Cross-Modal Progressive Perspective Matching Network for Remote Sensing Image-Text Retrieval

Chengyu Zheng, Xiu Li, Xinyue Liang, *Member, IEEE*, Lei Huang, *Member, IEEE*, Shan Du, *Senior Member, IEEE*, Jie Nie*, *Member, IEEE*, Junyu Dong, *Member, IEEE*

***Abstract*—Cross-modality based on remote sensing (RS) text-image retrieval has gained increasing attention in recent years due to its ability to leverage the rich semantics of images and the understandability of text to provide a more comprehensive description. Existing cross-modal retrieval methods typically apply self-attention or cross-attention mechanisms to identify important information in RS data, but they ignore the multi-view perception characteristic of geographical space in RS images. As a result, these retrieval models fail to locate the correct perspective in images according to the query text, ultimately leading to incorrect matching. In this work, a Cross-modal Progressive Perspective Matching Network (CPPMN) is proposed for remote sensing image-text retrieval by establishing a progressive perspective matching mechanism and semantic alignment to further improve the performance of the retrieval model. Specifically, the CPPMN framework consists of three core modules: the Compensation Network for Full Perspective Modeling (CN_FPM), the Graph Transformation for Individual Perspective Modeling (GT_IPM), and the Cascaded Transformer for Cross-modal Semantic Alignment (CT_CSA). The CN_FPM module utilizes all positive text samples as supervision signals to guide the feature extraction training process, aiming to capture full perspective information from images. Subsequently, the GT_IPM module transforms implicit-perspective feature representations into explicit-perspective cross-modal relationship graphs. This transformation enables the identification of specific perspective locations within the image according to the query sentence by analyzing graph density and connectivity. Finally, the CT_CSA module comprises a cascaded Transformer network that aligns features at the semantic level between cross-modal data The quantitative and qualitative experiments are conducted on four large-scale remote sensing cross-modal retrieval datasets to demonstrate the significant performance of adopting the progressive perspective matching mechanism and semantic alignment strategy.**

***Index Terms*—Perspective matching, semantic alignment, remote sensing, cross-modal retrieval.**

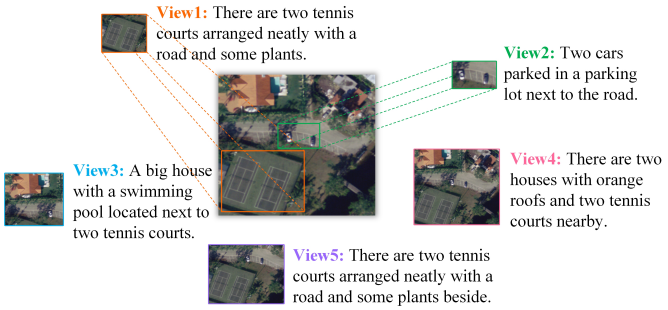

Fig. 1. Illustration of the multi-view perception of geographical space. An RS image can be described from multiple perspectives, such as view 1 in the figure focuses on local geographical objects, whereas view 4 shows the global information. The existing methods fail to consider the multi-view perception characteristic of geographical space, and thus cannot locate the correct perspective in image according to the query text, ultimately leading to incorrect matching and hindering performance improvement of retrieval models.

I. Introduction

In recent years, with the rapid development of satellite sensors, the amount of remote sensing (RS) data has shown explosive growth and is widely used in various fields [1]–[3]. Due to differences in data acquisition altitudes, capture devices, and application scenarios, remote sensing data presents considerable variation in multiple modalities, such as images, text, videos, etc. The proliferation of data and the span of modalities require researchers to employ data fusion technologies to promote the utilization efficiency of remote sensing data. As one of the important approaches to data fusion, RS cross-modal retrieval has attracted a great deal of attention because it can establish bridges between multiple modal data. By leveraging the rich semantics of images and the clarity of text for information complementation, image-text retrieval enhances data utilization and availability, thereby playing a crucial role in cross-modal retrieval tasks.

To perform cross-modal retrieval, the state-of-the-art models first extract high-level feature representations from both the query image (or text) and the data in the database. Then, a similarity calculation and ranking operation are carried out to identify the optimal matching pairs of image and text. Recently, deep learning-based image-text retrieval methods have gradually become mainstream due to their strong capability to exploit latent semantic information [4]–[6]. Compared with natural images captured by cameras, RS images contain more complex characteristics, such as dispersed targets, multiple scales, and a large amount of interference noise, which makes the cross-modal retrieval task more challenging. Some studies

This work was supported in part by the National Natural Science Foundation of China (62172376), the Regional Innovation and Development Joint Fund of the National Natural Science Foundation of China (U22A2068), and the Central Government Guides Local Science and Technology Development Fund (YDZX2022028).

Chengyu Zheng, Xiu Li, Xinyue Liang, Lei Huang, Jie Nie, and Junyu Dong are with the College of Information Science and Engineering, Ocean University of China, Qingdao, 266100, China.

Shan Du is with the Department of Computer Science, Mathematics, Physics, and Statistics, the university of British Columbia (Okanagan Campus), Kelowna, V1V 1V7, BC Canada.

Jie Nie is the corresponding author; Emails:niejie@ouc.edu.cn.

[7], [8] are dedicated to handling these characteristics to enhance the RS retrieval accuracy. Over the past two years, the development of pre-trained large models (such as CLIP [9], BLIP [10], and RemoteCLIP [11]) has created new opportunities for image-text retrieval. These large models offer significant advantages as they are trained on extensive annotated data for diverse tasks, leading to improved generalization and representation capabilities.

Despite significant performance improvements, the above methods fail to explore fine-grained information, which is crucial in cross-modal retrieval tasks as it enhances data distinguishability and ultimately leads to more accurate retrieval results. Several studies [12]–[15] have explored using Transformers [16]–[18] to develop local feature representations within patch spatial structures and adaptively extract global relationships. They first split the global data into patch-level representations and adapted the attention mechanism to establish interactive connections between local features. A drawback of these approaches is that the patch-level structure of the Transformer often fragments objects in RS images. The semantic-level extraction [19]–[21] is another effective way for fine-granularity modelling. It involves using latent semantics within the data or external knowledge as guidance to separate distinct semantics from complex backgrounds. Despite significant advancements in semantic-level extraction, there are still challenges in processing RS data with high inter-class and intra-class similarity. Fortunately, with the development of Fast R-CNN [22], RS images can be effectively represented at a refined object level, enabling models to focus on specific object attributes and reduce class ambiguity. For example, the study in [23] integrated the object detection network PP-YOLO [24] into RS cross-modal retrieval, thereby enhancing the performance of the retrieval model.

Although the above methods have made significant progress, they overlook a unique characteristic of RS images: the multi-view perception of geographical space. Because most remote sensing images are captured by sensors far from ground objects, they encompass a broad geographical range and offer multiple perspectives. As illustrated in Fig.1, an RS image can be described from various perspectives, such as view 1 focuses on local geographical objects, while view 4 emphasizes global information. In contrast, ordinary images captured by close-range devices such as cameras usually have a single viewpoint and a specific focus point, making multi-view perception less common. Existing deep learning-based cross-modal retrieval methods, such as [13], [14], typically apply self-attention or cross-attention mechanisms to identify important information in RS data, but they are not designed to extract feature representation from different perspectives. As a result, these retrieval models fail to locate the correct perspective in images according to the query text, ultimately leading to incorrect matching and hindering performance improvement of retrieval models.

In response to this challenge, we propose a Cross-modal Progressive Perspective Matching Network (CPPMN) based on progressive perspective modeling and semantic alignment, so as to achieve adaptive perspective matching of cross-modal data. Our model introduces three innovative modules: the Compensation Network Full Perspective Modeling (CN_FPM), the Graph Transformation Individual Perspective Modeling (GT_IPM), and the Cascaded Transformer Cross-modal Semantic Alignment (CT_CSA). Specifically, the CN_FPM module includes a compensation network, that uses positive text samples as supervision signals to guide the training process of image feature extraction. The utilization of diverse positive text samples is crucial, as they provide robust descriptions of images from multiple perspectives, effectively filtering out irrelevant information, such as noise and background. The compensation network integrates both saliency and semantics information from text samples, thereby ensuring the comprehensiveness and reliability of the supervision signal. The second module, GT_IPM, is designed to automatically locate individual perspectives within an image based on a given query sentence. GT_IPM initially converts image features into graph representations guided by the query sentence. Since the density and number of graph connections can explicitly reflect individual perspectives, an individual perspective extraction mechanism is employed. This mechanism first uses KL-divergence to determine the perspective to which the graph representation belongs, then selects a specialized network to train the image based on the identified perspective, ensuring both targeted and robust feature extraction. The third module, CT_CSA, develops a cascade Transformer network to achieve feature alignment at the semantic level between the cross-modal data. This network is capable of extracting effective information from individual perspective image features and highlighting salient information to filter out complex noise in the texts. Experiments conducted on four RS image-text retrieval datasets demonstrate the effectiveness of the proposed architecture.

