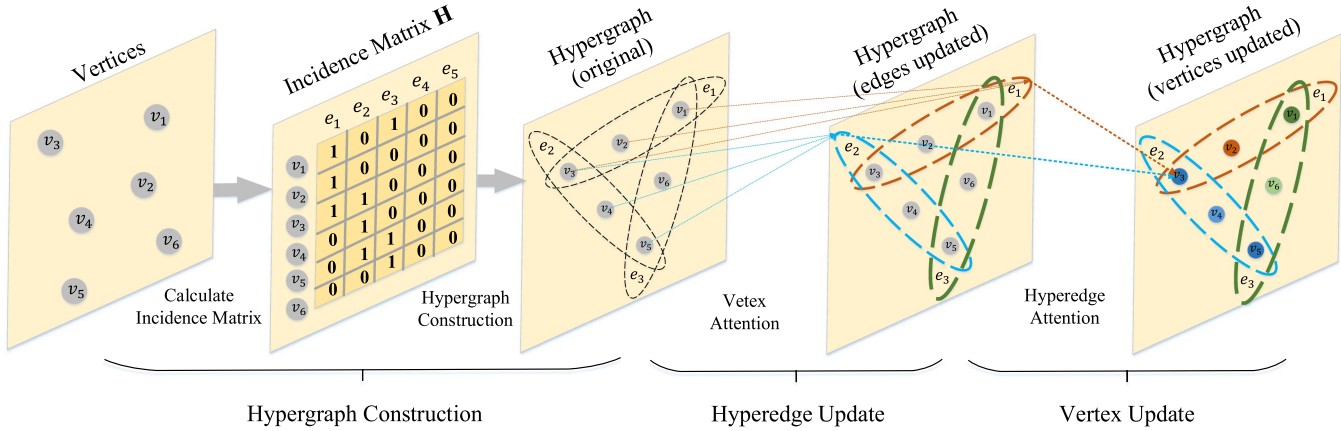


Fig. 3. Dynamic hypergraph learning algorithm. (a) Calculating incident matrix for building an original hypergraph. (b) Vertice features connected by a hyperedge are aggregated into the hyperedge by vertex attention. (c) Hyperedge representations are converged to their connected vertices via hyperedge attention.

entries defined as

$$\mathbf{H}_{ij} = \begin{cases} 1, & \text{if } \mathcal{V}_i \in \mathcal{E}_j \\ 0, & \text{if } \mathcal{V}_i \notin \mathcal{E}_j. \end{cases} \tag{5}$$

Dynamic hypergraph learning: According to the characteristics of multiobjective and multicategory in RS image, a well-designed dynamic hypergraph learning algorithm is introduced to automatically model the association relationships between multiple objects and cluster congeneric objects into a same hyperedge. As demonstrated in Fig. 3, the algorithm consists of three processes, i.e., hypergraph construction, hyperedge update, and vertex update.

Hypergraph construction: Given the feature matrix $\mathbf{M} = \{\mathbf{v}_i\}_{i=1}^{n}$ of RS image, each element/vector $\mathbf{v}_i$ in the matrix is regarded as a vertex $\mathcal{V}_i$. To reasonably connect the relevant features by a hyperedge, for each vertex, cosine distance $cosine(\mathbf{v}_i, \mathbf{v}_j)$ is employed as the measurement metric to cluster its nearest k vertices into a hyperedge $\mathcal{E}_i = \mathbf{v}_i \cup \{\mathbf{v}_m, \ldots, \mathbf{x}_{m+k-1}\}$. By this way, a hypergraph with n vertices and n hyperedges is formed, and the incidence matrix $\mathbf{H} \in \mathbb{R}^{n \times n}$ is square that the hyperedge weight default to 1.

Hyperedge update: After the hypergraph construction, the hyperedges need to be updated via gathering their connected vertex information. Based on this, we conceive a hyperedge update mechanism. Whereby the specific structure of hypergraph, each hyperedge is considered the intermediary of vertex feature updating. In other words, the update of vertices needs to first aggregate the information of its connected hyperedges rather than directly update with adjacent vertices.

With n vertices $\{\mathcal{V}_k\}_{k=1}^{n}$ connected by a hyperedge $\mathcal{E}_j$, hyperedge update aims to emphasis on the significant vertices by calculating the contribution of the vertices to the hyperedge and then aggregates them to update the hyperedge feature $\mathbf{e}_j$

$$\mathbf{e}_j^l = \sigma\left(\sum_{\mathcal{V}_k \in \mathcal{E}_j} \alpha_{jk} \mathbf{W}_v \mathbf{v}_k^{l-1}\right)$$

$$\alpha_{jk} = \frac{\exp\left(\mathbf{a}_v^{\mathrm{T}} \mathbf{z}_k\right)}{\sum_{\mathcal{V}_p \in \mathcal{E}_j} \exp\left(\mathbf{a}_v^{\mathrm{T}} \mathbf{z}_p\right)}$$

$$\mathbf{z}_k = \mathrm{ReLU}\left(\mathbf{W}^v \mathbf{v}_k^{l-1} + \mathbf{b}^v\right) \tag{6}$$

where σ is the nonlinearity activation, $\mathbf{W}_v$, $\mathbf{W}^v$, and $\mathbf{a}_v^{\mathrm{T}}$ are weight parameters, $\mathbf{b}^v$ denote the learnable bias.

Vertex update: Contrary to the hyperedge update process, vertex update is devised for converging the hyperedge information to update the connected vertices. Given a set of hyperedges $\mathcal{Y} = \{\cdots, \mathcal{E}_m, \ldots, \mathcal{E}_n, \cdots\}$ that are connected to a vertice $\mathcal{V}_k$, the update process of the vertice feature $\mathbf{v}_k$ can be formalized as

$$\mathbf{v}_k^l = \sigma\left(\sum_{\mathcal{E}_j \in \mathcal{Y}} \beta_{kj} \mathbf{W}_e \mathbf{e}_j^l\right)$$

$$\beta_{kj} = \frac{\exp\left(\mathbf{a}_e^{\mathrm{T}} \hat{\mathbf{z}}_j\right)}{\sum_{\mathcal{E}_p \in \mathcal{Y}} \exp\left(\mathbf{a}_e^{\mathrm{T}} \hat{\mathbf{z}}_p\right)}$$

$$\hat{\mathbf{z}}_j = \mathrm{ReLU}\left(\mathbf{W}^e \mathbf{e}_j^l\right) \tag{7}$$

where $\mathbf{v}_k^l$ refers to the updated feature of vertex $\mathcal{V}_k$ that gathers information from all of its connected hyperedges $\mathcal{Y}$. $\mathbf{W}_e$, $\mathbf{W}^e$, and $\mathbf{a}_e^{\mathrm{T}}$ are weight parameters, and $\mathbf{a}_e^{\mathrm{T}}$ is for the sake of measuring the significance of the hyperedges to vertex $\mathcal{V}_k$.

