

# Parameter-Efficient Transfer Learning for Remote Sensing Image-Text Retrieval

Yuan Yuan, *Senior Member, IEEE*, Yang Zhan, and Zhitong Xiong, *Member, IEEE*

**Abstract**—Vision-and-language pre-training (VLP) models have experienced a surge in popularity recently. By fine-tuning them on specific datasets, significant performance improvements have been observed in various tasks. However, full fine-tuning of VLP models not only consumes a significant amount of computational resources but also has a significant environmental impact. Moreover, as remote sensing (RS) data is constantly being updated, full fine-tuning may not be practical for real-world applications. To address this issue, in this work, we investigate the parameter-efficient transfer learning (PETL) method to effectively and efficiently transfer visual-language knowledge from the natural domain to the RS domain on the image-text retrieval task. To this end, we make the following contributions. 1) We construct a novel and sophisticated PETL framework for the RS image-text retrieval (RSITR) task, which includes the pretrained CLIP model, a multimodal remote sensing adapter, and a hybrid multimodal contrastive (HMMC) learning objective; 2) To deal with the problem of high intra-modal similarity in RS data, we design a simple yet effective HMMC loss; 3) We provide comprehensive empirical studies for PETL-based RS image-text retrieval. Our results demonstrate that the proposed method is promising and of great potential for practical applications. 4) We benchmark extensive state-of-the-art PETL methods on the RSITR task. Our proposed model only contains 0.16M training parameters, which can achieve a parameter reduction of 98.9% compared to full fine-tuning, resulting in substantial savings in training costs. Our retrieval performance exceeds traditional methods by 7-13% and achieves comparable or better performance than full fine-tuning. This work can provide new ideas and useful insights for RS vision-language tasks.

**Index Terms**—Parameter-Efficient Transfer Learning (PETL), adapter, cross-modal, remote sensing image-text retrieval.

## I. INTRODUCTION

WITH the development of Earth observation technology [1], remote sensing (RS) imagery is becoming more and more accessible, improving human’s perception of the Earth [2, 3]. However, how to efficiently convert RS imagery into actionable information is still significant research [4–7]. In order to fully exploit the potential of RS images in human-computer interaction, RS vision-language (VL) tasks have become a hot research topic in recent years. The different granularity of VL multi-modal tasks have been introduced into

This work was supported in part by grants from the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (No.CX2023030), the National Key R&D Program of China (No.2020YFB2103900), and the National Science Fund for Distinguished Young Scholars (No.61825603). (*Corresponding authors: Yuan Yuan and Zhitong Xiong.*)

Yang Zhan and Yuan Yuan are with the School of Artificial Intelligence, Optics, and Electronics (iOPEN), Northwestern Polytechnical University, Xi’an 710072, China (e-mail:zhanyang@mail.nwpu.edu.cn; y.yuan@nwpu.edu.cn).

Zhitong Xiong is with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany.

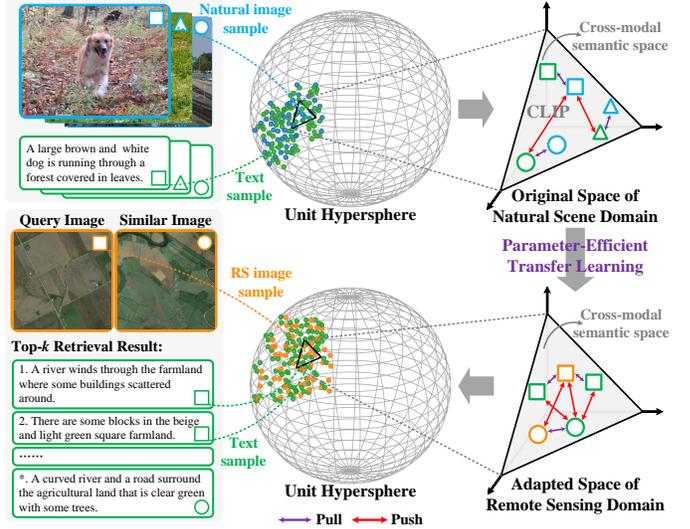


Fig. 1. The matched natural image-text samples have the same vector direction in the unit hypersphere space of the pre-trained CLIP model. The PETL method learns specific knowledge of RS domain to get the adapted space. In the original space of the natural scene domain, only the distance between different modalities is paid attention to. However, in the RS domain, it is necessary to consider samples with high intra-modal similarity to avoid the problem of matching errors.

RS data, including image level [8–11], object level [12–14], pixel level [15], and spatial-temporal level [16, 17]. These technologies have promising applications in urban planning, disaster monitoring, search and rescue activities, resource detection, and agricultural production [18–23].

Large vision-and-language pre-training (VLP) models have surged [24, 25] in recent years. In particular, contrastive vision-language pre-training (CLIP) [26] has shown great potential in multi-modal representations and can project natural image and text modalities into a joint semantic subspace. As shown in Fig. 1, the aligned image-text samples have the same vector direction on the unit hypersphere. Fine-tuning large VLP models has become a fundamental paradigm of research. However, the research of transferring the knowledge learned from image-text pairs in the natural domain to a more complex RS domain is still under-explored.

Meanwhile, VLP models in the RS domain have not been proposed due to the challenges of RS vision-language tasks. Although there are a large number of publicly available RS images, few of them are captioned and even fewer are multi-captioned. The fully fine-tuned CLIP model has achieved encouraging performance for RS image classification<sup>1</sup>. However,

<sup>1</sup><https://github.com/arampacha/CLIP-rsicc>

this approach is not feasible due to the heavy computation, memory storage, and excessive  $CO_2$  emissions. As RS images are updated, it is impractical to adapt with constant full fine-tuning on a daily basis. In this context, we explore a new research paradigm of parameter-efficient transfer learning (PETL) from the natural scene domain to the RS domain. Due to the outstanding representational capability of CLIP in VL, we use the CLIP model to study RS image-text retrieval (RSITR). This leads to a new research task, namely PETL-based RS image-text retrieval (PE-RSITR). The RSITR task can verify the performance of the adapted space of the RS domain, as shown in Fig. 1.

Nowadays, the mainstream methods of PETL are mainly divided into adapter [27] and prompt learning [28, 29]. PETL only fine-tunes a small number of parameters while keeping the parameters of the CLIP unchanged, which greatly reduces the computational cost while having comparable performance to full fine-tuning. However, existing works typically focus on downstream tasks from the same domain of the VLP models. This creates a limit that a strong VLP model with sufficient knowledge may not be available in an unknown specific domain (*e.g.*, remote sensing). Therefore, there are still many challenges to exploring PE-RSITR.

First, CLIP is pre-trained in the natural scene domain, which has domain gaps with the RS domain. To bridge the significant domain gap, the knowledge of CLIP needs to be transferred from "VL of the natural scene" to "VL of the RS". John von Neumann once said: with four parameters I can fit an elephant, and with five I can make him wiggle his trunk. Therefore, we attempt to design a method with a small number of trained parameters to explore the new knowledge of RS image-text efficiently while inheriting the prior knowledge structure of the natural scene domain appropriately. In addition, the RSITR task involves two modalities. If there is no cross-modal interaction mechanism, only suboptimal results can be obtained [30, 31]. Therefore, we further try to design a method that does not increase parameters and can accomplish cross-modal knowledge sharing. Finally, the RS image-text data is very different from the data of the natural domain. The RS images are collected by satellites from an overhead view and the intra-class similarity is extremely high due to the earth's texture. The visualization results of the textual similarity in the literature [10] show that the caption similarity is also high. Since this is not fully considered by existing methods, RSITR often results in the error of misalignment of similar RS images or captions.

To tackle the above problems, we propose a novel and sophisticated PE-RSITR framework. Although adapter-based methods have been widely explored in prior works, it is still non-trivial to design an effective adapter for the RSITR task. Based on extensive experiments and explorations, we make the following design decisions. 1) We design a more compact multimodal remote sensing adapter (MRS-Adapter) that has no skip connection and connects only once in parallel with the transformer block. 2) Inspired by the Cross-Modal Adapter [30], MRS-Adapter utilizes a linear layer for weight sharing. The shared linear layer enables the fine-grained information of RS image modality and text modality to interact, which can

enhance the RS vision language modality representation.

Furthermore, token-level data augmentation is designed to construct intra-modal positive pairs for RS images and texts. The method of data augmentation is to adopt the simple random dropout. By devising a simple yet efficient loss function of hybrid multi-modal contrastive constraints without increasing parameters, the distance between the query image (query text) and other similar images (similar texts) can be pushed to avoid matching errors. We have conducted extensive experiments on three commonly used datasets, *i.e.*, RSICD [11], RSITMD [10], and UCM [32]. First, we benchmark many state-of-the-art (SOTA) methods on the PE-RSITR task, and the adapter largely outperforms the prompt learning method. Secondly, the trained parameter of our proposed method is 0.16M, which can reduce 98.9% parameters of full fine-tuning and greatly save the training cost. Finally, our retrieval performance exceeds traditional methods by 7-13% and achieves comparable or better than full fine-tuning. Our PE-RSITR framework is both parameter-efficient and effective.

In general, our contributions can be summarized in the following aspects.