In summary, the contributions of this paper are as follows:

- We introduce a new challenge called the multi-view perception of geographical space in the RS cross-modal remote sensing task and present a novel cross-modal retrieval network CPPMN by integrating a progressive perspective modelling pattern.
- We propose a CN_FPM module, where the full perspective is extracted by applying all positive text samples as a supervision signal to guide the training process of feature extraction. We design a GT_IPM module by transferring implicit-perspective feature representation to explicit-perspective cross-modal relationship graphs, so as to achieve the individual perspective location in the image according to the query sentence. We propose a CT_CSA module with a cascade Transformer network to achieve feature alignment at the semantic level between the cross-modal data.
- We validate the effectiveness of the proposed method on four public RS image-text retrieval datasets to demonstrate the superiority of our approach.

II. Related Work

A. Image-text Retrieval of RS

The definition of RS image and text retrieval is that given a query image (or text), the retrieval networks need to find the

optimal matching text (or image) from the database. Typically, for the query data and the data in the database, the retrieval model first extracts high-level feature representations, and then performs similarity calculations and ranking operations to identify the best matching pairs. Deep learning-based image-text retrieval methods [25]–[27] have gradually become mainstream, benefiting from their powerful feature extraction capabilities. Compared with cameras capturing natural images, RS images contain more complex characteristics, such as dispersed targets, multiple scales, and a large amount of interference noise, which makes the cross-modal retrieval task more challenging. For instance, in 2019, Abdullah et al. [28] introduced the Deep Bidirectional Triplet Network (DBTN) for the inaugural resolution of the RS image-text retrieval challenge and released the TextRS dataset concurrently. Yuan et al. [7] developed an Asymmetric Multimodal Feature Matching Network (AMFMN) to address the multi-scale characteristics of RS targets, which concurrently extracts small-scale and large-scale feature representations from RS images. To reduce occupancy and overhead, Yuan et al. [8] created a lightweight crossmodal retrieval model, which considers multi-scale information of RS images and filters redundant information from both channel and spatial levels dynamically.

In recent years, the rise of pre-trained large models like CLIP [9], BLIP [10], and MiniGPT-4 [29] has paved the new way for advancements in image text retrieval. These models have showcased strong capabilities in feature representation through training on large sets of annotated data applied to various applications. Furthermore, to address the distinctiveness of remote sensing data, specific large models such as RemoteCLIP [11] and GeoRSCLIP [30] have been developed. For instance, Liu et al. [31] employed CLIP to encode remote sensing data and transfer its rich semantic knowledge from natural to remote sensing domains.

B. Fine-grained Image-text Retrieval of RS

To further explore fine-granularity information in remote sensing data, existing works are typically categorized into three groups: patch-level feature extraction, semantic-level feature extraction, and object-level feature extraction. Aiming at patch-level feature extraction, a lot of research is dedicated to applying Transformers [16]–[18], [32] for their inherent capability of exploring potential global relationships. For example, Tang et al. [13] maximized the benefits of the Transformer in the realm of remote sensing. They achieved this by crafting an information-interacting enhancing module that concurrently models the intrinsic relationships between RS images and texts. However, the patch spatial structure of the Transformer leads to the fragmentation of objects in the RS image. Semantic-level feature extraction methods usually focus on modeling different semantics separately through similarity calculation or semantic supervision. For instance, Lee et al. developed the Stacked Cross Attention Network (SCAN), wherein they initially derived representations for individual regions in the image based on similarity and subsequently compared them with the global sentence. Considering the complex multi-scale relationships of remote sensing images, Nie et al. [33] proposed a scale-relation joint decoupling network in semantic segmentation task, which can achieve separate modeling of same-scale relationships and cross-scale relationships, improving the reliability of feature representation. Afterwards, Zheng et al. [34] came up with the idea of converting the original feature representation into a correlation-based affinity matrix to enhance the anti-interference capacity of RS features and realize semantic decoupling. Although semantic-level feature extraction methods have achieved remarkable results, they still fail in some details or small targets, resulting in inapplicability in cross-modal retrieval due to the high similarity of remote sensing data. Therefore, Yuan et al. [23] incorporated PP-YOLO [24] into RS cross-modal retrieval and advanced GaLR by expressing image and text as object-level representation, thereby boosting the performance of the retrieval model. Subsequently, Yao et al. [35] proposed a method to capture the relationships among diverse objects in remote sensing images. This involved incorporating hypergraph learning and constructing hypergraph networks at different levels to tackle challenges related to multiple types, uneven distribution, and multiscale objects.

C. Cross-modal Alignment in Image-text Retrieval of RS

Recently, to address the heterogeneous gap in cross-modal data, more and more methods are focusing on semantic alignment and relationship construction between different modalities by leveraging attention mechanisms, graph convolution, and other techniques [36]–[38]. For example, Cheng et al. [19] first introduced a cross-attention mechanism to cross-modal retrieval and raised a Deep Semantic Alignment Network (DSAN), where the gating mechanism could discover and strengthen the underlying semantic relationship between remote sensing images and text. Lv et al. [39] proposed a novel approach known as the Fusion-based Correlation Learning Model (FCLM) to capture intermodality complementation and fusion cues. Then, Chen et al. [40] designed a Multiscale Salient image-guided text alignment (MSITA) network, which can address the issue of multi-scale targets in RS images, as well as achieve cross-modal alignment through image-guided text strategies. To further enhance the interaction between modalities, Zhao et al. [41] created a visual-language method by hiding certain information of the data, continuously learning subtle image-text relationships. In addition, Li et al. [42] proposed a new strategy called False Negative Elimination (FNE) to optimize the triple loss, which can more accurately reduce the distance between positive sample pairs and increase the distance between negative samples.

Despite the significance and value of the above cross-modal alignment methods in image-text retrieval tasks, they ignore a special characteristic of RS images: the multi-view perception of geographical space. As a result, these methods fail to locate the correct perspective in images according to the query text, ultimately leading to incorrect matching and hindering performance improvement of retrieval models. Therefore, this paper aims to address this issue and propose a novel network, named CPPMN, by introducing a progressive perspective modelling pattern.

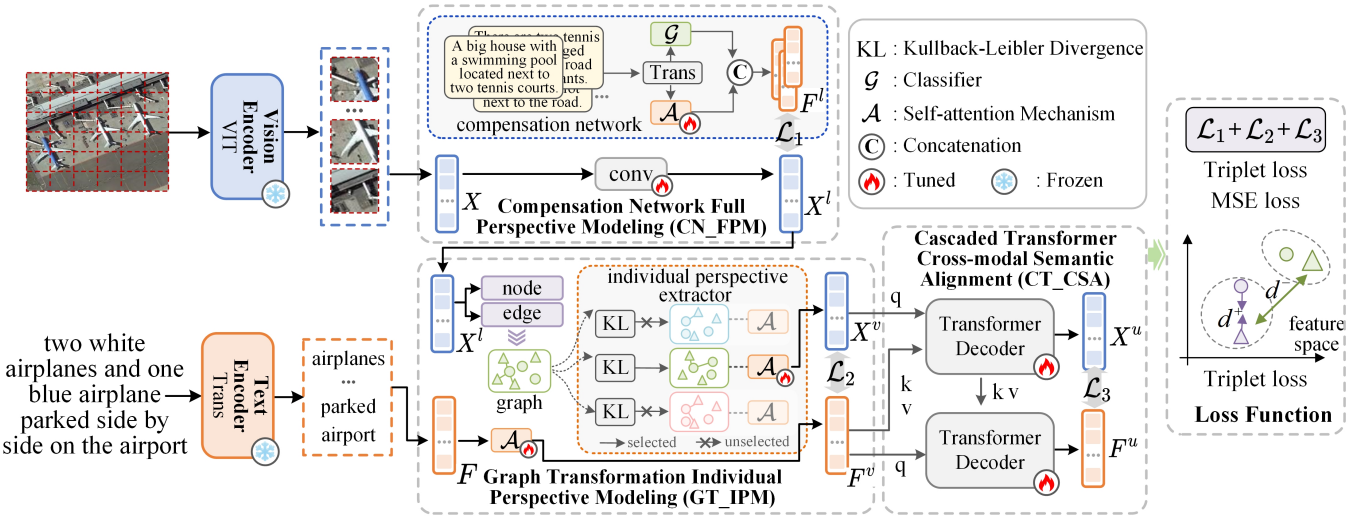

Fig. 2. The architecture of the proposed CPPMN. The CPPMN first applies pre-trained feature extractors to extract high-level features of the image and text, respectively. Subsequently, the generated image feature is fed into the CN_FPM module, which introduces a compensation network, integrating both saliency and semantics from all positive text samples as a supervision signal to guide the training process of image feature extraction, so as to separate the full perspective information from the RS images. Afterwards, the GT_IPM module is designed to initially convert image features into graph representations. Then, an individual perspective extraction process is applied, starting with KL-divergence to determine the perspective of the graph representation, followed by selecting a specialized network to train the image based on the identified perspective, ensuring targeted and robust feature extraction. Finally, the CT_CSA module is built by introducing a cascade Transformer network to achieve feature alignment at the semantic level between the cross-modal data.