B. Feature Extraction

RS image feature extraction: As for RS images, we resize them to 256×256 pixels, and randomly crop and rotate the images to extend the training samples.To avoid over-fitting due to the deep backbone, we apply the ResNet-18 model [32] pretrained on the ImageNet dataset [33], [34], [35] following [36] to extract last convolution layer's feature maps with size $512 \times 8 \times 8$, i.e.,

$$\mathbf{v}^g = \mathrm{ResNet}(I) \tag{8}$$

where $\mathbf{v}^g$ denotes global feature of RS image.

Although $\mathbf{v}^g$ contains the global information of the RS image, it still encounters the bottleneck in accurately expressing the multiscale properties of objects. To solve the above issue, we follow [4] to up-sampling the feature maps of the first three layers of ResNet-18 and concatenate these feature maps together as

low-level RS image features $\mathbf{V}_I^{low}$. In addition, the feature maps of the last two layers are sampled and connected as high-level RS image features $\mathbf{V}_I^{high}$.

Query text feature extraction: A query text can be regarded as a word sequence $\{x_i\}_{i=1}^n$. Considering the temporal information in query text, we first exploit BiGRU [37] as a text encoder to refine each embedded word $e(x_i)$ from forward and backward directions. Afterward, the generated bidirectional hidden states are averaged to avoid dimension amplification, as follows:

$$\begin{aligned}\overrightarrow{\mathbf{h}}_t &= \overrightarrow{\mathrm{GRU}}(e(x_i), \overrightarrow{\mathbf{h}}_{t-1})\\ \overleftarrow{\mathbf{h}}_t &= \overleftarrow{\mathrm{GRU}}(e(x_i), \overleftarrow{\mathbf{h}}_{t-1})\\ \mathbf{h}_t &= (\overrightarrow{\mathbf{h}}_t + \overleftarrow{\mathbf{h}}_t)/2 \end{aligned} \tag{9}$$

where $\mathbf{h}_t$ refers to the hidden state of word x_i containing forward and reverse query text information. All the hidden states $\{\mathbf{h}_t\}_{t=1}^n$ compose of the features of query text $\mathbf{V}_T \in \mathbb{R}^{n\times d}$.

C. Multiscale RS Image Hypergraph Networks

To handle the multiscale properties of objects in RS images, we develop the multiscale RS image hypergraph networks. Specifically, based on the extracted high-level RS image features $\mathbf{V}_I^{high}$ containing high-level semantic information, a high-level RS image hypergraph network is established to capture the relationships between high-level objects, cluster similar high-level objects into a hyperedge through dynamic hypergraph learning, and promote the information interaction between the objects. Similarly, the low-level RS image hypergraph network is also constructed to determine the relationships between low-level objects.

1) High-Level RS Image Hypergraph Network: The extracted high-level RS image features $\mathbf{V}_I^{high}$ are taken as the high-level hypergraph vertices $\mathcal{V}_I^{high}$, and the hypergraph construction method in IV-A is exploited to form the relevant vertices into hyperedges $\mathcal{E}_I^{high}$ through cosine similarity, and then the corresponding incidence matrix $\mathbf{H}_I^{high}$ is calculated

$$\mathcal{HG}_I^{high} = (\mathcal{V}_I^{high}, \mathcal{E}_I^{high}, \mathbf{W}_I^{high}) \leftarrow \mathrm{HG}(\mathbf{V}_I^{high}) \tag{10}$$

where $\mathcal{HG}_I^{high}$ refers to the constructed high-level RS image hypergraph, $\mathbf{W}_I^{high}$ is the weight of the hyperedge default to 1, and $HG(\cdot)$ is the method of constructing the hypergraph.

The constructed hypergraph models the relationships between high-level objects without making an object learn knowledge from other related objects. Thus, dynamic hypergraph learning in IV-A is adopted for hypergraph iterative updating. Specifically, the hyperedge update mechanism $\mathrm{HyperedgeUpdate}(\cdot)$ aggregates the relevant object features into their connected hyperedges following the contribution of these objects to the hyperedges

$$\mathbf{E}_I^{high} = \mathrm{HyperedgeUpdate}(\mathbf{V}_I^{high}, \mathbf{H}_I^{high}) \tag{11}$$

where $\mathbf{E}_I^{high}$ is the updated hyperedge features.

Since various hyperedges connect a vertex/object feature, the hyperedge that gathers relevant high-level object features is regarded as a relay station further to feedback the hyperedge information to the connected vertices

$$\hat{\mathbf{V}}_I^{high} = \mathrm{VertexUpdate}(\mathbf{E}_I^{high}, \mathbf{H}_I^{high}) \tag{12}$$

where $\hat{\mathbf{V}}_I^{high}$ is the updated vertex features that contains all high-level object information in RS image. Finally, the above process is repeated multiple times to obtain sufficient information for high-level objects.

2) Low-Level RS Image Hypergraph Network: Similar to the process of high-level RS image hypergraph network, first, the low-level RS image features $\mathbf{V}_I^{low}$ are considered as the vertices $\mathcal{V}_I^{low}$ of the low-level RS image hypergraph network. Thereafter, the hypergraph construction method $\mathrm{HG}(\cdot)$ is employed to cluster the relevant low-level RS objects into hyperedges $\mathcal{E}_I^{low}$, and the incidence matrix $\mathbf{H}_I^{low}$ is calculated. Eventually, the hyperedge $\mathrm{HyperedgeUpdate}(\cdot)$ and vertex $\mathrm{VertexUpdate}(\cdot)$ update mechanisms are repeatedly exploited to promote the iterative interaction of hyperedge and vertex features, to achieve the fusion of the most relevant information of other objects for each low-level object. All formulation processes are as follows:

$$\begin{aligned}\mathcal{HG}_I^{low} &= (\mathcal{V}_I^{low}, \mathcal{E}_I^{low}, \mathbf{W}_I^{low}) \leftarrow \mathrm{HG}(\mathbf{V}_I^{low})\\ \mathbf{E}_I^{low} &= \mathrm{HyperedgeUpdate}(\mathbf{V}_I^{low}, \mathbf{H}_I^{low})\\ \hat{\mathbf{V}}_I^{low} &= \mathrm{VertexUpdate}(\mathbf{E}_I^{low}, \mathbf{H}_I^{low}) \end{aligned} \tag{13}$$

where $\mathcal{HG}_I^{low}$, $\mathbf{E}_I^{low}$, and $\hat{\mathbf{V}}_I^{low}$ represent the constructed low-level RS image hypergraph, the updated hyperedges and vertices, respectively.[1]

Algorithm 1: Multiscale RS Image Hypergraph Networks.