- 1) We propose a novel and sophisticated PETL framework for the RS image-text retrieval task. Specifically, the proposed framework consists of the pretrained CLIP model, the MRS-Adapter, and a hybrid multi-modal contrastive learning objective.
- 2) We design a simple yet effective loss function: the hybrid multi-modal contrastive (HMMC) loss for PETL-based RS image-text retrieval. Experimental results prove that the proposed HMMC loss is effective in further improving the performance on top of the proposed MRS-Adapter.
- 3) We provide comprehensive empirical studies for the PETL-based RS image-text retrieval task. Our qualitative and quantitative results demonstrate that the proposed method is promising and of great potential for practical applications.
- 4) Extensive experiments show that our approach can achieve a parameter reduction of 98.9% without performance sacrifice compared to full fine-tuning. Our performance exceeds traditional methods by 7-13%. The comprehensive benchmark results are insightful for future research.

This paper is organized as follows. We review the related work of RS image-text retrieval, the VLP model, and parameter-efficient transfer learning in Section II. In Section III, we present our proposed PE-RSITR framework. Evaluation methods and extensive experiment results are shown in Section IV. Finally, we conclude this work in Section V.

## II. RELATED WORK

### A. Remote Sensing Image-Text Retrieval

With the development of RS vision-language cross-modal technology, RS image-text retrieval (RSITR) is becoming a major interest. RSITR can effectively verify the performance of VL modal representations. However, due to the complexity

of RS image-text data, there have been limited related works. Some works [33–36] employ CNN (*e.g.*, VGG, ResNet) to extract RS image features and RNN (*e.g.*, LSTM, BiLSTM) to extract text features. Hoxha et al. [34] first encoded the RS image and converted it to a caption, and finally calculates the similarity with the real captions to complete the matching. Rahhal et al. [35] proposed an unsupervised image-text retrieval method for RS imagery. To reduce the occupancy and overhead of the retrieval algorithm, Yuan et al. [9] proposed a lightweight RS multiscale crossmodal retrieval model (LW-MCR), and designed distillation loss and semi-supervised loss to enhance the retrieval performance. Yuan et al. [10] also proposed an asymmetric multimodal multi-source image retrieval method that uses the multiscale self-attention module to extract salient features of RS images and utilizes the features to guide the text representation. In the RSITR framework based on global and local information (GaLR), Yuan et al. [10] indicated that RSITR should focus not only on the global features of RS images but also on the local features reflecting object relationships and saliency. Recently, multilanguage transformer [37] demonstrated that loading CLIP pre-training model [26] can achieve promising performance on RS image-text retrieval.

### B. Vision-Language Pre-training Model

Large-scale VLP models are developing rapidly and have shown encouraging results on various downstream tasks [24]. According to the encoder type, VLP models are mainly classified into fusion encoder and dual encoder [25]. The fusion encoder takes image and text features as input and uses some fusion methods for VL interaction. The fusion encoder is mainly classified into single-stream and dual-stream structures. The single-stream structure (*e.g.*, OSCAR [38], XGPT [39], SimVLM [40]) concatenates multimodal features and uses the transformer encoder in a unified framework. However, the single-stream performs the self-attention directly on two modalities, ignoring the inter-modal interaction. Therefore the dual-stream structure (*e.g.*, ViLBERT [41], ALBEF [42]) performs the cross-attention using the transformer decoder. The fusion encoder relies on a large transformer for VL interaction modeling but the inference process can be very slow in solving matching tasks such as image-text retrieval. In contrast, the dual encoder (*e.g.*, CLIP [26], ALIGN [43]) uses some simple methods for VL interaction modeling and calculates similarity scores after projecting the image and text features into the same semantic subspace. This method is more efficient for retrieval tasks, *e.g.* CLIP shows amazing results for the image-text retrieval, but does not work well when dealing with VL understanding tasks. The current pre-training models for RS image research have also been greatly developed [44]. Zhang et al. [45] proposes a transfer learning method from natural scenes to the RS domain, which can achieve good results in many tasks (*e.g.*, scene classification, object detection, and land cover classification). Unfortunately, there is no sufficiently large and uniform RS image-text dataset to support VLP models for the RS domain.

### C. Parameter-Efficient Transfer Learning

Existing methods of PETL are broadly divided into two families, *i.e.* prompt learning and adapter, and are summarized in detail as follows.

1) *Prompt Learning*: Prompt learning [28, 29] was first proposed in natural language processing (NLP). When fine-tuning large language models, task-specific learnable vectors are added to the input. Unlike full fine-tuning, prompt learning can significantly reduce storage and computational costs and can achieve comparable performance. Prompt learning has been applied to computer vision (CV). Visual Prompt Tuning [46] (VPT) can efficiently fine-tune large-scale transformer models in vision. For CLIP-based image classification, Zhou et al. [47] added continuous learnable prompts on text labels for context optimization (CoOp) and then proposed to generate prompts using image features for conditional context optimization [48] (CoCoOp). Zang et al. [49] further combined the advantages of text prompt learning and visual prompt learning to propose a tiny network to jointly optimize prompt learning for different modalities. MaPLe [50] used the V-L coupling function to generate visual prompts on textual prompts to adapt both language and vision branches simultaneously. However, there are few prompt learning methods for multimodal tasks. CPT [51] proposed a new prompt learning paradigm for visual grounding by adding color prompts to text and images, respectively. CPT used the color-masked token of the target region and color text prompts to ground the object.

2) *Adapter*: Since Houshy et al. [27] proposed the adapter module to fine-tune large pretrained models in NLP, many improved methods [52–55] have shown good performance on NLP tasks. Adapter is a lightweight plug-and-play module. The pre-trained model is frozen during fine-tuning and the parameters of the adapter are updated. Recently, the adapter is widely used to fine-tune pre-trained models in CV. Chen et al. [56] proposed Conv-Adapter which replaces the original linear layers with convolutional layers, making it possible to efficiently fine-tune largescale ConvNets. Convpass [57] is also composed of convolutional layers, but it is the adapter for vision transformer. AdapterFormer [58] is the adapter based on the original linear layers for vision transformer, and it can adapt both image and video tasks efficiently. Pan et al. [59] proposed a new Spatio-Temporal Adapter (ST-Adapter) to accomplish PETL from image models to video tasks. With the appearance of CLIP, the adapter combined with CLIP further shows superior performance. CLIP-Adapter [60] employs an additional bottleneck layer with residual connections at the end of the image and text branches, respectively. Zhang et al. [61] proposed a training-free Tip-Adapter for CLIP for few-shot classification. SVL-Adapter [62] combines the complementary advantages of CLIP and self-supervised representation learning for image classification that are significantly different from common images. Previous works had focused on unimodal tasks and less on multimodal tasks. Although VL-Adapter [63] is proposed for image-text and video-text multimodal tasks, it utilizes CLIP for image encoding and uses Adapter only in the language model. The latest Cross-Modal Adapter [30] and UniAdapter [31] both propose cross-modal interaction

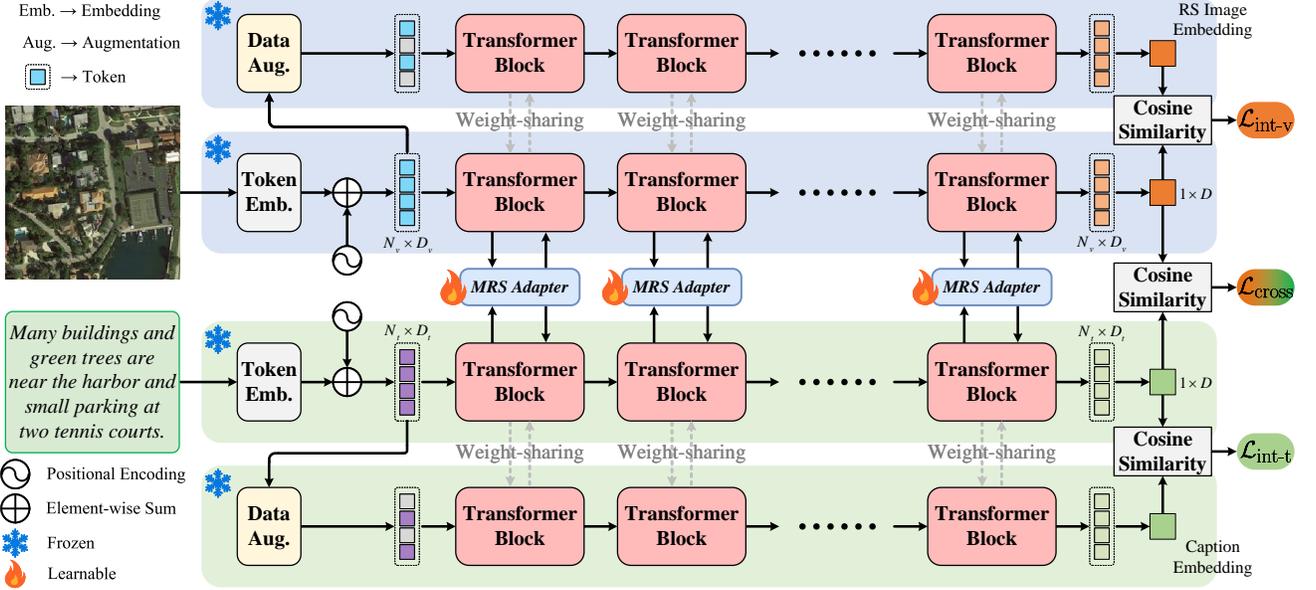


Fig. 2. Overall architecture of our proposed novel and sophisticated PE-RSITR framework. It mainly consists of three parts: the frozen CLIP pre-training backbone, the multimodal remote sensing adapter (MRS-Adapter) with cross-modal interaction, and a hybrid multi-modal contrastive (HMMC) loss.

mechanism and have shown good performance in multimodal tasks such as image/video text retrieval and visual question answering.