III. PROPOSED METHOD

In this section, we introduce the CPPMN, which is a novel approach specifically designed to enhance the image-text retrieval capabilities of RS. Initially, we provide a comprehensive overview of the network, outlining its motivations and overall structure. We then formulate the process of CPPMN, followed by a detailed explanation of the proposed modules. Finally, we explain the loss function used for network training.

A. Overview of the Proposed CPPMN

In order to correctly locate the perspective from multi-view perception RS image according to the query text to achieve accurate retrieval matching, we propose a novel network, namely CPPMN. The architecture of CPPMN is shown in Fig. 2, and it consists of three modules, namely CN_FPM, GT_IPM, and CT_CSA. The CN_FPM module introduces a compensation network, which integrates both saliency and semantics from all positive text samples as a supervision signal to guide the training process of image feature extraction, so as to separate the full perspective information from the RS images to avoid the interference of noises and backgrounds. The GT_IPM module initially converts image features into graph representations. In these graph representations, the density and number of graph connections can clearly reflect the individual perspectives described by the query sentence in the image. Therefore, an individual perspective extraction process is applied, starting with KL-divergence to determine the perspective of the graph representation, followed by selecting a specialized network to train the image based on the identified perspective, ensuring targeted and robust feature extraction. The CT_CSA module builds a cascade Transformer network to achieve feature alignment at the semantic level between the cross-modal data.

B. Feature Extraction

We will detail the feature representation in two aspects, including image feature extraction and text feature extraction.

1) Image Feature Extraction: In recent years, CLIP [9] has exhibited strong feature representation capabilities as a result of its training on extensive datasets, utilizing over 400 million parameters. Furthermore, in response to the unique characteristics of RS data, RemoteCLIP [11] was developed. It is built upon CLIP and trained on seventeen diverse and extensive RS datasets across three tasks. Its framework integrates Vision Transformer (VIT) [18] for images and Transformer [16] for text, proving to be effective in various downstream remote sensing tasks.

Thus, we utilize pre-trained VIT on RemoteCLIP as our image feature extractor. Specifically, RS image $I \in \mathbb{R}^{H\times W\times 3}$ is firstly divided into fixed-sized patches. Then these patches are encoded by VIT to get an image feature $X \in \mathbb{R}^{N\times D}$.

$$X = \mathcal{F}(I, \theta^I) \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the vision encoder, θ^I is the frozen weights.

2) Text Feature Extraction: As previously mentioned, the Transformer trained in RemoteCLIP is very beneficial for text extraction, so we apply it as our text extractor. It is important to note that masked self-attention was utilized in the text encoder to maintain the ability to incorporate language modelling as an auxiliary objective. More specifically, a sentence T is first embedded as word vectors $e(T)$, which is delivered to the Transformer to obtain text feature $F \in \mathbb{R}^{L\times D}$.

$$F = \mathcal{T}(T, \theta^T) \tag{2}$$

where $\mathcal{T}(\cdot)$ denotes the text encoder, θ^T is the frozen weights. Note that Trans in Fig. 2 is the abbreviation of Transformer.

C. Compensation Network Full Perspective Modeling

As mentioned above, RS images are captured by sensors positioned at a considerable distance from the Earth's surface, offering multi-view information. This poses challenges for cross-modal retrieval tasks, as models often struggle to align the correct image perspective with the corresponding query text, leading to inaccurate matches. Therefore, automatically locating the individual perspective in the image based on the query sentence is the main motivation of this paper. In order to achieve this goal, we believe that the primary work is to separate the multiple perspective information from the RS images to avoid the interference of noises and backgrounds in RS images. Considering that an image in the dataset has diverse positive text samples, which can describe images from full perspectives with high robustness, we design a compensation network by proposing the CN_FPM module, where the positive text samples are applied as a supervision signal to guide the training process of image feature extraction. The CN_FPM module extracts two types of information from text samples, including saliency and semantics, to ensure the comprehensiveness and reliability of the supervision signal.

Specifically, for the image feature X, we first utilize a convolutional neural network to further extract features and output X^l.

$$X^l = cov_{1\times1,s}(X; w^s), \tag{3}$$

where $cov_{1\times1,s}(\cdot;\cdot)$ represents the space convolution operation with 1×1 the convolution kernel, w^s is the parameters to be trained. Subsequently, generated X^l is regarded as the full perspective image feature since it is supervised by supervision signal F^l with MSE loss $\mathcal{L}^{mse}$.

$$\mathcal{L}_1 = \mathcal{L}^{mse}(X^l, F^l) \tag{4}$$

The supervision signal is obtained from a compensation network, which explores two types of information from text samples, termed saliency and semantics. The former is designed to extract effective descriptions from the positive text samples and filter out invalid backgrounds, while the latter is devised to view these texts globally. First, we define positive text samples with different perspectives as $\{T'_k\}_{k=1}^K$, where K represents the number of samples. Similar to the text feature generation process, word embeddings and the Transformer are utilized again to transfer word representations into features $\{F_k\}_{k=1}^K$. Subsequently, for saliency, we apply the self-attention mechanism on F_k to capture the saliency of samples and integrate the newly generated features to form a robust text supervision signal, which is formulated as follows:

$$\begin{aligned} F_k^g &= \mathcal{A}(F_k; w^g), \\ \mathcal{A}(F_k; w^g) &= (cov_{1\times1,c}(F_k; w^g) \odot F_k), \end{aligned} \tag{5}$$

where $\mathcal{A}(\cdot;\cdot)$ refers to performing self-attention mechanism on the former item. w^g is the convolution parameters in $\mathcal{A}(\cdot;\cdot)$ and $cov_{1\times1,c}(\cdot)$ represents the channel convolution operation. $\odot$ denotes dot product operation. Regarding semantics, we transfer the classifier pre-trained on the WikiText-2 dataset to generate the semantic representation of the texts.

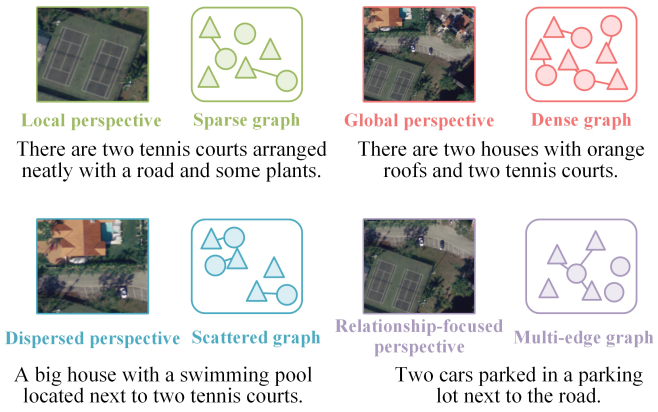

Fig. 3. The relationships between image perspectives and cross-modal relationship graphs by the analysis of the density and number of connections in the graph. For example, a sparse graph indicates a local perspective, while a dense graph represents a global perspective, and a scattered graph implies that the sentence describes the image from a dispersed perspective, etc.

$$F_k^c = \mathcal{G}(F_k; w^c), \tag{6}$$

where $\mathcal{G}(\cdot;\cdot)$ refers to the classifier implemented on the former item and w^c is the corresponding convolution parameter. Finally, the generated features F_g^k and F_c^k are combined to produce the robust supervision signal F^l.

$$F^l = \sum_{k=1}^{K} [F_k^g; F_k^c], \tag{7}$$

where $[\cdot;\cdot]$ denotes concatenation operation.

D. Graph Transformation Individual Perspective Modeling

In this section, we conceive a pattern to automatically locate the individual perspective in the image according to the query sentence. The individual perspective in RS involves multiple aspects, such as global perspective, local perspective, dispersed perspective, and relationship-focused perspective. In order to distinguish these various perspectives, we transfer feature representation to cross-modal relationship graphs, which hold the capability to explicitly display image perspectives according to the query sentence by analyzing the density and number of connections in the graph. For example, as shown in Fig. 3, a sparse graph indicates a local perspective, while a dense graph represents a global perspective, and a scattered graph implies that the sentence describes the image from a dispersed perspective, etc. Based on the identified perspective, the image is transmitted to a specific network, where the correct perspective information could be mined, so as to achieve precise alignment between text and image, further improving the accuracy of retrieval results.

1) Graph Representation: To construct cross-modal relationship graphs, we first define the nodes and edges. Since features are inherently in two modalities, image and text, we calculate the graphs in the following way.