Input: high-level and low-level RS image features $\mathbf{V}_I^{high}$, $\mathbf{V}_I^{low}$.
Output: the updated vertex features $\hat{\mathbf{V}}_I^{high}$ and $\hat{\mathbf{V}}_I^{low}$ that contains all high-level and low-level object information in RS image, respectively

High-Level RS Image Hypergraph Network.
  Step 1. Construction:
    $\mathcal{HG}_I^{high} = (\mathcal{V}_I^{high}, \mathcal{E}_I^{high}, \mathbf{W}_I^{high}) \leftarrow \mathrm{HG}(\mathbf{V}_I^{high})$
  Step 2. Update:
    $\mathbf{E}_I^{high} = \mathrm{HyperedgeUpdate}(\mathbf{V}_I^{high}, \mathbf{H}_I^{high})$
    $\hat{\mathbf{V}}_I^{high} = \mathrm{VertexUpdate}(\mathbf{E}_I^{high}, \mathbf{H}_I^{high})$
Low-Level RS Image Hypergraph Network.
  Step 1. Construction:
    $\mathcal{HG}_I^{low} = (\mathcal{V}_I^{low}, \mathcal{E}_I^{low}, \mathbf{W}_I^{low}) \leftarrow \mathrm{HG}(\mathbf{V}_I^{low})$
  Step 2. Update:
    $\mathbf{E}_I^{low} = \mathrm{HyperedgeUpdate}(\mathbf{V}_I^{low}, \mathbf{H}_I^{low})$
    $\hat{\mathbf{V}}_I^{low} = \mathrm{VertexUpdate}(\mathbf{E}_I^{low}, \mathbf{H}_I^{low})$

D. Textual Fully Connected Graph Network

The query text is composed of various words, and different terms are of varying importance to the retrieval task. For instance, some entities (such as "*plane*" in Fig. 2) play a decisive

[1]Noting that the procedure of multiscale RS image hypergraph networks is summarized in Algorithm 1.

role in RS image retrieval. In addition, there are internal relationships between words, e.g., "*gray*" for modifying "*plane*." Therefore, to model the relationships between arbitrary pairwise terms and capture the contribution of other words to a word, we elaborate a textual fully connected graph network.

The text features $\mathbf{V}_T$ extracted by BiGRU are utilized as the graph's vertices, and an edge connects arbitrary pairwise vertices to build the fully connected graph. The mutual contribution of words determines the weight of the edge. Given a fully connected graph with n vertices, the significance of an edge/contribution can be calculated as follows:

$$\begin{aligned} s_k &= \mathbf{w}_k^T \mathbf{m}_k + \mathbf{w}_v^T \mathbf{v} + b_k \\ a_j &= exp(s_j)/\textstyle\sum_{k=1}^{n} exp(s_k) \end{aligned} \tag{14}$$

where $\mathbf{v} \in \mathbf{V}_T$ is a vertex in the fully connected graph, $\mathbf{m}_k \in \mathbf{V}_T$ is a vertex sharing the same edge with $\mathbf{v}$, and a_j represents the weight of the edge connecting $\mathbf{v}$ and $\mathbf{m}_k$, which is used to measure the significance of $\mathbf{m}_k$ to $\mathbf{v}$.

With the weight of each edge in the fully connected graph, each vertex/word aggregates information from other vertices/words according to their importance

$$\hat{\mathbf{v}}_i^t = \sum_{j=1}^{n} a_j \mathbf{m}_j \tag{15}$$

where $\hat{\mathbf{v}}_i^t$ is the ith vertex feature that converges all the word information in query text. To make the words learn more fine-grained information from other words, we repeat the above process multiple times and receive the query text feature matrix $\hat{\mathbf{V}}_T = \{\hat{\mathbf{v}}_i^t\}_{i=1}^n$.

E. Cross-Modal Matching

There is usually a coreference relationship between the objects in the RS image and the entities in query text, as shown in Fig. 1, the two planes in the RS image correspond to "*grey plane*" and "*blue plane*" in the query text, respectively. To grasp the coreference relationship and improve the retrieval accuracy, we conceive a cross-modal matching method, which is divided into two modules, i.e., dynamic multiscale feature fusing and image-induced multimodal fusing, as shown in Figs. 4 and 5 respectively.

1) Dynamic Multiscale Feature Fusing: Given the updated high-level features $\hat{\mathbf{V}}_I^{high}$, low-level features $\hat{\mathbf{V}}_I^{low}$, and global feature $\mathbf{v}^g$, the intention of the module is to dynamically fuse the above three features to solve the multiscale problem of RS images. Specifically, for the updated high-level features, the convolution neural network(CNN) with built-in 1×1 kernel $\mathrm{Conv}_{1\times1}(\cdot)$ is adopted for preliminary encoding and is activated by $\mathrm{ReLu}(\cdot)$. Then, the average pooling $\mathrm{Avg}(\cdot)$ is used to reduce the feature dimension, and finally, the convolutional neural network is adopted again for deep encoding, as follows:

$$\begin{aligned} \hat{\mathbf{v}}_I^{high} &= \mathrm{Avg}(\mathrm{ReLu}(\mathrm{Conv}_{1\times1}(\hat{\mathbf{V}}_I^{high}))) \\ \dot{\mathbf{v}}_I^{high} &= \mathrm{Conv}_{1\times1}(\hat{\mathbf{v}}_I^{high}) \end{aligned} \tag{16}$$

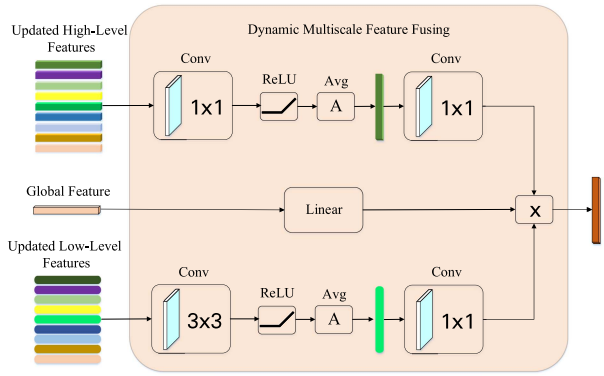


Fig. 4. Dynamic multiscale feature fusing module. Its purpose is to dynamically fuse the updated high-level, low-level, and global features to solve the multiscale problem of RS image.