### III. METHODOLOGY

To capitalize a large VLP model of natural scenes for RS vision-language tasks with domain differences, such as RSITR, the intrinsic gap in different domains need to be filled. This section illustrates our proposed PE-RSITR framework in detail, and the overall framework is shown in Fig. 2.

#### A. Preliminary

In this subsection, we briefly describe how to process RS images and caption embeddings through the CLIP pre-training model and introduce the basic structure of the adapter.

**Multimodal Encoder.** We use the CLIP pre-training model as the primary multimodal encoder, including two branches of image encoder (ViT-B/32) and text encoder with the same structure. Concretely, given a cross-modal RS image-query dataset  $\mathcal{O} = \{(i_n, t_n)\}_{n=1}^N$ , where there are  $N$  pairs of image-text positive pairs  $(i_n, t_n)$ . To simplify the notations, we denote  $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$  and  $\mathbf{T} = \{w_m\}_{m=1}^M$  ( $M$  is the sentence length) as single instances of RS image and query text modality, respectively, where  $H \times W \times 3$  denotes the size of the RS image and  $w_m$  represents the  $m$ -th word. The basic architecture of the image encoder is shown in the blue background branch in Fig. 2. First, a convolution layer is used to generate the patch tokens  $\mathbf{I}_{patch} \in \mathbb{R}^{\frac{H}{s} \cdot \frac{W}{s} \times D_v}$ , where  $s$  is the stride of the backbone network and also the patch size. An additional classification (CLS) token  $\mathbf{I}_{cls} \in \mathbb{R}^{1 \times D_v}$  is added to the token sequence, and the vision transformer adds the positional embedding  $\mathbf{I}_{pos} \in \mathbb{R}^{N_v \times D_v}$  into each token. The RS image tokens are defined as

$$\mathbf{I}_0 = [\mathbf{I}_{cls}; \mathbf{I}_{patch}] + \mathbf{I}_{pos}, \quad (1)$$

where  $\mathbf{I}_0 \in \mathbb{R}^{N_v \times D_v}$ ,  $N_v = 1 + \frac{H}{s} \cdot \frac{W}{s}$ , and  $D_v$  is the hidden dimension of the vision transformer. Afterward, the RS image tokens  $\mathbf{I}_0$  pass through 12 layers of stacked transformer blocks. The  $l$ -th transformer block can be represented as

$$\hat{\mathbf{I}}_l = \text{MHA}(\text{LN}(\mathbf{I}_{l-1})) + \mathbf{I}_{l-1}, \quad (2)$$

$$\mathbf{I}_l = \text{MLP}(\text{LN}(\hat{\mathbf{I}}_l)) + \hat{\mathbf{I}}_l, \quad (3)$$

where  $\hat{\mathbf{I}}_l$  and  $\mathbf{I}_l$  respectively indicate the output of multi-head attention (MHA) and feed-forward network (FNN) modules. Finally, the CLS token is used as the global RS image embedding  $\mathbf{v}$  which is linearly projected into the  $D$ -dimensional cross-modal semantic space and then  $L_2$  normalized.

The green background branch in Fig. 2 is the text encoder. The text encoder performs token embedding similarly. First, the caption is tokenized using lower-cased byte pair encoding (BPE), denoted as  $\mathbf{T}_{token} \in \mathbb{R}^{N_t \times D_t}$ . The token sequence for each caption starts with a  $[BOS]$  token and ends with a  $[EOS]$  token. Afterwards, the positional embedding  $\mathbf{T}_{pos} \in \mathbb{R}^{N_t \times D_t}$  is added to each token. The text tokens are defined as

$$\mathbf{T}_0 = \mathbf{T}_{token} + \mathbf{T}_{pos}, \quad (4)$$

where  $\mathbf{T}_0 \in \mathbb{R}^{N_t \times D_t}$  and  $D_t$  is the hidden dimension of the text transformer. After the word embedding, the tokens  $\mathbf{T}_0$  are sent to 12 layers of stacked transformer blocks. The  $l$ -th transformer block can be represented as

$$\hat{\mathbf{T}}_l = \text{MHA}(\text{LN}(\mathbf{T}_{l-1})) + \mathbf{T}_{l-1}, \quad (5)$$

$$\mathbf{T}_l = \text{MLP}(\text{LN}(\hat{\mathbf{T}}_l)) + \hat{\mathbf{T}}_l, \quad (6)$$

where  $\hat{\mathbf{T}}_l$  and  $\mathbf{T}_l$  respectively indicate the output of MHA and FNN modules. Finally, the highest layer of the transformer at the  $[EOS]$  token is used as the global caption embedding  $\mathbf{t}$  which is linearly projected into the  $D$ -dimensional cross-modal semantic space and then  $L_2$  normalized. The only

difference between the image encoder and the text encoder is that the hidden dimension  $D_v=768$  while  $D_t=512$ .

**Adapter.** Inspired by the success of Adapter [27] in NLP, more and more adapter-based methods have shown promising performance in both CV and VL tasks. The adapter consists of a bottleneck that contains few parameters relative to the original pre-training model. Specifically, the adapter first uses a down-projection linear layer with parameters  $\mathbf{W}_{down} \in \mathbb{R}^{d \times \hat{d}}$  ( $\hat{d} \ll d$ ) to project the input features onto a low-dimensional representation. Then a non-linear activation function is used, commonly the ReLU activation function. Finally, an up-projection linear layer with parameters  $\mathbf{W}_{up} \in \mathbb{R}^{\hat{d} \times d}$  projects the features back to the input size before adding the skip connection. Formally, given an input feature  $\mathbf{X} \in \mathbb{R}^{N_{in} \times d}$ , the adapted feature  $\hat{\mathbf{X}} \in \mathbb{R}^{N_{in} \times d}$  can be calculated as

$$\mathbf{X}^{down} = \text{ReLU}(\mathbf{X} \cdot \mathbf{W}_{down}), \quad (7)$$

$$\text{Adapter}(\mathbf{X}) = \tilde{\mathbf{X}} = s \cdot \mathbf{X}^{down} \cdot \mathbf{W}_{up} + \mathbf{X}, \quad (8)$$

where  $s$  is a scalar scale factor that controls the effect of the adapter. The original adapter scheme is inserted sequentially into the MHA and FFN of the transformer block. The feature adaptation process at the  $l$ -th layer can be written as

$$\hat{\mathbf{X}}_l = \text{Adapter}(\text{MHA}(\text{LN}(\mathbf{X}))) + \mathbf{X}, \quad (9)$$

$$\mathbf{X}_l = \text{Adapter}(\text{MLP}(\text{LN}(\hat{\mathbf{X}}_l))) + \hat{\mathbf{X}}_l. \quad (10)$$

### B. MRS-Adapter

Various CLIP-based adapter methods show great potential for VL tasks. The core of extending CLIP to the RS domain to accomplish RSITR lies in exploring the VL expert knowledge in the RS domain efficiently while appropriately inheriting the VL prior knowledge structure of the natural scene domain. This work is still under-explored. If the adapter is directly extended from NLP to both modalities of VL, it can only lead to sub-optimal results due to the lack of cross-modal interactions. This point was also verified in the recent works [30, 31]. Therefore, we attempt to design an adapter that can share knowledge between RS image modality and text modality without increasing parameters. Finally, we found that the same cross-modal interaction as Cross-Modal Adapter is the most effective way and can reduce parameters. Our MRS-Adapter is extremely similar to Cross-Modal Adapter, but our scheme is more concise. MRS-Adapter has no skip connection and is only connected in parallel with the FFN module, which can further reduce the number of adapters. The specific structure is shown in Fig. 3. Formally, the input of  $l$ -th layer is the  $\hat{\mathbf{I}}_l \in \mathbb{R}^{N_v \times D_v}$  and  $\hat{\mathbf{T}}_l \in \mathbb{R}^{N_t \times D_t}$  of the MHA module output, and the adapted features can be obtained as follows:

$$\hat{\mathbf{I}}_l^{down} = \text{ReLU}(\hat{\mathbf{I}}_l \cdot \mathbf{W}_{down}^v), \quad (11)$$

$$\hat{\mathbf{I}}_l^{\text{MRS-Adapter}} = [\hat{\mathbf{I}}_l^{down} \cdot \mathbf{W}_{up}^v; \hat{\mathbf{I}}_l^{down} \cdot \mathbf{W}_{up}^{share}], \quad (12)$$

$$\hat{\mathbf{T}}_l^{down} = \text{ReLU}(\hat{\mathbf{T}}_l \cdot \mathbf{W}_{down}^t), \quad (13)$$

$$\hat{\mathbf{T}}_l^{\text{MRS-Adapter}} = [\hat{\mathbf{T}}_l^{down} \cdot \mathbf{W}_{up}^t; \hat{\mathbf{T}}_l^{down} \cdot \mathbf{W}_{up}^{share}], \quad (14)$$

where  $d \ll D_v$ ,  $d \ll D_t$ ,  $0 < r < D_v$ ,  $0 < r < D_t$ ,  $\mathbf{W}_{down}^v \in \mathbb{R}^{D_v \times d}$  and  $\mathbf{W}_{down}^t \in \mathbb{R}^{D_t \times d}$  are the modality-specific down-projection weights of two branches,  $\mathbf{W}_{up}^{share} \in$