Nodes: For the generate image feature X^l and text feature F, the concatenation operation is performed to generate cross-modal features $G \in \mathbb{R}^{(N+L)\times D}$.

$$G = [X^l; F], \tag{8}$$

where $[\cdot;\cdot]$ denotes concatenation operation between the two items. Then, we regard the spatial pixels of the features as nodes. Thus, a set of nodes $G = \left\{g^1, ..., g^i, ..., g^{(N+L)}\right\}$ is obtained, and each node encodes different regions or words in images and texts.

Edges: In order to acquire the relations between nodes, we refer to the work of [43]. The pairwise affinities of features are calculated as edges of the graph, formulating as follows:

$$R(g^i, g^j) = \phi(g^i)\varphi(g^j)^T, \tag{9}$$

where $\phi(g^i) = w^\phi g^i$ and $\varphi(g^j) = w^\varphi g^j$. The parameters w^ϕ and w^φ are trained by back propagation. Due to the nodes of the graphs being generated by the concatenation from images and texts, the above formula can also be written as:

$$R(g^i, g^j) = R(X^l, F) = \begin{bmatrix} \phi(X^l)\varphi(X^l)^T & \phi(X^l)\varphi(F)^T \\ \phi(F)\varphi(X^l)^T & \phi(F)\varphi(F)^T \end{bmatrix} \tag{10}$$

Our goal is to extract perspectives from images based on text, thus we only need to create a text-to-image graph structure.

$$R(X^l, F) = \begin{bmatrix} 0 & 0 \\ \phi(F)\varphi(X^l)^T & 0 \end{bmatrix} \tag{11}$$

The formula is then expressed as:

$$R(X^l, F) = \phi(F)\varphi(X^l)^T \tag{12}$$

2) Individual Perspective Extractor: In this section, we build a graph dictionary containing various graph structures. The individual perspective extractor compares the generated graph structure $R(X^l, F)$ with the graph structures $\{G_p\}_{p=1}^P$ in this dictionary. This comparison helps determine the graph structure to which it belongs, thus inferring the image perspective the text depicts. Specifically, the graph structure is first passed into the convolutional neural network to mine potential information. Here, we rewrite $R(X^l, F)$ as R for simplicity.

$$R^s = cov_{1\times 1,s}(R; w^r), \tag{13}$$

where w^r refers to the trainable parameters in CNN. Then, KL(Kullback-Leibler)-divergence D^{KL} is utilized to compare the R^s with the graphs in the dictionary to deduce the perspective.

$$\mathbb{D}_p^{KL}(R^s \parallel G_p) = \sum_{q=1}^{Q} R^s(q) ln(\frac{R^s(q)}{G_p(q)}), \tag{14}$$

$$o = \underset{p}{argmin}(\mathbb{D}_p^{KL}(R^s \parallel G_p)), \tag{15}$$

where p represents the index of graph types in the dictionary and q is the matrix row vector in R^s and G_p. Eq. 14 and 15 can be interpreted as follows: for the generated R^s, we calculate its KL-divergence with each G_p in the dictionary. Then the most matching graph is selected with the largest KL-divergence, and recorded in o. o plays the role of an index. Through o, the model identifies the dictionary/perspective described by the text in the image, and transfers X^l to the corresponding feature extractor.

$$X^u = \mathcal{A}(X^l; w_o^x), \tag{16}$$

where w_o^x is the convolution parameters in $\mathcal{A}(\cdot;\cdot)$. It should be noted that currently, only the feature extractor corresponding to "o" is involved in the training process. By applying the above strategy, we can determine the graph dictionary that R^s belongs to. In turn, the graphs in the dictionary also rely on R^s for updating.

$$G^p = G^o + R^s \tag{17}$$

An additional clarification is that R^s for each data can only update its corresponding graph dictionary. The text feature F is also input to the feature extractor corresponding to filter out noise information.

$$F^u = \mathcal{A}(F; w^f) \tag{18}$$

Finally, the generated image and text features X^u and F^u are trained with triplet loss $\mathcal{L}^{tri}$.

$$\mathcal{L}_2 = \mathcal{L}^{tri}(X^u, F^u) \tag{19}$$

E. Cascaded Transformer Cross-modal Semantic Alignment

As introduced above, the GT_IPM module is designed to automatically locate individual perspectives in an image based on the query sentence. Building on the GT_IPM module, the CT_CSA module is developed to align features at the semantic level between image and text features. Given that the Transformer decoder can achieve semantic alignment by calculating cross-modal similarity, we designed a cascaded Transformer network. In this network, the first stage extracts information relevant to the query text from individual perspective image features, while the second stage filters out complex noise within the text features.

1) Transformer Decoder: The Transformer decoder consists of N = 2 identical stacks, and its structure follows the work [16]. Each stack of the decoder contains three sub-layers, namely multi-head attention, multi-head cross-attention and feed-forward network. The residual connections are employed around each of the sub-layers, accompanied by layer normalization (Add & Norm), which we only represent in the figure for simplicity. To improve the comprehensibility of the proposed method, here, we introduce the attention and multi-head attention in detail. The attention function contains three main elements, namely $query$ and a set of key - $value$ pairs. The output is calculated as a weighted sum of the $values$, where the weight assigned to each $value$ is computed by a compatibility function of the $query$ with the corresponding key. The attention output is computed as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{20}$$

where Q, K and V represent the matric of $query$, key and $value$, respectively. $\sqrt{d_k}$ is the scaling factor. Instead of performing single attention, multi-head attention holds a better

performance due to its capability of focusing on different representations jointly on different parts:

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(Head_1,...,Head_H)W_H, \\ Head_i &= Attention(Qw_i^q, Kw_i^k, Vw_i^v), \end{aligned} \tag{21}$$

where w^q, w^k, and w^v are the parameter matrices. The feed-forward network is the last sub-layer in each stack, and is composed of two liner networks with ReLU activation:

$$FFN(x) = max(0, xw_1 + b_1)w_2 + b_2 \tag{22}$$

2) Text-guided Image Feature Extraction : This unit utilizes the standard Transformer decoder structure. In order to extract effective information in the image according to the query sentence, multi-head cross-attention is utilized, where $query$ is derived from X^u, key and $value$ are obtained from the text feature F^u. Each Transformer decoder layer t updates the $queries$ X_{t-1}^u from the output previous stack:

$$\begin{aligned} X_t^{mid} &= MultiHead(X_{t-1}^u, F^u, F^u), \\ X_t^u &= FFN(X_t^{mid}), \end{aligned} \tag{23}$$

where X_t^{mid} represents the outputs of the middle layer decoder in the Transformer. Through the above process, the effective semantic information in images is emphasized based on the query text, and we obtain the robust image feature X^v, which refers to the output of the last decoder layer.

3) Image-guided Text Feature Extraction: Contrary to the text-guided image feature extraction unit, this unit considers F^u as $query$ and robust image feature X^v as key and $value$ to filter out complex noise information by utilizing multi-head cross-attention. Thus, the $queries$ F_{i-1} are updated in layers i of the Transformer decoder as follows:

$$\begin{aligned} F_t^{mid} &= MultiHead(F_{t-1}^u, X^v, X^v), \\ F_t &= FFN(F_t^{mid}), \end{aligned} \tag{24}$$

Similar to the text-guided image feature extraction unit, the F_t^{mid} is an interactive variable computed by multi-head attention. After that, F^u is transferred to robust text feature F^v, where F^v is the output of the last decoder layer. By utilizing cross-attention to calculate cross-modal correlations, the model finally achieves reliable feature representation and cross-modal alignment. Finally, the generated X^v and F^v are sent into triplet loss $\mathcal{L}^{tri}$.

$$\mathcal{L}_3 = \mathcal{L}^{tri}(X^v, F^v) \tag{25}$$

F. Loss Function

As usual, the triple loss is applied as a loss function for our cross-modal retrieval modal since it can achieve a more refined measure of difference. The purpose of triple loss is to increase the distance between a sample and the corresponding negative sample while reducing the distance between the sample and its positive sample as much as possible. [44] proposes a bidirectional triple loss for text-image matching:

$$\begin{aligned} \mathcal{L}^{tri}(X,F) = &\textstyle\sum_{\hat{T}}[\alpha - S(X,F) + S(X,\hat{F})]_+ \\ &+ \textstyle\sum_{\hat{I}}[\alpha - S(X,F) + S(\hat{X},F)]_+, \end{aligned} \tag{26}$$

where α refers to the margin and $[\cdot]_+ = max(\cdot, 0)$. $S(X,F)$ represent the similarity between positive sample pairs of image and text. $S(X,\hat{F})$ is processed for all negative sentences $\hat{F}$ of a given image X, and $S(\hat{X},F)$ considers all negative images $\hat{X}$ of a given sentence F. To save the calculation cost, the loss is usually computed in each batch rather than in all training sets. In Eq. 4, the MSE loss is utilized, which focuses on minimizing the distance between the image and text features. It can be expressed as:

$$\mathcal{L}^{mse}(X^l, F^l) = \frac{1}{D}\sum_{d=1}^{D}(x_d^l - f_d^l) \tag{27}$$

where $X^l = \{x_1^l, ..., x_D^l\}$ and $F^l = \{f_1^l, ..., f_D^l\}$. The total loss can be written as

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 \tag{28}$$

IV. Experimental Results and Analysis

In this section, to evaluate the efficacy of the proposed method, we conduct experiments. We begin with a brief overview of the datasets, metrics, implementation details, and baselines. Subsequently, we perform experiments on the proposed approach, including ablation studies and comparisons with state-of-the-art methods.