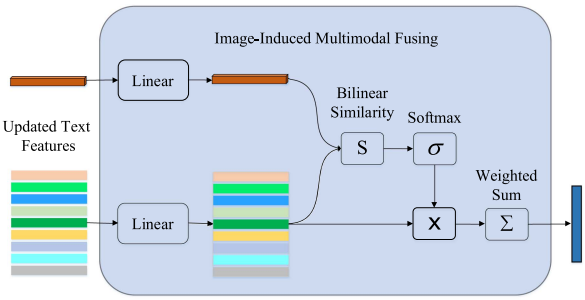


Fig. 5. Image-induced multimodal fusing module. It aims to establish the feature association between RS image and query text.

where $\dot{\mathbf{v}}_I^{high}$ represents the condensed feature containing large-scale object information.

The processing of the updated low-level features is consistent with that of the high-level features. The only difference is that 3×3 convolution kernel is utilized instead of 1×1 in CNN to ensure the consistency of feature dimensions and facilitate subsequent fusion, that is

$$\begin{aligned} \hat{\mathbf{v}}_I^{low} &= \mathrm{Avg}(\mathrm{ReLu}(\mathrm{Conv}_{3\times3}(\hat{\mathbf{V}}_I^{low}))) \\ \dot{\mathbf{v}}_I^{low} &= \mathrm{Conv}_{1\times1}(\hat{\mathbf{v}}_I^{low}) \end{aligned} \tag{17}$$

where $\dot{\mathbf{v}}_I^{low}$ represents the fine-grained feature covering the information of small-scale objects.

Finally, $\dot{\mathbf{v}}_I^{high}$, $\dot{\mathbf{v}}_I^{low}$, and $\mathbf{v}^g$ are multiplied at the element level to obtain the final representation $\dot{\mathbf{v}}_I$ of the RS image.

2) Image-Induced Multimodal Fusing: To establish the feature association between RS image and query text, we design an image-induced multimodal fusing module to guide the RS image feature that integrates the high-level, low-level, and global features to locate the relevant or significant features in the query text. First, the updated RS image $\dot{\mathbf{v}}_I$ and text features $\hat{\mathbf{V}}_T = \{\hat{\mathbf{v}}_i^t\}_{i=1}^n$ are projected by affine transformation. Afterward, the bilinear similarity is exploited to measure the correlation between them. Finally, the features that match the RS image in the query text are weighted and summed to obtain a new feature.

This yields

$$s_k = \hat{\mathbf{v}}_i^t \mathbf{W}_v^T \dot{\mathbf{v}}_I + b_k$$
$$a_j = exp(s_j) / \sum_{k=1}^{n} exp(s_k)$$
$$\dot{\mathbf{v}}_T = \sum_{j=1}^{n} a_j \hat{\mathbf{v}}_i^t \tag{18}$$

where $\dot{\mathbf{v}}_T$ represents the query text feature condensed according to the correlation intensity with RS image.

F. Triplet Loss

We choose the triplet loss as the loss function following [4] for increasing the distance between the sample and its corresponding negative samples and making the distance between the sample and its positive samples as close as possible:

$$L(I,T) = \sum_{\widehat{T}} [\alpha - S(I,T) + S(I,\widehat{T})]_+ + \sum_{\widehat{I}} [\alpha - S(I,T) + S(\widehat{I},T)]_+ \tag{19}$$

where α represents the margin, $[x]_+ = \max(x, 0)$, and $S(I,T)$ represent the similarity of the RS image and text.

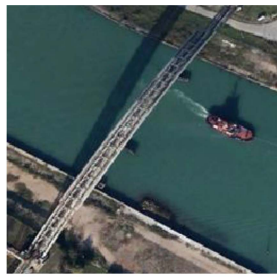
V. Experiments

A. Settings

Datasets: In this article, we select two public datasets (please refer to Fig. 6), i.e., RSICD and RSITMD datasets, to verify the model's effectiveness. RSICD dataset [31] is a large-scale and diverse RS image caption dataset containing 10 921 images and 30 scenes, and has become the preference for RS image caption tasks. RSITMD dataset [4] is a fine-grained dataset dedicated to RS cross-modal text-image retrieval. Some images in this dataset are selected from the RSICD dataset, while others are from Google Earth, including 4743 images, 23 715 captions, and 24 scenes.

Settings: All the experiments are performed on pyTorch [41], running on a Tesla V100 GPU with 32G memory. For the RS image, the image embedding dimension size is 512. The word embedding dimension is set to 256, and the hidden layer of the BiGRU is set to 512. In this manner, the dimension of the RS image and query text can be kept consistent for the subsequent feature interactions. In terms of hypergraph construction, for high-level and low-level RS image hypergraph networks, the number of vertices connected by a hyperedge is fixed at 6. In pursuit of a balance of complexity and efficiency, the update times of textual fully connected graph network and multiscale RS image hypergraph networks are set to twice. Adam is selected as the optimizer to train the network up to 50 epochs with the batch size set to 128. During the training period, the learning rate was adjusted to $1e^{-4}$, and was decreased by 0.7 every 5 epochs. For evaluation indicators, R@K (K=1, 5, and 10) and mR are applied to evaluate the performance of the proposed model. R@K represents the percentage of ground truth that appears in topK results. Moreover, to reasonably evaluate the model's overall performance, we also use the average of six recall rates to obtain mR.

(a)



Caption1: The narrow arch bridge spans a straight river with red boats and three black boats.
Caption2: A red boat is sailing in the river over which is an briage.
Caption3: The bridge crosses the jade green river.
Caption4: A narrow arch bridge spans a straight river with a red boat and three black boats.
Caption5: Some green plants are in two sides of a river with a bridge over it and a red boat in it.

(b)

Caption1: Four planes were disappeared at the airport.
Caption2: Four planes were scattered at the airport.
Caption3: Four planes were scattered throughout the airport.

Fig. 6. Illustration of the samples of RSICD and RSITMD datasets. Each RS image is attached with several descriptions. (a) The example of RSICD dataset. (b) The example of RSITMD dataset.