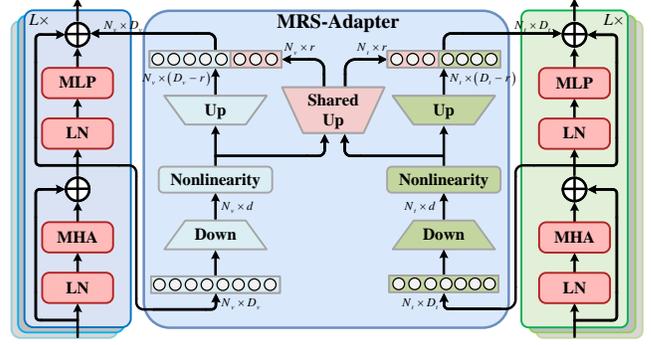


Fig. 3. The implementation details of our MRS-Adapter. MRS-Adapter is inserted in parallel with the FFN module and adds a modality-shared up-projection to connect the original modality-specific up-projection linear layer for cross-modal weight sharing.

$\mathbb{R}^{d \times r}$  is modality-shared weights, and  $\mathbf{W}_{up}^v \in \mathbb{R}^{d \times D_v - r}$  and  $\mathbf{W}_{up}^t \in \mathbb{R}^{d \times D_t - r}$  are the modality-specific up-projection weights of two branches. Finally, the MRS-Adapter is connected in parallel with the FFN module at each layer of both branches, and the feature adaptation process at  $l$ -th layer can be written as

$$\hat{\mathbf{I}}_l = \text{MHA}(\text{LN}(\mathbf{I}_{l-1})) + \mathbf{I}_{l-1}, \quad (15)$$

$$\mathbf{I}_l = \text{MLP}(\text{LN}(\hat{\mathbf{I}}_l)) + \hat{\mathbf{I}}_l + \hat{\mathbf{I}}_l^{\text{MRS-Adapter}}, \quad (16)$$

$$\hat{\mathbf{T}}_l = \text{MHA}(\text{LN}(\mathbf{T}_{l-1})) + \mathbf{T}_{l-1}, \quad (17)$$

$$\mathbf{T}_l = \text{MLP}(\text{LN}(\hat{\mathbf{T}}_l)) + \hat{\mathbf{T}}_l + \hat{\mathbf{T}}_l^{\text{MRS-Adapter}}. \quad (18)$$

MRS-Adapter adding an  $r$ -dimensional linear layer for weight sharing can directly reduce  $d \times r$  parameters. The shared up-projection enables the fine-grained information of RS image modality and text modality to interact, which can enhance the RS vision language modality representation. MRS-Adapter can learn the VL knowledge specific to the RS domain and efficiently extend the natural scene domain to the RS domain.

### C. HMMC Learning Objective

In image-text retrieval tasks, the bi-directional triplet loss established by Faghri et al. [64] has become the mainstream loss function. The bi-directional triplet loss can pull the distance between this sample and the positive sample of another modality closer while pushing the distance between it and the negative sample of another modality farther. However, this framework can only constrain the inter-modal samples and does not consider the intra-modal samples. In particular, RS images are characterized by extremely high intra-class similarity, and the visualization results of the textual similarity in the literature [10] show that the textual similarity is also extremely high. Therefore, RSITR often results in the error of retrieving misalignment of similar RS images or captions. To cope with the problem, token-level data augmentation is adopted to construct intra-modal positive pairs for RS images and texts. We design a simple yet effective hybrid multi-modal contrastive loss function. The framework of inter-modal and intra-modal cooperative constraints is shown in Fig. 2. Inspired

by contrastive learning [65], the method of data augmentation is to adopt the simple random dropout.

**Intra-Modal Constraints.** The bi-directional triplet loss with hard negatives is used, and the similarity between sample pairs is measured by cosine similarity. For RS image modality, data augmentation is performed after token embedding, and a token-level positive pair  $(I_0, I_0^+)$  is obtained for each RS image  $I$ . The process is denoted as

$$I_0^+ = \text{random\_dropout}(I_0). \quad (19)$$

Then we perform the intra-modal constraint of the RS image to pull the distance from the query image and other similar images. We denote the final embedding of  $I_0^+$  as  $v^+$ . For a positive pair  $(v, v^+)$ , the intra-modal triplet loss we adopt is:

$$\begin{aligned} \mathcal{L}_{\text{intra-v}}(v, v^+) &= \sum_{\hat{v}^+} [\alpha_v - \cos(v, v^+) + \cos(v, \hat{v}^+)]_+ \\ &+ \sum_{\hat{v}} [\alpha_v - \cos(v, v^+) + \cos(\hat{v}, v^+)]_+, \end{aligned} \quad (20)$$

where  $\alpha_v$  is the margin of the intra-modal constraint of the RS image and  $[x]_+ = \max(x, 0)$ . Similarly, the text branch also uses random dropout to generate a token-level positive pair  $(T_0, T_0^+)$  for each caption  $T$ , calculated as follows:

$$T_0^+ = \text{random\_dropout}(T_0). \quad (21)$$

Then we perform the intra-modal constraint of the caption text to pull the distance from the query caption and other similar captions. We denote the final embedding of  $T_0^+$  as  $t^+$ . For a positive pair  $(t, t^+)$ , the intra-modal triplet loss is:

$$\begin{aligned} \mathcal{L}_{\text{intra-t}}(t, t^+) &= \sum_{\hat{t}^+} [\alpha_t - \cos(t, t^+) + \cos(t, \hat{t}^+)]_+ \\ &+ \sum_{\hat{t}} [\alpha_t - \cos(t, t^+) + \cos(\hat{t}, t^+)]_+, \end{aligned} \quad (22)$$

where  $\alpha_t$  is the margin of the intra-modal constraint of the caption text.

**Cross-Modal Constraint.** The multi-modal alignment of RS image-text is promoted by relying on the global similarity of RS image-text. We compute the cross-modal constraint loss with

$$\begin{aligned} \mathcal{L}_{\text{cross}}(v, t) &= \sum_{\hat{t}} [\lambda - \cos(v, t) + \cos(v, \hat{t})]_+ \\ &+ \sum_{\hat{v}} [\lambda - \cos(v, t) + \cos(\hat{v}, t)]_+, \end{aligned} \quad (23)$$

where  $\lambda$  is the margin of the cross-modal constraint.

**Overall Objective.** By combining hard-negative-based intra-modal constraints loss with cross-modal constraints loss, we obtain the HMMC loss:

$$\mathcal{L}(v, t) = \mathcal{L}_{\text{cross}}(v, t) + \mathcal{L}_{\text{intra-v}}(v, v^+) + \mathcal{L}_{\text{intra-t}}(t, t^+). \quad (24)$$

## IV. EXPERIMENTS

In this section, we first describe the dataset, evaluation metrics, and experimental setup details in Section IV-A and Section IV-B. Further, Section IV-C introduces the SOTA approaches and provides comparisons of retrieval performance. Section IV-D conducts result analyses. In Section IV-E, we perform sufficient ablation experiments. Finally, we present some visualization results to further analyze in Section IV-F.

### A. Dataset and Evaluation Metrics

We evaluate our proposed PE-RSITR framework on the three widely used RS image-text datasets: RSICD [11], RSITMD [10], and UCM [32]. RSICD is the dataset with the largest number of samples, while RSITMD is the dataset with more fine-grained captions and more challenges. UCM requires the model to be robust because of the small numbers.

Two evaluation metrics Recall at  $K$  ( $R@K$ ,  $K=1, 5$ , and  $10$ ) and mean recall (mR) are exploited to assess our model.  $R@K$  aims to calculate the ratio of queries that successfully retrieve the ground truth as one of the first  $K$  results.  $mR$  represents the average of  $R@K$  for both the text retrieval and image retrieval, which evaluates the overall retrieval performance and can be formulated in the equation below,

$$mR = \frac{\underbrace{(R@1 + R@5 + R@10)}_{\text{Text retrieval}} + \underbrace{(R@1 + R@5 + R@10)}_{\text{Image retrieval}}}{6}. \quad (25)$$

### B. Implementation Details

All experiments in this work are conducted on one NVIDIA RTX 3090 24GB GPU. We follow the data partitioning approach of Yuan et al. [10] and use 80%, 10%, and 10% of the dataset as the training set, validation set, and test set, respectively. For the RS image, we resize the image size to a fixed size of  $224 \times 224$  for training. We set the dimension  $d$  and  $r$  to 64 and the probability of random dropout to 0.2. The margin  $\lambda$ ,  $\alpha_v$ ,  $\alpha_t$  are set to 0.2 for the triplet loss calculation. We set the initial learning rate of our network to 0.0002 for trained parameters and weight decay by 0.7 every 20 epochs. During training, we adopt the Adam optimizer to train our network with a batch size of 16 for 30 epochs. To make the experiment more convincing, we follow the works in GaLR [8] and MCRN [66] to conduct the experiments and report results. We leverage k-fold cross-validation to obtain an average result, and k is set to 5.