A. Datasets

We perform experiments on four benchmark RS datasets for cross-modal image-text retrieval: RSICD, RSITMD, UCM-Captions, and Sydeny-Captions.

The RSICD dataset is used for the remote sensing image captioning task [45]. The spatial sizes of the data files are fixed to 224 $\times$ 224 pixels with various resolutions. The total number of remote sensing images is 10921, with five sentences of descriptions per image.

The RSITMD dataset is supplied by [46] for RS cross-modal image-text retrieval. The images in the RSITMD dataset are selected from the RSICD dataset and provide a total of 23715 captions for 4743 images. The RSITMD dataset has more granular and diverse captions than the RSICD dataset.

The UCM-Captions is built by [47], based on the UCMerced-LandUse dataset [48]. It contains land use images in 21 categories, with 100 images per category. The spatial sizes of the data files are 256 $\times$ 256 pixels, and the pixel resolution is 0.3m. The UCM-Captions dataset exploited five different sentences to describe every image.

The Sydney-Captions dataset is also provided by [47], which is based on the Sydney dataset [49]. The spatial sizes of the data files are 18 000 $\times$ 14 000 pixels, and the pixel resolution is 0.5 m. Similar to the UCM-Captions dataset, five different sentences were given to describe each image.

B. Evaluation Metrics

We perform two types of image-text matching tasks: 1) sentence retrieval, involving the retrieval of ground-truth sentences related to the query image (I2T); and 2) image retrieval,

TABLE I
COMPARISONS OF IMAGE-TEXT RETRIEVAL EXPERIMENTS ON RSICD AND RSITMD DATASETS.

Type	Method	Backbone Vision/Text	RSICD			RSITMD		
			Image to Text R@1 / R@5 / R@10	Text to Image R@1 / R@5 / R@10	mR	Image to Text R@1 / R@5 / R@10	Text to Image R@1 / R@5 / R@10	mR
N	SCAN t2i	ResNet/GRU	4.42 / 11.20 / 17.68	4.02 / 11.54 / 18.60	11.24	10.59 / 28.72 / 38.41	10.04 / 29.54 / 42.91	26.70
	SCAN i2t	ResNet/GRU	5.90 / 13.21 / 19.96	3.86 / 16.83 / 26.49	14.38	10.84 / 25.86 / 37.26	9.62 / 29.80 / 41.06	25.74
	CAMP-triplet	ResNet/GRU	5.22 / 13.05 / 21.02	4.30 / 17.10 / 27.57	14.71	11.82 / 27.40 / 38.12	8.69 / 27.17 / 43.60	26.13
	CAMP-bce	ResNet/GRU	4.50 / 10.08 / 16.48	3.03 / 15.12 / 23.05	12.04	9.21 / 22.56 / 35.73	6.81 / 25.65 / 40.05	23.34
	MTFN	ResNet/GRU	5.01 / 12.86 / 21.30	4.96 / 12.14 / 29.12	14.23	10.80 / 27.68 / 36.40	9.82 / 30.28 / 48.27	27.21
	SGRAF	ResNet/GRU	3.93 / 13.36 / 22.78	3.42 / 14.53 / 26.72	14.12	5.97 / 18.81 / 30.09	6.02 / 24.25 / 41.81	21.16
	NAAF	ResNet/GRU	5.92 / 16.31 / 27.23	4.21 / 17.20 / 26.33	16.20	10.50 / 30.75 / 41.86	9.13 / 30.38 / 44.25	27.81
R	SAM i2t	CNN/RNN	6.41 / 20.10 / 32.89	5.91 / 18.08 / 29.73	18.85	13.27 / 30.09 / 44.91	9.87 / 33.89 / 50.66	30.45
	SAM t2i	CNN/RNN	5.91 / 18.59 / 30.63	4.82 / 18.10 / 33.44	18.58	11.73 / 27.25 / 40.06	11.07 / 35.03 / 50.94	29.35
	LW-MCR	SqueezeNet/RNN	3.66 / 13.63 / 21.68	3.77 / 15.32 / 28.20	14.38	6.64 / 17.92 / 30.31	5.53 / 25.51 / 44.34	21.71
	AMFMN	ResNet/GRU	5.39 / 15.32 / 28.82	5.05 / 18.19 / 29.07	16.97	11.37 / 27.70 / 39.29	9.04 / 32.46 / 49.68	28.26
	GaLR	ResNet+ppyolo/GRU	4.67 / 17.47 / 27.63	4.65 / 20.20 / 34.80	18.24	13.27 / 31.85 / 41.15	10.92 / 35.3 / 52.96	30.91
	MCRN	CNN/GRU	5.58 / 18.38 / 30.56	4.87 / 18.92 / 33.25	18.59	8.85 / 23.67 / 35.84	8.05 / 30.97 / 49.60	26.16
	CMFM	ResNet/GRU	5.40 / 16.47 / 27.54	4.15 / 17.99 / 31.25	17.13	8.85 / 23.89 / 37.17	7.12 / 31.95 / 50.88	26.64
	PIR	Swin Trans/Bert	10.52 / 26.44 / 39.52	6.83 / 23.97 / 39.09	24.40	19.03 / 41.15 / 53.32	13.72 / 42.48 / 62.88	38.76
	HVSA	CNN/RNN	8.60 / 22.78 / 33.21	5.95 / 22.53 / 34.57	21.27	14.60 / 32.96 / 46.46	10.84 / 38.01 / 57.35	33.37
	KAMCL	ResNet/GRU	8.14 / 22.14 / 33.39	6.35 / 23.70 / 38.87	22.10	14.82 / 29.65 / 43.58	10.53 / 35.8 / 55.84	31.70
	DOVE*	ResNet/GRU	8.66 / 22.35 / 34.95	6.04 / 23.95 / 40.35	22.72	16.81 / 36.80 / 49.93	12.20 / 44.13 / 66.50	37.73
P	IEFT	Trans/Trans	4.86 / 19.76 / 34.48	6.77 / 23.88 / 38.15	21.32	14.82 / 33.63 / 49.56	10.98 / 37.61 / 58.25	34.14
	CLIP	Trans/Trans	10.52 / 25.25 / 39.16	9.33 / 30.01 / 46.04	26.72	14.60 / 36.28 / 49.56	14.42 / 43.65 / 64.65	37.19
	RemoteCLIP	Trans/Trans	16.65 / 35.13 / 47.39	14.73 / 40.14 / **56.96**	35.17	27.21 / 50.22 / 63.93	24.20 / 57.21 / **74.96**	49.62
	GLISA*	Trans/Trans	**19.52** / **40.44** / 52.28	14.75 / 39.50 / 55.46	36.99	28.21 / 50.94 / 62.50	20.50 / 54.80 / 72.19	48.41
	CPPMN	Trans/Trans	19.12 / 39.57 / **53.89**	**15.61** / **41.22** / **56.96**	**37.72**	**30.97** / **57.30** / **68.80**	**27.08** / **60.40** / **74.96**	**53.25**

entailing the retrieval of ground-truth images related to the query text (T2I). The evaluation metric used is the Recall at K ($R@K$, with K=1, 5, and 10), which represents the proportion of ground truth appearing in the top K results. Additionally, we calculate the average mR, based on the recall rates of $R@K$ proposed by Huang et al. [50], to provide a more comprehensive assessment of the model's performance.

C. Implementation Details

In this subsection, we mainly introduce the implementation details of the proposed method. The channel D of the image feature X and text feature F is 768. For triple loss (Eq. 26), we conduct a large number of ablation experiments and take the value of α to be 0.2. For the text classifier, we use BERT, pre-trained on the WikiText-2 dataset. All the models are implemented with PyTorch and the Adam optimizer with a 0.0002 learning rate. We set the batch size of training to 30 and trained the model for 50 epochs. All the experiments are conducted on a server with one NVIDIA 3090 24GB.