B. Baselines

We select several previous state-of-the-art models, which are specially oriented to image-text matching, as the comparison baselines to verify the effectiveness of our model, as follows.

VSE++ [42]: In [42], image information and text information are embedded into the same space by using convolution network and recursive network, and utilizing triple loss to train image-text matching model.

SCAN [19]: SCAN, which on the foundation of VSE++, applies faster RCNN [50] to extract image features and attempts to align the corresponding objects in the RS image and query text.

CAMP [43]: CAMP introduces an adaptive message passing mechanism to control the flow of information transmission between different modes adaptively and uses the fused features to calculate the matching degree of image and text.

MTFN [36]: MTFN leverages rank decomposition to construct a multimodal fusion network for calculating the distance of embedded features.

TABLE I
EXPERIMENTS OF SENTENCE-TO-IMAGE RETRIEVAL AND IMAGE-TO-SENTENCE RETRIEVAL ON RSICD TEST SET

RSICD dataset	Sentence-to-Image Retrieval			Image-to-Sentence Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	3.38	9.51	17.46	2.82	11.32	18.10	10.43
SCAN t2i	4.39	10.90	17.64	3.91	16.20	26.49	13.25
SCAN i2t	5.85	12.89	19.84	3.71	16.40	26.73	14.23
CAMP-Triplet	5.12	12.89	21.12	4.15	15.23	27.81	14.39
CAMP-BCE	4.20	10.24	15.45	2.72	12.76	22.89	11.38
MTFN	5.02	12.52	19.74	4.90	17.17	29.49	14.81
HyperMatch	**7.14**	**20.04**	**31.02**	**6.08**	**20.37**	**33.82**	**19.75**

TABLE II
EXPERIMENTS OF SENTENCE-TO-IMAGE RETRIEVAL AND IMAGE-TO-SENTENCE RETRIEVAL ON RSITMD TEST SET

RSITMD dataset	Sentence-to-Image Retrieval			Image-to-Sentence Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	10.38	27.65	**39.60**	7.79	24.87	38.67	24.83
SCAN t2i	10.18	**28.53**	38.49	**10.10**	28.98	43.53	26.64
SCAN i2t	11.06	25.88	39.38	9.82	29.38	42.12	26.28
CAMP-Triplet	**11.73**	26.99	38.05	8.27	27.79	44.34	26.20
CAMP-BCE	9.07	23.01	33.19	5.22	23.32	38.36	22.03
MTFN	10.40	27.65	36.28	9.96	31.37	45.84	26.92
HyperMatch	**11.73**	28.10	38.05	9.16	**32.31**	**46.64**	**27.66**

C. Comparisons

Table I summarizes the experimental results of HyperMatch on the RSICD dataset. It can be seen from Table I that the proposed HyperMatch achieves significantly improved performance compared with the state-of-the-art models in both sentence-to-image and image-to-sentence retrieval. In mR metric, HyperMatch outperforms the optimal CAMP-Triplet model by 5.36%. In using a sentence as a query to retrieve RS images, HyperMatch improves by 2.02%, 7.15%, and 9.90% in R@1, R@5, and R@10 metrics, respectively. The experimental results demonstrate that given a query sentence, HyperMatch can better match RS images related to a sentence. At the same time, retrieving sentences whereby an RS image improves performance by 1.93%, 5.14%, and 6.01%, respectively, verifies image-to-sentence retrieval effectiveness. In addition, the experimental results on the RSITMD dataset (please refer to Table II) also show that the performance of HyperMatch receives competitive performance in most indicators, e.g., it exceeds the MTFN model with the best performance by 0.74% in mR indicator.

HyperMatch can achieve superior performance on the selected datasets, mainly including the following aspects. On the one hand, aiming at the situation of multitypes, uneven distribution, and multiscales of objects in an RS image, the high-level and low-level RS image hypergraph networks are well-designed to model the relationships between objects at different scales, and cluster the features of similar objects into a same hyperedge. On the other hand, an undirected fully connected graph network is conceived to quantify the mutual contribution of words in the query text. Furthermore, the constructed cross-modal matching module learns the coreference relationships between the objects in the RS image and the entities in the query text.

D. Ablations

To explore the importance of each pivotal component of the proposed model, ablation experiments are performed in the selected two datasets in this section. The results are summarized in Tables III and IV.

From the experimental results in Table III, it can be observed that when the high-level RS image hypergraph network is eliminated from the model, the performance of in sentence-to-image retrieval task is decreased by 0.6%, 1.42%, and 1.56%, respectively on the R@1, R@5, R@10 indicators. In the image-to-sentence retrieval task, the performance also declined by 0.59%, 0.81%, and 1.5%, respectively. The main reason for the performance degradation is that the objects in RS images possess the characteristics of multitype and multiscale. Once the high-level RS image hypergraph network is removed, the relationships between large-scale objects will not be modeled, nor can the similar type of large-scale objects be clustered into the same hyperedges through dynamic hypergraph learning to realize the information interaction between objects. In addition,

TABLE III
ABLATION EXPERIMENTS ON RSICD TEST SET

RSICD dataset							
Configuration	Sentence-to-Image Retrieval			Image-to-Sentence Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
w/o High-Level RS Image Hypergraph Network	6.54	18.62	29.46	5.49	19.56	32.32	18.66
w/o Low-Level RS Image Hypergraph Network	**7.77**	18.02	27.53	4.99	**20.53**	33.41	18.71
w/o Textual Fully Connected Graph Network	5.58	17.84	28.45	5.25	18.38	30.77	17.71
w/o Dynamic Multiscale Feature Fusing	6.67	17.93	27.99	5.12	19.46	32.09	18.21
w/o Dynamic Image-Induced Multimodal Fusing	5.94	17.20	27.90	4.92	18.75	30.83	17.59
HyperMatch (full)	7.14	**20.04**	**31.02**	**6.08**	20.37	**33.82**	**19.75**