### C. Comparisons with State-of-the-art Methods

In this experiment, we comprehensively compare our proposed method with traditional cross-modal retrieval methods and CLIP-based methods.

**Traditional methods:** Following the previous literature, we also compare the proposed method with the progressive image-text retrieval models (VSE++ [64], SCAN [67], CAMP [68], MTFN [69], LW-MCR [9], AMFMN [10], GaLR [8]) on three RS image-text datasets. For these methods, we use the results in three literature [8–10]. In addition, we have added two latest RS image-text retrieval methods.

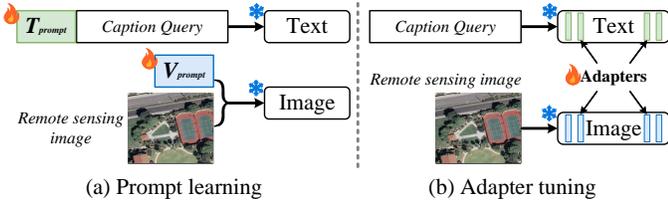


Fig. 4. (a): Prepending a sequence of learnable prompt tokens to the input tokens of the visual or text encoder. Only these added prompt tokens are updated during fine-tuning; (b): Insert a lightweight adapter into the visual or text encoder and update the parameters of the adapter during fine-tuning.

- *MCRN* [66]: MCRN constructs a multi-source cross-modal retrieval network capable of image modality, text modality, and audio modality alignment based on shared networks of pattern memory and generative adversarial theory.
- *CABIR* [70]: CABIR proposes a cross-attention model based on region-level semantic features of RS images, with textual semantics to allocate weights and filter redundant features for image regions.

**CLIP-based methods:** Single-Language [37] is loaded with CLIP pre-training parameters for training. In addition, we benchmark extensive efficient and commonly used PETL approaches in our PE-RSITR task. In order to assess the merits of our proposed method, we report our performance and compare it with the following methods.

- *Zero-shot CLIP* [26]: Testing directly on the sample of the unseen RS domain.
- *Linear Probe*: Adding an extra linear layer on top of each of the two branches of the backbone and freezing all the parameters except the parameters in the linear layer.
- *Full Fine-tuning*: Fully updating all the parameters. In the experiments of the RSICD dataset, Full Fine-tuning utilizes the weight of the CLIP-rsicc model<sup>1</sup>.
- *Prompt Learning*: As shown in Fig. 4(a), prepending a sequence of learnable prompt tokens to the input tokens, and only these added prompt tokens are updated during fine-tuning. Specifically, we compare CoOp [47] (added in the text tokens) and VPT [46] (added in the visual tokens). Following Jiang et al. [30], we applied both the visual and text tokens, called VL-Prompt.
- *Adapter Tuning*: Adapter is a lightweight plug-and-play module. The pre-trained model is frozen during fine-tuning and the parameters of adapters are updated, as shown in Fig. 4(b). CLIP-Adapter [60] employs an additional bottleneck layer to learn new features and make residual connections. Adapter [27] is for the language transformer, Convpass [57] is the convolutional adapter for ViT, and AdaptFormer [58] is for adapting ViT to different image and video tasks. UniAdapter [31] and Cross-Modal Adapter [30] are multimodal adapters. UniAdapter cannot be used directly due to the different dimensions of the visual branch and text branch. We use a linear layer to project visual features to 512 dimensions.

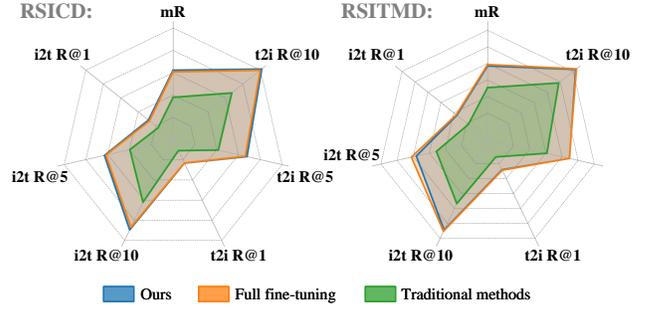


Fig. 5. Comparison with the RSITR results of full fine-tuning and the best results of traditional methods on RSICD and RSITMD datasets.

#### D. Results Analysis

**Results on three datasets.** Tables I, II, and III show the test results on the three datasets: RSICD, RSITMD, and UCM. We present results using various traditional methods and CLIP-based fine-tuning methods. Fig. 5 shows the retrieval results of full fine-tuning and the best results of traditional methods on RSICD and RSITMD datasets. Our method outperforms the traditional methods by 7-13% and even surpasses the full fine-tuning. CLIP possesses a more complex and larger network structure with a higher number of parameters. Compared to traditional methods, PETL, based on CLIP, can achieve significant performance gains by leveraging the powerful generalization ability and rich visual-linguistic prior knowledge obtained through pre-training on massive datasets of natural scenes.

We provide comprehensive empirical studies for PETL-based RS image-text retrieval. We observe that our PE-RSITR framework outperforms other works on the RSICD dataset. Compared with full fine-tuning, our results are improved by about 1%. Currently, the RSITMD dataset is the highest fine-grained and most challenging RS image-text data. Our approach can obtain optimal and suboptimal results, achieving comparable results to full fine-tuning. The UCM dataset challenges the robustness of the model due to its small size. As the results are shown in Table III, our method has the most optimal results and exceeds the effect of full fine-tuning. This result demonstrates the robustness of our PE-RSITR framework.

Single Language adopts the dual-transformer structure and loads the pre-trained parameters of CLIP to train. It achieves an average improvement of 8% in retrieval performance over the best traditional methods on the three datasets. Zero-shot CLIP, Linear probe, and Full fine-tuning serve as the fundamental baselines for PETL. Zero-shot CLIP directly employs the pre-trained model for testing, resulting in the poorest performance on the three datasets. Full fine-tuning updates all parameters and theoretically yields optimal results. However, due to the limited data size of UCM, full fine-tuning tends to overfit and fails to achieve the best performance, as shown in Table III.

**Performance analysis on prompt learning.** In prompt learning, the VPT achieves superior performance, surpassing traditional methods. The CoOp and VL-Prompt can only achieve performance similar to traditional methods. Due to the significant visual feature gaps between natural and RS

TABLE I

THE RETRIEVAL PERFORMANCE OF STATE-OF-THE-ART METHODS ON RSICD TEST SET. THE BEST PERFORMANCE OF TRADITIONAL METHODS IS WITH BOLD. RED AND BLUE REPRESENT THE BEST AND SECOND PERFORMANCE IN CLIP-BASED AND TRANSFER LEARNING METHODS.