D. Baselines

As for baselines, nineteen image-text retrieval methods are compared, including five classic image-text retrieval networks applied on natural data, abbreviated as "N", which are SCAN [51], CAMP [52], MTFN [53], SGRAF [54], NAAF [55], and nine networks dedicated to RS image-text retrieval, termed "R", including SAM [19], LW-MCR [8], AMFMN [7], GaLR [23], MCRN [56], CMFM [57], PIR [15], HVSA [58] KAMCL [59] and DOVE [60]. In addition, we compare the results with four pre-trained large models, abbreviated as "P", namely IEFT [13], CLIP [9], RemoteCLIP [11] and GLISA [37]. It should be noted that since some methods do not have publicly available codes, we directly quote the results provided by the authors in the published manuscript and mark them with the symbol "*".

E. Comparison with State-of-the-Art Methods

To prove the effectiveness of the proposed CPPMN, we compare our method with the mainstream nineteen methods on four datasets. The relevant results are shown in TABLE I and II, we give the following analysis:

Results on RSICD: The first table of TABLE I presents comparative experimental results on the RSICD dataset, where we can see that NAAF outperforms all classic cross-modal retrieval models, achieving 16.20 accuracy on the mR evaluation metric. Besides, regarding retrieval models designed for processing remote sensing data, PIR surpasses other methods, with a precision of 24.40. For pre-trained large models, our CPPMN achieves outstanding achievements, improving by 1.97% compared to the current best retrieval model GLISA. This superior expression is because the proposed CN_FPM and GT_IPM modules can realize progressive perspective modelling from full to distinguishable perspective, and the GT_IPM module holds the capability of semantic alignment on cross-modal data. It is important to note that all models perform poorly on this dataset, possibly due to its low pixel resolution, resulting in poor data quality.

Results on RSITMD: The comparative experiments on the RSITMD dataset are shown in the right table of TABLE I. We can observe that NAAF still obtains the best retrieval results among all classic models since it considers both mismatched and matched textual fragments with negative mining strategies.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3535365

TABLE II
COMPARISONS OF IMAGE-TEXT RETRIEVAL EXPERIMENTS ON UCM AND SYDNEY DATASETS.

Type	Method	Backbone Vision/Text	UCM			SYDNEY		
			Image to Text R@1 / R@5 / R@10	Text to Image R@1 / R@5 / R@10	mR	Image to Text R@1 / R@5 / R@10	Text to Image R@1 / R@5 / R@10	mR
N	SCAN t2i	ResNet/GRU	13.89 / 45.81 / 68.59	13.11 / 50.69 / 78.21	45.05	19.14 / 50.71 / 74.12	17.81 / 55.94 / 76.57	49.05
	SCAN i2t	ResNet/GRU	12.81 / 46.99 / 69.11	12.59 / 46.89 / 72.71	43.52	20.42 / 54.16 / 67.62	16.13 / 57.61 / 75.98	48.65
	CAMP-triplet	ResNet/GRU	11.20 / 44.29 / 65.72	9.89 / 46.11 / 77.29	42.42	22.83 / 50.46 / 75.88	15.31 / 43.14 / 70.39	46.34
	CAMP-bce	ResNet/GRU	15.09 / 47.18 / 68.59	11.93 / 47.22 / 76.98	44.50	15.64 / 49.65 / 71.33	11.59 / 51.34 / 76.16	45.95
	MTFN	ResNet/GRU	10.68 / 47.62 / 64.31	14.74 / 52.65 / 81.13	45.19	21.71 / 51.58 / 69.03	14.13 / 55.96 / 78.62	48.51
	SGRAF	ResNet/Bi-GRU	4.76 / 36.19 / 63.81	10.76 / 48.67 / 78.38	40.43	15.52 / 43.10 / 62.07	7.93 / 46.55 / 73.10	41.38
	NAAF	ResNet/GRU	12.38 / 49.52 / 64.76	11.52 / 44.38 / 66.29	41.48	24.14 / 56.90 / 68.97	17.24 / 46.90 / 75.86	48.34
R	SAM i2t	CNN/RNN	13.78 / 46.77 / 67.15	11.32 / 49.27 / 81.81	45.02	24.12 / 55.17 / 75.93	13.41 / 54.46 / 76.62	49.95
	SAM t2i	CNN/RNN	11.92 / 45.23 / 70.95	11.42 / 53.41 / 83.12	46.01	27.64 / 56.87 / 72.35	19.01 / 57.93 / 79.96	52.29
	LW-MCR	SqueezeNet/RNN	13.81 / 40.10 / 59.05	10.38 / 44.86 / 74.95	40.53	17.24 / 43.10 / 63.79	16.21 / 56.21 / 75.17	45.29
	AMFMN	ResNet/GRU	14.48 / 51.17 / 67.35	14.49 / 51.82 / 80.71	46.67	20.69 / 50.73 / 75.23	13.45 / 60.00 / 82.70	50.47
	MCRN	CNN/GRU	12.38 / 44.29 / 68.09	11.71 / 50.48 / 92.76	46.62	25.86 / 55.17 / 65.52	15.52 / 54.83 / 80.34	49.54
	CMFM	ResNet/GRU	13.33 / 50.00 / 66.67	11.43 / 49.24 / 74.57	44.21	17.24 / 51.72 / 62.07	15.17 / 55.52 / 81.38	47.18
	HVSA	CNN/RNN	14.29 / 39.52 / 59.52	13.71 / 48.48 / 73.33	41.48	13.79 / 48.28 / 62.07	14.48 / 55.52 / 77.24	45.23
P	IEFT	Trans/Trans	11.90 / 47.62 / 75.71	13.63 / 53.03 / 90.45	48.72	22.41 / **63.79** / 72.41	18.59 / 54.65 / 79.55	51.90
	CLIP	Trans/Trans	17.63 / 51.43 / 74.29	14.57 / 57.05 / 94.95	51.65	22.41 / 55.17 / 67.24	16.90 / 56.55 / 80.34	49.77
	RemoteCLIP	Trans/Trans	15.71 / 50.95 / 80.47	17.71 / 63.33 / **98.00**	54.36	20.69 / 56.89 / 74.14	21.03 / 64.83 / 82.41	53.33
	CPPMN	Trans/Trans	**25.71 / 63.81 / 86.67**	**21.33 / 66.00 / 98.00**	**59.70**	**27.64 / 63.79 / 75.93**	**23.79 / 65.86 / 86.90**	**55.57**

TABLE III
ABLATION STUDIES

Method	RSITMD		
	Image to Text R@1 / R@5 / R@10	Text to Image R@1 / R@5 / R@10	mR
FE	27.21 / 50.22 / 63.93	24.20 / 57.21 / **74.96**	49.62
FE + CN_FPM	30.31 / 52.88 / 66.37	23.19 / 57.65 / 73.50	50.65
FE + CN_FPM + GT_IPM	30.53 / 53.09 / 65.48	25.17 / 59.12 / 73.83	51.20
FE + CN_FPM + CT_CSA	**31.86** / 53.09 / 67.59	25.82 / 58.94 / 74.25	51.93
FE + CN_FPM + GT_IPM+ CT_CSA	30.97 / **57.30** / **68.80**	**27.08 / 60.40 / 74.96**	**53.25**

For models in the field of RS, PIR also exhibits the best performance, especially on the $R@1$ evaluation metric. This is because PIR leverages prior knowledge to guide the adaptive learning of vision and text representations, thereby enhancing vision-language understanding. Among all methods, our method makes optimum outputs on all evaluation metrics, and achieves a retrieval accuracy of 53.25 mR, demonstrating the effectiveness of the proposed model.

Results on UCM: The left table of TABLE II lists the relevant results on the UCM dataset. For the classic cross-modal retrieval model, MTFN exhibits the highest cross-modal retrieval result of 45.19 on the mR evaluation metric. Regarding retrieval methods in the RS field, AMFMN achieved a result of 46.67 on the mR evaluation metric, slightly outperforming other methods. While RemoteCLIP achieves satisfactory results, it is important to highlight that our method surpasses it by 9.82%, representing a significant improvement. This approach considers the multi-view perception of geographical space, using all positive text samples as a supervisory signal to guide feature extraction and capture the full perspective in images, while applying graph transformation to explicitly explore individual perspective information.

Results on SYDNEY: It can be observed from the last table of TABLE II that all retrieval models trained on the UCM dataset yield better accuracy than on other datasets. The SCAN t2i and SAM t2i achieved the best performance in the natural and remote sensing cross-modal retrieval models, respectively, with 49.05 and 52.29 on the mR. Besides, the proposed model demonstrates competitive advantages across various evaluation metrics, such as R@1, R@5, and R@10, and in both text-to-image and image-to-text retrieval tasks. Regarding mR, the CPPMN achieves 4.20% gains compared with the most prominent retrieval results output by RemoteCLIP. The results on SYDNEY indicate the availability and feasibility of the proposed modules.