TABLE IV
ABLATION EXPERIMENTS ON RSITMD TEST SET

RSITMD dataset							
Configuration	Sentence-to-Image Retrieval			Image-to-Sentence Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
w/o High-Level RS Image Hypergraph Network	9.40	24.66	36.39	8.67	31.26	47.94	26.39
w/o Low-Level RS Image Hypergraph Network	10.39	26.10	36.50	9.02	31.90	47.61	26.92
w/o Textual Fully Connected Graph Network	9.29	23.78	35.61	8.31	28.45	44.00	24.91
w/o Dynamic Multiscale Feature Fusing	12.83	25.66	34.51	9.02	28.27	43.14	25.57
w/o Dynamic Image-Induced Multimodal Fusing	7.96	23.23	34.07	8.62	29.38	42.92	24.36
HyperMatch (full)	**11.73**	28.10	38.05	9.16	**32.31**	**46.64**	**27.66**

an interesting phenomenon is that the performance degradation after removing the module is not particularly obvious, mainly because the low-level RS image hypergraph network, which can model the relationship between small-scale objects, has not been eliminated, compensating for the performance degradation caused by the elimination of high-level RS image hypergraph network. Analogously, while eliminating the low-level RS image hypergraph network, the performance on mR metric decreases by 1.09%, which verifies the ability of the module to model the relationships between small-scale objects and aggregate information between objects based on the relationships. Noting that the improvement brought by low-level hypergraphs is not as significant as that brought by high-level hypergraphs, especially on image-to-text retrieval. We attribute it to the fact that high-level hypergraphs absorb high-level semantic information; compared with more implicit and localized underlying semantic information, the relational modeling of high-level information plays a more important role for the model to recognize global image information, which is more conducive to be mined by text features.

The textual fully connected graph network regards words as vertices and the contribution of words to each other as edges. By utilizing the self-attention mechanism on the fully connected graph network, each word can aggregate information according to the importance of other words. Therefore, when the module is removed, the performance decreases significantly, e.g., on the sentence-to-image retrieval task, in terms of R@1, R@5, R@10 indicators, the performance decreased by 1.56%, 2.2%, and 2.57%, respectively.

The original intention of the dynamic multiscale feature fusing module is to combine the large-scale and small-scale object features learned from high-level and low-level RS image hypergraph networks. Also, the global feature containing global information of the RS image is dynamically fused to deal with multiscale problems of objects. Thus, after removing this module, the performance decreased significantly, e.g., by 1.54% in mR metric.

To establish feature association between remote sensing image and query text, a dynamic image-induced multimodal fusing module is designed to guide the positioning of the most relevant or significant features in the query text by integrating the high-level, low-level, and global features of RS image. After the module is removed, the performance degrades obviously, e.g., in sentence-to-image and image-to-sentence retrieval, R@1, R@5, and R@10 decrease by (1.2%, 2.84%, 3.3%) and (1.16%, 1.62%, 2.99%), respectively, which illustrates the importance of this module for cross-modal remote sensing image retrieval.

Table IV shows the ablation experiment results on the RSITMD dataset, from which it can be observed that the elimination of the five vital components aforementioned decreases

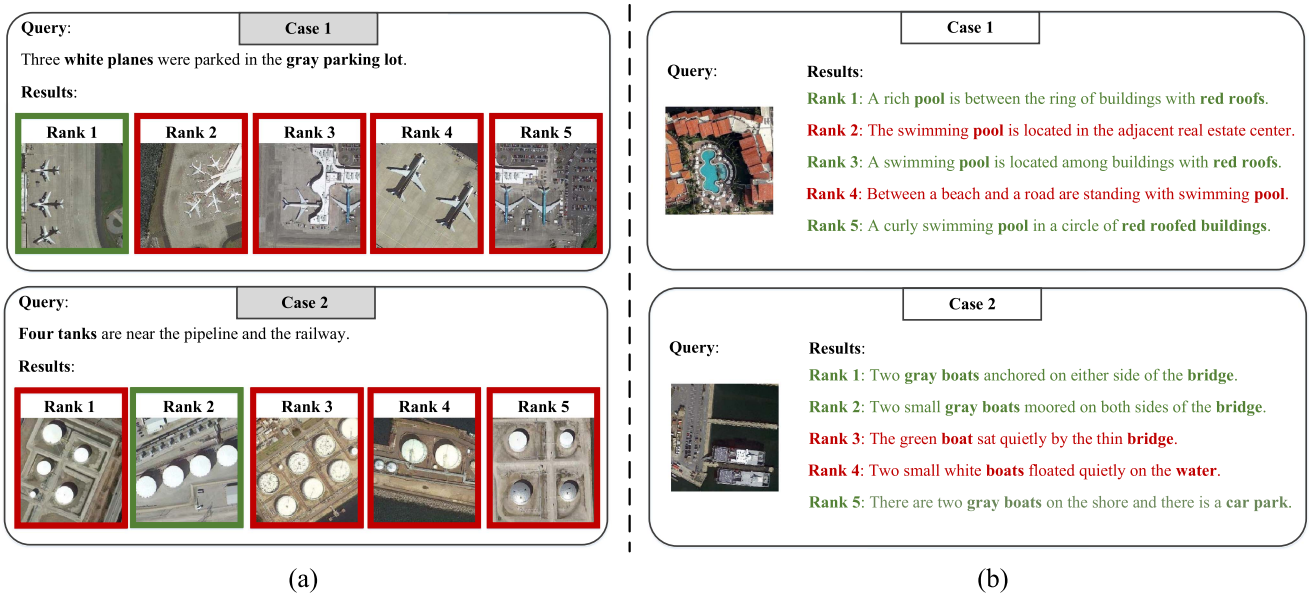


Fig. 7. Examples of selected text-to-image retrieval and image-to-text retrieval. (a) Two cases of retrieving RS images by query text, showing the top five most relevant retrieval results, of which the green box frames the ground truth. (b) It demonstrates two cases of treating RS images as queries to match relevant captions and also returns the top five results. The results with green font represent the ground truths. (a) Text-to-image retrieval. (b) Image-to-text retrieval.

the mR metric by 1.27%, 0.74%, 2.75%, 2.09%, and 3.3%, respectively. The phenomenon and analysis of performance degradation are the same as those in Table III, and will not be described here for simplicity. Note that the decline is most significant when removing the textual fully connected graph network and dynamic image-induced multimodal fusing, illustrating the necessity of self-attention-based weighted aggregation of entity information in text and the effectiveness of learning feature associations between RS image and query text.

E. Case Study

To intuitively demonstrate the proposed model's performance in text-to-image and image-to-text retrieval, we select several examples (as shown in Fig. 7) for analysis of the two tasks.