Methods	Reference	Backbone (image/text)	Params	Text retrieval			Image retrieval			mR
				R@1	R@5	R@10	R@1	R@5	R@10	
<i>Traditional methods</i>										
VSE++ [64]	<i>BMVC'18</i>	ResNet18, GRU	15.78M	3.38	9.51	17.46	2.82	11.32	18.10	10.43
SCAN t-i [67]	<i>ECCV'18</i>	Faster R-CNN, biGRU	13.68M	4.39	10.90	17.64	3.91	16.20	26.49	13.25
SCAN i-t [67]	<i>ECCV'18</i>	Faster R-CNN, biGRU	13.68M	5.85	12.89	19.84	3.71	16.40	26.73	14.23
CAMP-triplet [68]	<i>ICCV'19</i>	Faster R-CNN, biGRU	27.03M	5.12	12.89	21.12	4.15	15.23	27.81	14.39
CAMP-bce [68]	<i>ICCV'19</i>	Faster R-CNN, biGRU	27.03M	4.20	10.24	15.45	2.72	12.76	22.89	11.38
MTFN [69]	<i>ACM MM'19</i>	Faster R-CNN, biGRU	77.90M	5.02	12.52	19.74	4.90	17.17	29.49	14.81
LW-MCR-b [9]	<i>TGRS'22</i>	SqueezeNet, -	1.65M	4.57	13.71	20.11	4.02	16.47	28.23	14.52
LW-MCR-d [9]	<i>TGRS'22</i>	SqueezeNet, -	1.65M	3.29	12.52	19.93	4.66	17.51	30.02	14.66
LW-MCR-u [9]	<i>TGRS'22</i>	SqueezeNet, -	1.65M	4.39	13.35	20.29	4.30	18.85	32.34	15.59
AMFMN-soft [10]	<i>TGRS'22</i>	ResNet18, biGRU	35.94M	5.05	14.53	21.57	5.05	19.74	31.04	16.02
AMFMN-fusion [10]	<i>TGRS'22</i>	ResNet18, biGRU	35.94M	5.39	15.08	23.40	4.90	18.28	31.44	16.42
AMFMN-sim [10]	<i>TGRS'22</i>	ResNet18, biGRU	35.94M	5.21	14.72	21.57	4.08	17.00	30.60	15.53
MCRN [66]	<i>JAG'22</i>	ResNet18, biGRU	52.35M	6.59	19.40	30.28	5.03	19.38	32.99	18.95
CABIR [70]	<i>AS'22</i>	ResNet152, BERT+biGRU	-	<b>8.59</b>	16.27	24.13	<b>5.42</b>	<b>20.77</b>	<b>33.58</b>	18.12
GaLR w/o MR [8]	<i>TGRS'22</i>	ResNet18, biGRU	46.89M	6.50	18.91	29.70	5.11	19.57	31.92	18.62
GaLR with MR [8]	<i>TGRS'22</i>	ResNet18, biGRU	46.89M	6.59	<b>19.85</b>	<b>31.04</b>	4.69	19.48	32.13	<b>18.96</b>
<i>CLIP-based methods</i>										
Single Language [37]	<i>JSTARS'22</i>	CLIP(ViT-B-32)	151M	10.70	29.64	41.53	9.14	28.96	44.59	27.42
Zero-shot CLIP [26]	<i>ICML'21</i>	CLIP(ViT-B-32)	0.00M	6.77	15.37	23.15	5.01	15.75	24.21	15.04
Linear probe [26]	<i>ICML'21</i>	CLIP(ViT-B-32)	1.05M	8.46	24.41	37.72	7.81	25.89	42.47	24.46
Full fine-tuning <sup>1</sup>	-	CLIP(ViT-B-32)	151M	<b>13.54</b>	<b>30.83</b>	<b>43.46</b>	<b>11.55</b>	<b>33.14</b>	<b>49.83</b>	<b>30.39</b>
CoOp [47]	<i>IJCV'22</i>	CLIP(ViT-B-32)	0.10M	6.32	15.89	27.60	4.31	17.89	31.73	17.29
VPT [46]	<i>ECCV'22</i>	CLIP(ViT-B-32)	0.46M	7.23	19.40	30.77	6.94	25.26	40.73	21.72
VL-Prompt	-	CLIP(ViT-B-32)	0.47M	6.47	17.05	28.68	6.69	22.62	36.16	19.61
Adapter [27]	<i>ICML'19</i>	CLIP(ViT-B-32)	0.17M	8.73	24.73	37.81	8.43	26.02	43.33	24.84
CLIP-Adapter [60]	<i>arXiv'21</i>	CLIP(ViT-B-32)	0.52M	7.11	19.48	31.01	7.67	24.87	39.73	21.65
Convpass [57]	<i>arXiv'22</i>	CLIP(ViT-B-32)	0.14M	6.54	19.67	32.78	7.03	23.08	39.15	21.38
AdaptFormer [58]	<i>NIPS'22</i>	CLIP(ViT-B-32)	0.17M	12.46	28.49	41.86	9.09	29.89	46.81	28.10
Cross-Modal Adapter [30]	<i>arXiv'22</i>	CLIP(ViT-B-32)	0.16M	11.18	27.31	40.62	9.57	30.74	48.36	27.96
UniAdapter [31]	<i>arXiv'23</i>	CLIP(ViT-B-32)	0.55M	12.65	30.81	42.74	9.61	30.06	47.16	28.84
Ours	-	CLIP(ViT-B-32)	0.16M	<b>14.13</b>	<b>31.51</b>	<b>44.78</b>	<b>11.63</b>	<b>33.92</b>	<b>50.73</b>	<b>31.12</b>

TABLE II

THE RETRIEVAL PERFORMANCE OF STATE-OF-THE-ART METHODS ON RSITMD TESE SET. THE BEST PERFORMANCE OF TRADITIONAL METHODS IS WITH BOLD. RED AND BLUE REPRESENT THE BEST AND SECOND PERFORMANCE IN CLIP-BASED AND TRANSFER LEARNING METHODS.

Methods	Reference	Backbone (image/text)	Params	Text retrieval			Image retrieval			mR
				R@1	R@5	R@10	R@1	R@5	R@10	
<i>Traditional methods</i>										
VSE++ [64]	<i>BMVC'18</i>	ResNet18, GRU	15.78M	10.38	27.65	39.60	7.79	24.87	38.67	24.83
SCAN t-i [67]	<i>ECCV'18</i>	Faster R-CNN, biGRU	13.68M	10.18	28.53	38.49	10.10	28.98	43.53	26.64
SCAN i-t [67]	<i>ECCV'18</i>	Faster R-CNN, biGRU	13.68M	11.06	25.88	39.38	9.82	29.38	42.12	26.28
CAMP-triplet [68]	<i>ICCV'19</i>	Faster R-CNN, biGRU	27.03M	11.73	26.99	38.05	8.27	27.79	44.34	26.20
CAMP-bce [68]	<i>ICCV'19</i>	Faster R-CNN, biGRU	27.03M	9.07	23.01	33.19	5.22	23.32	38.36	22.03
MTFN [69]	<i>ACM MM'19</i>	Faster R-CNN, biGRU	77.90M	10.40	27.65	36.28	9.96	31.37	45.84	26.92
LW-MCR-b [9]	<i>TGRS'22</i>	SqueezeNet, -	1.65M	9.07	22.79	38.05	6.11	27.74	49.56	25.55
LW-MCR-d [9]	<i>TGRS'22</i>	SqueezeNet, -	1.65M	10.18	28.98	39.82	7.79	30.18	49.78	27.79
LW-MCR-u [9]	<i>TGRS'22</i>	SqueezeNet, -	1.65M	9.73	26.77	37.61	9.25	34.07	54.03	28.58
AMFMN-soft [10]	<i>TGRS'22</i>	ResNet18, biGRU	35.94M	11.06	25.88	39.82	9.82	33.94	51.90	28.74
AMFMN-fusion [10]	<i>TGRS'22</i>	ResNet18, biGRU	35.94M	11.06	29.20	38.72	9.96	34.03	52.96	29.32
AMFMN-sim [10]	<i>TGRS'22</i>	ResNet18, biGRU	35.94M	10.63	24.78	41.81	<b>11.51</b>	34.69	<b>54.87</b>	29.72
MCRN [66]	<i>JAG'22</i>	ResNet18, biGRU	52.35M	13.27	29.42	41.59	9.42	35.53	52.74	30.33
GaLR w/o MR [8]	<i>TGRS'22</i>	ResNet18, biGRU	46.89M	13.05	30.09	<b>42.70</b>	10.47	36.34	53.35	31.00
GaLR with MR [8]	<i>TGRS'22</i>	ResNet18, biGRU	46.89M	<b>14.82</b>	<b>31.64</b>	42.48	11.15	<b>36.68</b>	51.68	<b>31.41</b>
<i>CLIP-based methods</i>										
Single Language [37]	<i>JSTARS'22</i>	CLIP(ViT-B-32)	151M	19.69	40.26	54.42	17.61	49.73	66.59	41.38
Zero-shot CLIP [26]	<i>ICML'21</i>	CLIP(ViT-B-32)	0.00M	9.29	26.33	37.39	7.79	23.67	38.89	23.89
Linear probe [26]	<i>ICML'21</i>	CLIP(ViT-B-32)	1.05M	17.02	33.12	48.35	13.33	41.80	63.89	36.25
Full fine-tuning	-	CLIP(ViT-B-32)	151M	<b>24.16</b>	<b>47.12</b>	<b>61.28</b>	<b>20.40</b>	<b>50.53</b>	<b>68.54</b>	<b>45.33</b>
CoOp [47]	<i>IJCV'22</i>	CLIP(ViT-B-32)	0.10M	12.19	30.69	42.82	9.16	33.85	54.35	30.51
VPT [46]	<i>ECCV'22</i>	CLIP(ViT-B-32)	0.46M	14.98	32.05	40.15	15.97	41.35	60.35	34.14
VL-Prompt	-	CLIP(ViT-B-32)	0.47M	12.81	31.28	42.64	12.61	36.20	58.84	32.40
Adapter [27]	<i>ICML'19</i>	CLIP(ViT-B-32)	0.17M	13.75	27.64	39.96	12.89	40.09	59.91	32.37
CLIP-Adapter [60]	<i>arXiv'21</i>	CLIP(ViT-B-32)	0.52M	12.83	28.84	39.05	13.30	40.20	60.06	32.38
Convpass [57]	<i>arXiv'22</i>	CLIP(ViT-B-32)	0.14M	16.03	30.16	40.26	12.05	38.66	58.11	32.55
AdaptFormer [58]	<i>NIPS'22</i>	CLIP(ViT-B-32)	0.17M	16.71	30.16	42.91	14.27	41.53	61.46	34.81
Cross-Modal Adapter [30]	<i>arXiv'22</i>	CLIP(ViT-B-32)	0.16M	18.16	36.08	48.72	16.31	44.33	64.75	38.06
UniAdapter [31]	<i>arXiv'23</i>	CLIP(ViT-B-32)	0.55M	19.86	36.32	51.28	17.54	44.89	65.46	39.23
Ours	-	CLIP(ViT-B-32)	0.16M	<b>23.67</b>	<b>44.07</b>	<b>60.36</b>	<b>20.10</b>	<b>50.63</b>	<b>67.97</b>	<b>44.47</b>

TABLE III

THE RETRIEVAL PERFORMANCE OF STATE-OF-THE-ART METHODS ON UCM TESE SET. THE BEST PERFORMANCE OF TRADITIONAL METHODS IS WITH BOLD. RED AND BLUE REPRESENT THE BEST AND SECOND PERFORMANCE IN CLIP-BASED AND TRANSFER LEARNING METHODS.