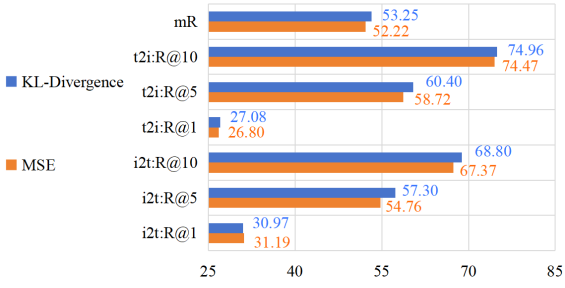

Fig. 4. Performance of KL-divergence

TABLE IV
PERFORMANCE OF $\mathcal{G}$ AND $\mathcal{A}$ IN CN_FPM

Method	RSITMD		
	Image to Text R@1 / R@5 / R@10	Text to Image R@1 / R@5 / R@10	mR
w/o $\mathcal{A}$, $\mathcal{G}$	29.68 / 56.62 / 66.99	26.43 / 60.00 / 73.89	52.27
w/o $\mathcal{G}$	30.57 / 56.96 / 67.48	26.22 / 60.21 / 74.23	52.61
w/o $\mathcal{A}$	30.46 / 57.14 / 67.48	26.82 / 59.89 / 74.77	52.76
CN	**30.97 / 57.30 / 68.80**	**27.08 / 60.40 / 74.96**	**53.25**

F. Ablation Studies

In this subsection, ablation experiments are conducted to verify the effectiveness of the proposed modules. The experimental results are listed in TABLE III, in which FE refers to the network only utilizing the feature extraction module, and CN_FPM, GT_IPM, and CT_CSA indicate the proposed three modules. "+" means that the network integrates and adopts the corresponding modules.

As shown in the first row of the table, FE has the lowest accuracy for the mR evaluation metric at 49.62. The integration of module CN_FPM promotes the performance of the network, with a 2.08% improvement in mR. Besides, from the third and fourth rows, we can see that both GT_IPM and CT_CSA modules contribute to the upgrade of retrieval results, increasing the mR evaluation metric by 3.18% and 4.64%, respectively. After integrating all modules into the network, the model achieves peak performance, reaching an accuracy of 53.25. This is because the model not only performs progressive perspective modeling from full to individual perspectives but also achieves semantic alignment across cross-modal data. Therefore, these results clearly illustrate the effectiveness and advancement of each designed module.

G. Performance of KL-Divergence

We perform the following experiments to demonstrate the effectiveness of KL-divergence in the GT_IPM module. As explained earlier, the KL-divergence is calculated to compare the generated graph with the graphs in the dictionary to determine the perspective indicated by the text in the image. Apart from using KL-divergence, we also utilize MSE to achieve this objective, and the relevant results are illustrated in the bar chart. In Fig. 4, the orange line and blue line represent the retrieval results of MSE and KL-divergence, respectively. i2t denotes the image-to-text retrieval pattern, while i2t refers to the text-to-image.

Based on the chart, we observe that KL-divergence outperforms MSE, particularly in terms of the R@5 and R@10 evaluation metrics. Specifically, it achieves an accuracy of 60.40 and 74.96 in the t2i task, and 57.30 and 68.80 in the i2t task, respectively. Regarding the average evaluation metric mR, KL-divergence improves the MSE by 1.97%, benefiting from its capability to measure the difference between two distributions. Therefore, we choose KL-divergence to build our retrieval network in the GT_IPM module.

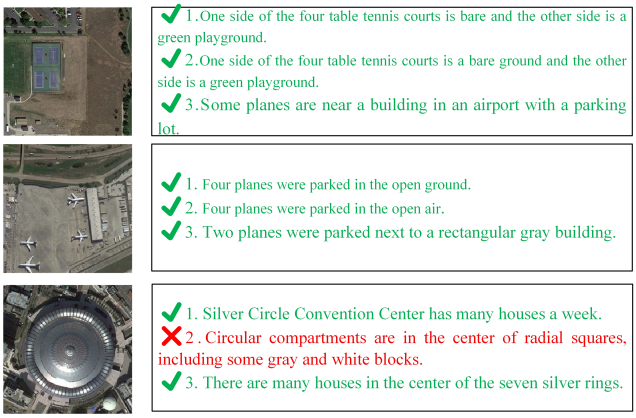

Fig. 5. Top-3 image-to-text retrieval results on RSITMD dataset. The ground truth texts are marked with green checks, and the wrong results are indicated by cross marks.

TABLE V
COMPLEXITY ANALYSIS

Type	Method	Parameters	Time(s)
N	SCAN	37M	0.0081
	CAMP	38M	0.0061
	MTFN	69M	0.0057
	SGRAF	28M	0.0085
	NAAF	23M	0.0058
R	SAM	17M	0.0474
	LW-MCR	12M	0.0024
	AMFMN	37M	0.0025
	GaLR	47M	0.0043
	MCRN	52M	0.0037
	CMFM	41M	0.0026
	HVSA	34M	0.0025
P	RemoteCLIP	427M	0.0114
	CPPMN	37M	0.0069

H. Performance of $\mathcal{G}$ and $\mathcal{A}$ in CN_FPM

In this section, to select an optimal mechanism for extracting useful information from positive text samples to generate supervision signals, the following ablation experiments are performed, and the experimental results are shown in TABLE IV. w/o means that the network did not perform the operation indicated by the following item. For example, w/o $\mathcal{A}$ in the table represents that the network does not perform the self-attention mechanism to extract saliency information. CN refers to the proposed compensation network, which applies both $\mathcal{A}$ and $\mathcal{G}$ to explore saliency and semantics.

As can be seen from the last row in the table, the compensation network delivers the most favourable outcomes. The exclusion of $\mathcal{A}$ and $\mathcal{G}$ from the model leads to a decline in performance, which is reduced by 0.92% and 1.20%, respectively, compared with the compensation network. This clearly demonstrates the importance of both $\mathcal{A}$ and $\mathcal{G}$ and their indispensability. Finally, the model produces the worst results when $\mathcal{A}$ and $\mathcal{G}$ are not applied.

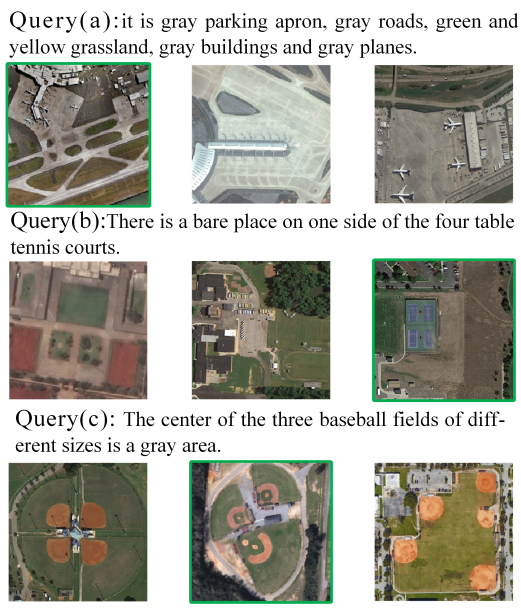

Fig. 6. Top-3 text-to-image retrieval results on RSITMD dataset. The matched images are annotated in green boxes.

I. Complexity Analysis

We also performed complexity analysis experiments, as demonstrated in TABLE V. "Parameters" represents the trainable parameters in the network, and "Time" refers to the average time needed to train the network on one sample pair.

The table shows that the parameters of the large pre-trained model, categorized "P", are significantly higher than those of the models not pre-trained on the large model, labelled "N" and "R". Since our model leverages the pre-trained encoder from RemoteCLIP, the parameter count remains efficient. In terms of time, the proposed CPPMN has a moderate matching time of 0.0069 compared to all models. Consequently, our model strikes a balance between efficiency and effectiveness.

J. Qualitative Analysis

In this subsection, we visualize the retrieval results for some of the RSITMD datasets, as shown in Fig. 5 and 6. It can be seen from Fig. 5 that for the task of image-to-text, our model can achieve correct retrieval. Although there are some wrong results, the retrieved text can also describe the image. As can be seen from (a) and (b) in Fig. 6, our model can also be well implemented for the task of text-to-image. Although (c) in Fig. 6 retrieves the wrong image, the retrieved image still fits the text description. In conclusion, from the visualization results, our model can perform the retrieval task very well.

V. Conclusion

In this paper, considering the multi-view perception characteristics of geographical space, we have introduced a novel CPPMN for RS cross-modal retrieval tasks to enhance retrieval model performance. To this end, we have devised three modules: CN_FPM, GT_IPM, and CT_CSA. The first two aim to achieve progressive perspective matching, while the third addresses semantic alignment. Specifically, the first module, CN_FPM, includes a compensation network where positive text samples serve as supervisory signals, guiding the image feature extraction process to separate full perspective information from RS images and avoid interference from noise and backgrounds. The second module, GT_IPM, automatically locates individual perspectives within an image based on a given query sentence by converting implicit-perspective feature representations into explicit cross-modal relationship graphs. The third module, CT_CSA, introduces a cascaded Transformer network to align features at the semantic level across cross-modal data.