Case 1 in Fig. 7(a) shows the retrieved five RS images that are most relevant to the content of query text, that is, "*Three white planes were parked in the gray parking lot* ." From the retrieved results, we can see that the proposed HyperMatch can accurately retrieve the best suitable RS image (i.e., Rank 1) consistent with the ground truth according to the query text, which shows the superior ability of the model in text-image retrieval. In addition, the remaining retrieved RS images are also highly related to the content of the query text. In particular, the objects in the RS images are in accordance with the keywords of the query text (e.g., "*planes*" and "*parking lot*"), which verifies the rationality of the retrieval results. For the retrieval results of Case 2, the ground truth is ranked second. Even so, the content of the Rank 1 is highly similar to that of the ground truth, e.g., there are four water tanks in both RS images. The remaining three retrieved RS images also contain the key entities in the query text, that is, the "*tank*, " which further verifies the ability of the model to retrieve RS images by query text.

Fig. 7(b) illustrates the retrieved relevant captions according to an RS image. It can be found from Case 1 that all three ground truths are included in the five top-ranked results retrieved through an RS image, and the keyword "*pool*" in the retrieved nonground-truths, i.e., Rank 2 and Rank 4, is also in keeping with the object in the RS image. The retrieval situations in Case 2 are similar to Case 1. On the one hand, all three ground truths are retrieved. On the other hand, the remaining two captions that are not within the ground truths are also related to the content of the RS image, such as "*boat*" and "*bridge*" in Rank 3 and "*boats*" and "*water*" in Rank 4. Two cases demonstrate the competitive performance of the proposed model on image-to-text retrieval.

F. Visualization

To visually show whether the proposed model can accurately locate critical components (such as object positions) in RS images according to query text, we verify this capability in the semantic localization task, which refers to locating the regions that best match the query text in a large scene. Following the work proposed in [4], we first use the various sliding window to cut the large scene image to maintain the multiscale characteristics of the object. Afterward, the similarity between each patch obtained after segmentation and the query text is calculated to form a probability map. After that, the obtained probability distributions are combined, and the median filter is utilized to remove the impact noise in the probability map to ensure that the results are robust. Finally, the probability map is fused with the original RS image to generate a located image that can intuitively display the semantic positioning ability of the model. Fig. 8 illustrates the selected two examples.

Example (a) in Fig. 8 aims to locate football grounds surrounded by cars and houses from a large scene RS image. From

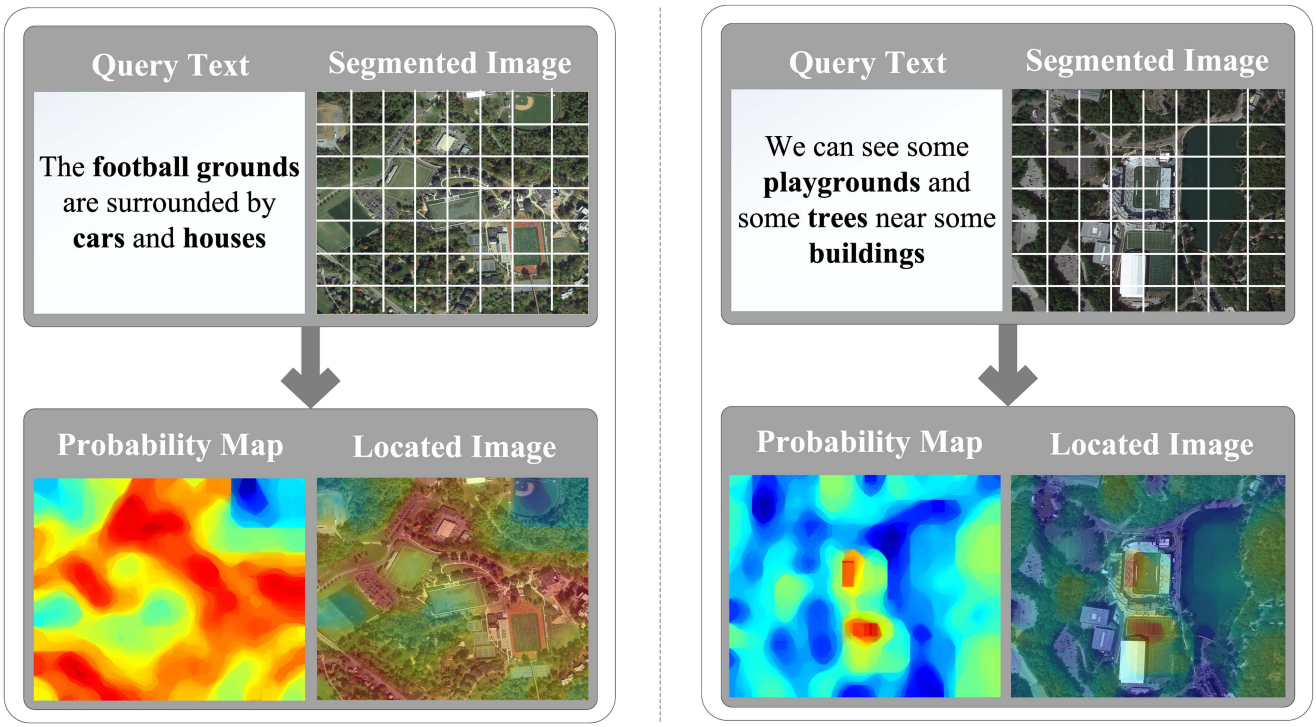


Fig. 8. Visualization of semantic localization results. Either (a) or (b) contains the query text, the segmented RS image, the probability map, and the located RS image. Among them, the segmented RS image is cut into multiple patches of different scales in various ways according to the method proposed in [4]. Probability map refers to the probability distribution heat map formed by concatenating the similarity between each patch in the segmented image with the query text. The located image is generated by "fusing" the probability map with the original RS image, which is convenient for visually discovering the places in the RS image that are related to the query text.

the located image, we can observe that several football grounds in the RS image are located. From the probability map, one can also find that the parts with high probability (coloured in orange and red) form a "circle" shape, which perfectly corresponds to the keyword "*surrounded*" in the query text, demonstrating that the proposed model can not only locate the objects in the large scene RS image according to the query text, the spatial relationships between the objects can also be understood. In example (b), we attempt to locate the playgrounds and trees near the buildings. From the located image, it can be found that the two playgrounds are accurately located. Also, the two playgrounds are given the highest probability (the deepest colour) in the probability map, which confirms the model's capability in the semantic localization task.