Methods	Reference	Backbone (image/text)	Params	Text retrieval			Image retrieval			mR
				R@1	R@5	R@10	R@1	R@5	R@10	
<i>Traditional methods</i>										
VSE++ [64]	<i>BMVC'18</i>	ResNet18, GRU	15.78M	12.38	44.76	65.71	10.10	31.80	56.85	36.93
SCAN t-i [67]	<i>ECCV'18</i>	Faster R-CNN, biGRU	13.68M	14.29	45.71	67.62	12.76	50.38	77.24	44.67
SCAN i-t [67]	<i>ECCV'18</i>	Faster R-CNN, biGRU	13.68M	12.85	47.14	69.52	12.48	46.86	71.71	43.43
CAMP-triplet [68]	<i>ICCV'19</i>	Faster R-CNN, biGRU	27.03M	10.95	44.29	65.71	9.90	46.19	76.29	42.22
CAMP-bce [68]	<i>ICCV'19</i>	Faster R-CNN, biGRU	27.03M	14.76	46.19	67.62	11.71	47.24	76.00	43.92
MTFN [69]	<i>ACM MM'19</i>	Faster R-CNN, biGRU	77.90M	10.47	47.62	64.29	<b>14.19</b>	52.38	78.95	44.65
LW-MCR-b [9]	<i>TGRS'22</i>	SqueezeNet, -	1.65M	12.38	43.81	59.52	12.00	46.38	72.48	41.10
LW-MCR-d [9]	<i>TGRS'22</i>	SqueezeNet, -	1.65M	15.24	<b>51.90</b>	62.86	11.90	50.95	75.24	44.68
LW-MCR-u [9]	<i>TGRS'22</i>	SqueezeNet, -	1.65M	<b>18.10</b>	47.14	63.81	13.14	50.38	79.52	45.35
AMFMN-soft [10]	<i>TGRS'22</i>	ResNet18, biGRU	35.94M	12.86	<b>51.90</b>	66.67	<b>14.19</b>	51.71	78.48	45.97
AMFMN-fusion [10]	<i>TGRS'22</i>	ResNet18, biGRU	35.94M	16.67	45.71	68.57	12.86	53.24	79.43	46.08
AMFMN-sim [10]	<i>TGRS'22</i>	ResNet18, biGRU	35.94M	14.76	49.52	68.10	13.43	51.81	76.48	45.68
CABIR [70]	<i>AS'22</i>	ResNet152, BERT+biGRU	-	15.17	45.71	<b>72.85</b>	12.67	<b>54.19</b>	<b>89.23</b>	<b>48.30</b>
<i>CLIP-based methods</i>										
Single Language [37]	<i>JSTARS'22</i>	CLIP(ViT-B-32)	151M	<b>19.04</b>	53.33	77.61	<b>19.33</b>	<b>64.00</b>	91.42	54.12
Zero-shot CLIP [26]	<i>ICML'21</i>	CLIP(ViT-B-32)	0.00M	10.95	34.76	55.71	7.33	35.14	54.57	33.08
Linear probe [26]	<i>ICML'21</i>	CLIP(ViT-B-32)	1.05M	13.33	52.71	77.62	14.43	59.42	90.28	51.30
Full fine-tuning	-	CLIP(ViT-B-32)	151M	17.14	<b>55.24</b>	79.52	13.90	56.95	91.81	52.43
CoOp [47]	<i>IJCV'22</i>	CLIP(ViT-B-32)	0.10M	7.19	42.76	74.19	9.78	48.34	87.41	44.95
VPT [46]	<i>ECCV'22</i>	CLIP(ViT-B-32)	0.46M	8.86	47.81	78.62	13.64	55.61	94.05	49.77
VL-Prompt	-	CLIP(ViT-B-32)	0.47M	7.97	45.48	75.96	12.51	51.57	91.52	47.50
Adapter [27]	<i>ICML'19</i>	CLIP(ViT-B-32)	0.17M	10.17	49.73	80.08	18.20	62.47	<b>95.77</b>	52.74
CLIP-Adapter [60]	<i>arXiv'21</i>	CLIP(ViT-B-32)	0.52M	9.83	42.52	66.79	14.61	51.03	84.14	44.82
Convpass [57]	<i>arXiv'22</i>	CLIP(ViT-B-32)	0.14M	16.46	54.79	78.78	14.24	58.67	94.51	52.91
AdaptFormer [58]	<i>NIPS'22</i>	CLIP(ViT-B-32)	0.17M	16.92	54.39	77.46	18.74	<b>63.49</b>	91.16	53.69
Cross-Modal Adapter [30]	<i>arXiv'22</i>	CLIP(ViT-B-32)	0.16M	13.77	50.57	<b>81.41</b>	17.43	61.45	95.48	53.62
UniAdapter [31]	<i>arXiv'23</i>	CLIP(ViT-B-32)	0.55M	14.46	55.19	<b>83.95</b>	16.74	61.43	<b>95.76</b>	<b>54.59</b>
Ours	-	CLIP(ViT-B-32)	0.16M	<b>22.71</b>	<b>55.81</b>	80.33	<b>18.82</b>	62.84	93.72	<b>55.71</b>

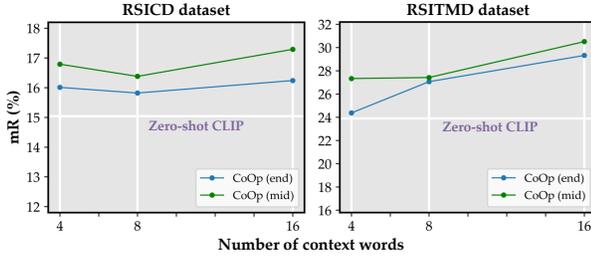


Fig. 6. Ablation on CoOp’s context length. We vary the number of the context length for CoOp-end and CoOp-mid.

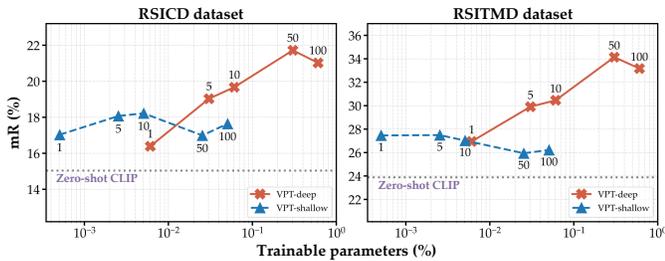


Fig. 7. Ablation on VPT’s prompt length. We vary the number of prompts for VPT-deep and VPT-shallow.

domains, adding learnable prompts into visual tokens (VPT) results in a 3-5% higher retrieval performance compared to text tokens (CoOp). However, the VL-Prompt directly adds prompts to both modalities, without considering multimodal interaction. The increase in prompts makes it more susceptible to overfitting and reduces its generalization ability.

Prompt length is the only additional hyper-parameter needed

to tune for prompt learning compared to full fine-tuning. We explore the impact of different prompt lengths on this task and analyze the performance of prompt learning in detail. Specifically, we follow the setting of CoOp [47] to vary the context length from 4 to 8 to 16 on RSICD and RSITMD datasets. “end” or “mid” means putting the learnable token in the end or middle. The results are shown in Fig. 6, which shows that the learnable token in the middle position is better than the end, and more context tokens lead to better performance. However, too many tokens may lead to overfitting. Likewise, we follow the setting of VPT [46] to vary the prompt length from 1 to 100 to explore two variants: VPT-deep and VPT-shallow. As shown in Fig. 7, VPT-deep significantly outperforms VPT-shallow. The prompt length has a large impact on the VPT-deep results and shows that the optimal prompt length is 50 for different datasets. More prompts will cause overfitting and more resource consumption.

**Performance analysis on adapter.** Observing the results in Tables I, II, and III, we analyze the performance of adapter tuning. CLIP-Adapter [60] only adds a bottleneck layer at the end of the CLIP, making the transfer to the RS domain insufficient. The Adapter [27] for language models can only tune textual features alone on the CLIP, and the methods [57, 58] for vision models can only tune visual features alone. Using the adapter to adapt the representation in a single branch (image or text) of CLIP is sub-optimal because it does not allow the flexibility to dynamically adapt both representation spaces on the RSITR task. UniAdapter and Cross-Modal Adapter are multimodal adapters that support multimodal interaction. However, their structures are complex

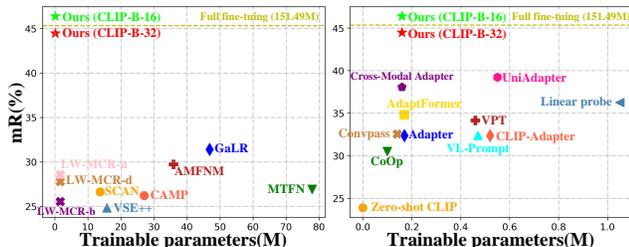


Fig. 8. Retrieval performance vs. the number of trainable parameters on the RSITMD dataset. *Left*: Comparison with traditional retrieval methods. *Right*: Comparison with CLIP-based methods.

and fail to consider the problem of high intra-class and inter-class similarity in RS data. The Cross-Modal Adapter [30] for multimodal tasks converges faster than UniAdapter [31], but the performance of UniAdapter is 1-2% higher.