We have conducted serious experiments to verify the performance of CPPMN on four large-scale remote sensing cross-modal retrieval datasets. Experimental results show that our model achieves excellent performance compared to nineteen other baselines. Specifically, the proposed CPPMN outperforms the best model, RemoteCLIP, by 7.25%, 7.32%, 9.82%, and 4.20% in the mR evaluation metric on the RSICD, RSITMD, UCM, and SYDNEY datasets, respectively. In the future, we will continue to study the multi-view perception characteristics of RS images to optimize the proposed model for better applications in cross-modal retrieval tasks.

References

[1] Cong Bai, Minjing Zhang, Jinglin Zhang, Jianwei Zheng, and Shengyong Chen. Lscidmr: Large-scale satellite cloud image database for meteorological research. *IEEE Transactions on Cybernetics*, 2021.

[2] Keiller Nogueira, Samuel G Fadel, Ícaro C Dourado, Rafael de O Werneck, Javier AV Muñoz, Otávio AB Penatti, Rodrigo T Calumby, Lin Tzy Li, Jefersson A dos Santos, and Ricardo da S Torres. Exploiting convnet diversity for flooding identification. *IEEE Geoscience and Remote Sensing Letters*, 15(9):1446–1450, 2018.

[3] Cong Bai, Dongxiaoyuan Zhao, Minjing Zhang, and Jinglin Zhang. Multimodal information fusion for weather systems and clouds identification from satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:7333–7345, 2022.

[4] Yuan Sun, Zhenwen Ren, Peng Hu, Dezhong Peng, and Xu Wang. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 2023.

[5] Jie Guo, Meiting Wang, Yan Zhou, Bin Song, Yuhao Chi, Wei Fan, and Jianglong Chang. Hgan: Hierarchical graph alignment network for image-text retrieval. *IEEE Transactions on Multimedia*, 2023.

[6] Chong Liu, Yuqi Zhang, Hongsong Wang, Weihua Chen, Fan Wang, Yan Huang, Yi-Dong Shen, and Liang Wang. Efficient token-guided image-text retrieval with consistent multimodal contrastive training. *IEEE Transactions on Image Processing*, 2023.

[7] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv preprint arXiv:2204.09868*, 2022.

[8] Zhiqiang Yuan, Wenkai Zhang, Xuee Rong, Xuan Li, Jialiang Chen, Hongqi Wang, Kun Fu, and Xian Sun. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2021.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[11] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[12] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[13] Xu Tang, Yijing Wang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Interacting-enhancing feature transformer for cross-modal remote sensing image and text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[14] Yaxiong Chen, Jirui Huang, Shengwu Xiong, and Xiaoqiang Lu. Integrating multisubspace joint learning with multilevel guidance for cross-modal retrieval of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024.

[15] Jiancheng Pan, Qing Ma, and Cong Bai. A prior instruction representation framework for remote sensing image-text retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 611–620, 2023.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[19] Qimin Cheng, Yuzhuo Zhou, Peng Fu, Yuan Xu, and Liang Zhang. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4284–4297, 2021.

[20] Fuzhong Zheng, Xu Wang, Luyao Wang, Xiong Zhang, Hongze Zhu, Long Wang, and Haisu Zhang. A fine-grained semantic alignment method specific to aggregate multi-scale information for cross-modal remote sensing image retrieval. *Sensors*, 23(20):8437, 2023.

[21] Hailong Ning, Bin Zhao, and Yuan Yuan. Semantics-consistent representation learning for remote sensing image–voice retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

[22] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[23] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[24] Xiang Long, Kaipeng Deng, Guanzhong Wang, Yang Zhang, Qingqing Dang, Yuan Gao, Hui Shen, Jianguo Ren, Shumin Han, Errui Ding, et al. Pp-yolo: An effective and efficient implementation of object detector. *arXiv preprint arXiv:2007.12099*, 2020.

[25] Xu Tang, Dabiao Huang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Prior-experience-based vision-language model for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[26] Zheng Wang, Xing Xu, Jiwei Wei, Ning Xie, Yang Yang, and Heng Tao Shen. Semantics disentangling for cross-modal retrieval. *IEEE Transactions on Image Processing*, 33:2226–2237, 2024.

[27] Dongwon Kim, Namyup Kim, and Suha Kwak. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23422–23431, 2023.

[28] Taghreed Abdullah, Yakoub Bazi, Mohamad M Al Rahhal, Mohamed L Mekhalfi, Lalitha Rangarajan, and Mansour Zuair. Textrs: Deep bidirectional triplet network for matching text to remote sensing images. *Remote Sensing*, 12(3):405, 2020.

[29] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[30] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300*, 2023.

[31] An-An Liu, Bo Yang, Wenhui Li, Dan Song, Zhengya Sun, Tongwei Ren, and Zhiqiang Wei. Text-guided knowledge transfer for remote sensing image-text retrieval. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[32] Yi Bin, Haoxuan Li, Yahui Xu, Xing Xu, Yang Yang, and Heng Tao Shen. Unifying two-stream encoders with transformers for cross-modal retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3041–3050, 2023.

[33] Jie Nie, Chengyu Zheng, Chenglong Wang, Zijie Zuo, Xiaowei Lv, Shusong Yu, and Zhiqiang Wei. Scale–relation joint decoupling network for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022.

[34] Chengyu Zheng, Jie Nie, Zhaoxin Wang, Ning Song, Jingyu Wang, and Zhiqiang Wei. High-order semantic decoupling network for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[35] Fanglong Yao, Xian Sun, Nayu Liu, Changyuan Tian, Liangyu Xu, Leiyi Hu, and Chibiao Ding. Hypergraph-enhanced textual-visual matching network for cross-modal remote sensing image retrieval via dynamic hypergraph learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:688–701, 2022.

[36] Yi Bin, Wenhao Shi, Jipeng Zhang, Yujuan Ding, Yang Yang, and Heng Tao Shen. Non-autoregressive cross-modal coherence modelling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3253–3261, 2022.

[37] Gang Hu, Zaidao Wen, Yafei Lv, Jianting Zhang, and Qian Wu. Global-local information soft-alignment for cross-modal remote-sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[38] Yi Bin, Yujuan Ding, Bo Peng, Liang Peng, Yang Yang, and Tat-Seng Chua. Entity slot filling for visual captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):52–62, 2021.

[39] Yafei Lv, Wei Xiong, Xiaohan Zhang, and Yaqi Cui. Fusion-based correlation learning model for cross-modal remote sensing image retrieval. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[40] Yaxiong Chen, Jinghao Huang, Xiaoyu Li, Shengwu Xiong, and Xiaoqiang Lu. Multiscale salient alignment learning for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[41] Zuopeng Zhao, Xiaoran Miao, Chen He, Jianfeng Hu, Bingbing Min, Yumeng Gao, Ying Liu, and Kanyaphakphachsorn Pharksuwan. Masking-based cross-modal remote sensing image-text retrieval via dynamic contrastive learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[42] Haoxuan Li, Yi Bin, Junrong Liao, Yang Yang, and Heng Tao Shen. Your negative may not be true negative: Boosting image-text matching with false negative elimination. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 924–934, 2023.

[43] Ao Chen and Yue Zhou. An attention enhanced graph convolutional network for semantic segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 734–745. Springer, 2020.

[44] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[45] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

[46] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[47] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. Deep semantic understanding of high resolution remote sensing image. In *2016 International conference on computer, information and telecommunication systems (Cits)*, pages 1–5. IEEE, 2016.

[48] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010.

[49] Fan Zhang, Bo Du, and Liangpei Zhang. Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4):2175–2184, 2014.

[50] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

[51] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[52] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5764–5773, 2019.

[53] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia*, pages 12–20, 2019.

[54] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1218–1226, 2021.

[55] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15661–15670, 2022.

[56] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Yongqiang Mao, Ruixue Zhou, Hongqi Wang, Kun Fu, and Xian Sun. Mcrn: A multi-source cross-modal retrieval network for remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 115:103071, 2022.

[57] Hongfeng Yu, Fanglong Yao, Wanxuan Lu, Nayu Liu, Peiguang Li, Hongjian You, and Xian Sun. Text-image matching for cross-modal remote sensing image retrieval via graph neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:812–824, 2022.

[58] Weihang Zhang, Jihao Li, Shuoke Li, Jialiang Chen, Wenkai Zhang, Xin Gao, and Xian Sun. Hypersphere-based remote sensing cross-modal text-image retrieval via curriculum learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[59] Zhong Ji, Changxu Meng, Yan Zhang, Yanwei Pang, and Xuelong Li. Knowledge-aided momentum contrastive learning for remote-sensing image text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[60] Qing Ma, Jiancheng Pan, and Cong Bai. Direction-oriented visual-semantic embedding model for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.