VI. Conclusion

Cross-modal RS image retrieval is to retrieve RS images using other modalities such as text or query other modalities via RS images. The multiscale and multicategory characteristics of objects in RS images make it difficult to match the short query text, further restricting the performance of RS image retrieval. The hyperedge in a hypergraph can connect the arbitrary number of vertices and have significant advantages in representing high-order complex relationships in data. In recent years, hypergraph learning has attracted extensive attention and developed rapidly. Therefore, this article introduces it into cross-modal RS image retrieval and proposes a HyperMatch to realize the accurate matching between RS image and query text by learning the spatial relationships between objects in RS image, the contribution relationships between words in query text, and the corresponding relationships between the objects in RS image and the entities in query text.

Specifically, high-level and low-level RS image hypergraph networks are constructed, respectively, to model the relationships between objects of different scales and cluster similar object features into the same hyperedge. For the construction of a hypergraph, cosine similarity is utilized as the metric to measure the correlation of features in the RS image. For the dynamic update of a hypergraph, vertex attention, and hyperedge attention are designed to realize the dynamic alternating update of vertices and hyperedges. Experiments on the published RSICD and RSITMD datasets verify the effectiveness of HyperMatch in cross-modal RS image retrieval. In the future, we will explore the feasibility of applying hypergraph learning to other multimodal tasks, such as modeling high-order relationships within and between modalities.

References

[1] C.-R. Shyu, M. Klaric, G. J. Scott, A. S. Barb, C. H. Davis, 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.

[2] Y. Chen and X. Lu, "A deep hashing technique for remote sensing image-sound retrieval," *Remote Sens.*, vol. 12, no. 1, pp. 1–84, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/1/84
[3] 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.
[4] 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, pp. 1–19, May 2021, Art. no. 4404119.
[5] 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.
[6] G. Hoxha, F. Melgani, and J. Slaghenauffi, "A new CNN-RNN framework for remote sensing image captioning," in *Proc. Mediterranean Middle-East Geosci. Remote Sens. Symp.*, 2020, pp. 1–4.
[7] 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.
[8] 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.
[9] R. Hang, X. Qian, and Q. Liu, "Cross-modality contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 1, pp. 1–12, Jul. 2022, Art. no. 5532812.
[10] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 1, pp. 1–12, Sep. 2022, Art. no. 5412012.
[11] R. Hang, Q. Liu, and Z. Li, "Spectral super-resolution network guided by intrinsic properties of hyperspectral imagery," *IEEE Trans. Image Process.*, vol. 30, no. 1, pp. 7256–7265, Aug. 2021.
[12] 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.
[13] 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.
[14] W. Xiong, Z. Xiong, Y. Cui, and Y. Lv, "A discriminative distillation network for cross-source remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 4, pp. 1234–1247, Mar. 2020.
[15] W. Xiong, Y. Lv, X. Zhang, and Y. Cui, "Learning to translate for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4860–4874, Jul. 2020.
[16] Z. Yuan et al., "A lightweight multi-scale crossmodal text-image retrieval method in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 6, pp. 1–19, Oct. 2021, Art. no. 5612819.
[17] D. Wang, L. Wang, S. Song, G. Huang, and A. Du, "Fusion layer attention for image-text matching," *Neurocomputing*, vol. 442, pp. 249–259, 2021.
[18] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1908–1917.
[19] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Comput. Vis. - ECCV 2018: 15th Eur. Conf.*, Munich, Germany, Sep. 8–14, 2018, pp. 212–228. [Online]. Available: https://doi.org/10.1007/978-3-030-01225-0_13
[20] N. Sarafianos, X. Xu, and I. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2020, pp. 5813–5823.
[21] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10918–10927.
[22] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2623–2631.
[23] N. Messina, F. Falchi, A. Esuli, and G. Amato, "Transformer reasoning network for image-text matching and retrieval," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 5222–5229.
[24] M.-D. Nguyen, B. T. Nguyen, and C. Gurrin, "A deep local and global scene-graph matching for image-text retrieval," in *New Trends Intell. Softw. Methodologies, Tools Techn. – Proc. 20th Int. Conf. New Trends Intell. Softw. Methodologies, Tools Techn.*, in Frontiers in Artificial Intelligence and Applications, H. Fujita and H. Pérez-Meana, Eds., vol, 337, SoMeT 202, Cancun, Mexicovol. IOS Press, Sep. 21–23, 2021, pp. 510–523. [Online]. Available: https://doi.org/10.3233/FAIA210049
[25] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7181–7189.
[26] R. Zhang, Y. Zou, and J. Ma, "Hyper-sagnn: A self-attention based graph neural network for hypergraphs," 2019.
[27] J. Zhang, Y. Chen, X. Xiao, R. Lu, and S.-T. Xia, "Learnable hypergraph Laplacian for hypergraph learning," in *Proc. ICASSP 2022-2022 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4503–4507.
[28] M. Xueqi et al., "Hypergraph p -Laplacian regularization for remotely sensed image recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1585–1595, Mar. 2019.
[29] Y. Duan, H. Huang, and Y. Tang, "Local constraint-based sparse manifold hypergraph learning for dimensionality reduction of hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 613–628, Jan. 2021.
[30] X. Wei, L. Cai, B. Liao, and T. Lu, "Local-view-assisted discriminative band selection with hypergraph autolearning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2042–2055, Mar. 2020.
[31] F. Luo, L. Zhang, X. Zhou, T. Guo, and T. Yin, "Sparse-adaptive hypergraph discriminant analysis for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1082–1086, Jun. 2020.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[33] 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.
[34] 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.
[35] 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.
[36] T. Wang, X. Xu, Y. Yang, A. Hanjalic, and J. Song, "Matching images and text with multi-modal tensor fusion and re-ranking," in the *Proc. 27th ACM Int. Conf.*, 2019, pp. 12–20.
[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[38] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8697–8710.
[39] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[40] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, Sep. 2019.
[41] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.
[42] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE: Improving visual-semantic embeddings with hard negatives," 2017.
[43] Z. Wang et al., "Camp: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, vol. 1, no. 1, 2019, pp. 5763–5772.

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, Beijing, China. 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.

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.

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 from 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.

Changyuan Tian received the B.Sc. degree in electronic information science and technology from Harbin Engineering University, Harbin, China, in 2021. He is currently working toward the Ph.D. degree in signal and information processing with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include deep learning and graph neural network.

Liangyu Xu received the B.Eng. degree in electronic information science and technology from Shandong University, Qingdao, China, in 2021. 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, learning with hypergraphs, and video prediction.

Leiyi Hu received the B.Eng. degree in electronic information science and technology from Tongji University, Shanghai, China, in 2021. 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 and multimodal learning.

Chibiao Ding (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in 1997.

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