The retrieval performance of adapter tuning is 5.9%, 2.4%, and 4.7% higher than that of prompt learning on the three datasets. Prompt learning is an input-related method and adapter tuning is a network-related method. Due to the significant domain gap between the pre-training CLIP and the RS domain, achieving efficient parameter transfer through adding learnable prompts to the input is limited. To adapt to the different knowledge structures of RS image-text and natural image-text, the adapter can adjust the network structure properly to align the RS image and text modality to obtain significant performance gain.

**Trained parameter efficiency.** As shown in Fig. 8(a), our method requires the smallest trainable parameters compared to the traditional methods. Even the lightweight model LW-MCR [9] requires 1.65 M parameters, our method requires only 0.16 M to achieve SOTA performance. This suggests that finetuning VL pre-training models in the natural domain have great potential for tasks in the RS domain. For the RS image-text retrieval task, our proposed PETL framework is both parameter-efficient and effective. Among the various PETL methods in Fig.8(b), the trainable parameters of our method are not the smallest, but it achieves the optimal performance with 98.9% reduction of the full fine-tuning parameters. Our method (CLIP-B-32) achieves nearly the performance of full fine-tuning on the RSITMD dataset and CLIP-B-16 exceeds it. This shows that it is essential to design the adapter and the PETL framework based on the high intra-modal similarity in RS data, which can achieve better performance.

### E. Ablation Study

We conduct detailed ablation experiments to validate the effectiveness of the PE-RSITR framework. We systematically analyze the influence of the CLIP backbone network, the proposed MRS-Adapter, and the HMMC loss function on the experiment results. In the following subsections, we conduct experiments mainly on the RSITMD dataset. The performance of T2I retrieval is measured by the sum of  $R@K$  ( $K=1, 5, \text{ and } 10$ ) of text retrieval and I2T retrieval is measured by the sum of  $R@K$  ( $K=1, 5, \text{ and } 10$ ) of image retrieval.

**Vision backbones.** Fig. 9 shows the results on the three datasets for the two ViT backbones of CLIP, *i.e.*, ViT-B/32

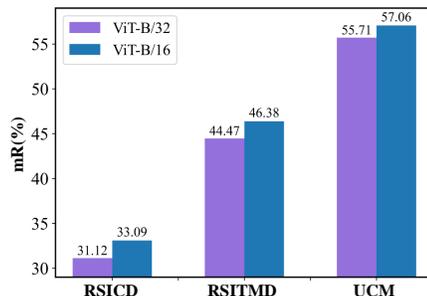


Fig. 9. Investigations with ViT-B/32 and ViT-B/16 backbones of CLIP.

Method	Visual	Textual	T2I retrieval	I2T retrieval	mR
w/o Share		✓	107.58	120.25	37.97
w/o Share	✓		114.51	130.32	40.81
w/o Share	✓	✓	116.47	136.16	42.11
Ours	✓	✓	<b>128.10</b>	<b>138.70</b>	<b>44.47</b>

and ViT-B/16. The results are expected: the more advanced the backbone, the better the performance.

**MRS-Adapter.** Since MRS-Adapter involves dual-modal branches and cross-modal interaction, it is necessary to verify that multimodal branches fine-tuned together are more suitable for the task. We applied it to the visual or textual branch alone or to both branches without the sharing mechanism as comparison experiments. As shown in Table IV, the retrieval performance of applying MRS-Adapter to a single branch is particularly poor, because it cannot adapt the space of the RS domain on both modalities at the same time. This result proves that MRS-Adapter should be used in both visual and text branches. When MRS-Adapter is applied to both branches, the absence of the sharing mechanism can degrade performance. The weight-sharing benefits the realignment of the VL feature space of CLIP in the RSITR task and can reduce the trained parameters. Specifically, MRS-Adapter reduces the parameters of 0.1M. In summary, MRS-Adapter helps to release the power of pre-trained CLIP models and improve RSITR performance.

**Bottleneck dimension  $d$ .** The bottleneck dimension  $d$  is an important parameter of MRS-Adapter. We conduct experiments with different sizes of bottleneck, as shown in Fig. 10(left). As the  $d$  increases, the parameters of MRS-Adapter increase. The retrieval performance is poor when the  $d$  is less than 64. It is possible that down-projection loses too much information by projecting features to the lower dimensional space. The best performance is achieved when the  $d$  is 64, and the performance starts to decrease as the dimension increases further. Therefore the  $d$  of MRS-Adapter is set to 64.

**Weight-sharing dimension  $r$ .** After the bottleneck dimension  $d$  is set to 64, we further explore the impact of the weight-sharing dimension  $r$ . We conduct experiments with different sizes of shared weights, as shown in Fig. 10(right). The overall effect of the weight-sharing dimension on the results is small. The poor retrieval performance when the  $r$  is small may be due to less knowledge sharing between RS image modality and text modality, which cannot fully facilitate modal alignment. The best performance is achieved when the  $r$  is 64, and

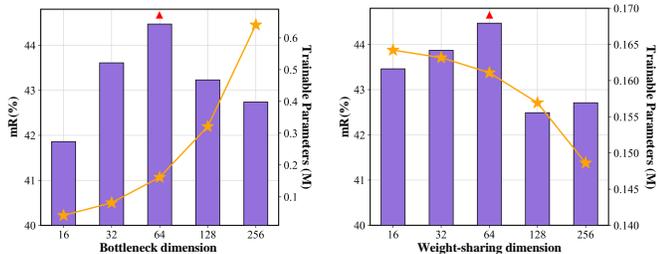


Fig. 10. *Left*: Ablations of the bottleneck dimension. *Right*: Ablations of the weight-sharing dimension.

TABLE V  
EXPERIMENT OF THE HMMC LOSS WITH DIFFERENT MARGINS  $\lambda$  AND  $\alpha$ .

$\lambda$	$\alpha=\alpha_v=\alpha_t$	T2I retrieval	I2T retrieval	mR
0.2	0.1	119.91	137.94	42.97
0.2	0.2	<b>128.10</b>	<b>138.70</b>	<b>44.47</b>
0.2	0.3	121.55	136.57	43.02
0.2	0.4	121.08	134.48	42.51
0.1	0.2	119.68	135.73	42.57
0.2	0.2	<b>128.10</b>	<b>138.70</b>	<b>44.47</b>
0.3	0.2	119.86	136.92	42.80
0.4	0.2	118.72	136.28	42.50

the performance starts to decrease as the dimension increases further. Therefore the  $r$  of MRS-Adapter is set to 64.

**Margins  $\lambda$  and  $\alpha$ .** In our HMMC learning objective, the margin parameters serve as crucial hyper-parameters. To observe the influence of parameters  $\lambda$ ,  $\alpha_v$ , and  $\alpha_t$  on retrieval results, we set up a series of experiments using a controlled variable approach. Regarding the intra-modal constraint, we set  $\alpha_v$  equal to  $\alpha_t$ . As shown in Table V, the results indicate that the model achieves the best retrieval performance, as measured by mR, when the margin parameters  $\lambda$  and  $\alpha$  are set to 0.2.

**HMMC loss.** We have carried out experiments on the hybrid multi-modal contrastive loss, *i.e.*, the inter-modal and intra-modal cooperative constraint framework. We set up four sets of comparisons to observe the effect of increasing the intra-modal constraints on the results, as shown in Table VI. The result shows that the retrieval performance is significantly improved after adding the intra-modal constraints. However, when the intra-modal constraint is added to the text branch only, text-to-image retrieval becomes worse, while image retrieval has the best result. Likewise, when the intra-modal constraint is added to the visual modality only, image-to-text retrieval becomes worse, while text retrieval has the best result. The RSITR task requires both text retrieval and image retrieval. Only by adding two intra-modal constraints can achieve a balanced result, and the overall retrieval performance is optimal. This result proves that the proposed HMMC loss is effective in further improving the performance on top of the proposed MRS-Adapter.

TABLE VI  
ABLATIONS OF THE HYBRID MULTI-MODAL CONTRASTIVE LOSS.

$\mathcal{L}_{intra-v}$	$\mathcal{L}_{intra-t}$	T2I retrieval	I2T retrieval	mR
		113.29	132.14	40.91
	✓	121.76	<b>141.02</b>	43.80
✓		<b>128.47</b>	134.81	43.88
✓	✓	128.10	138.70	<b>44.47</b>

## F. Qualitative Results

To get an intuition of how the RS image and text are aligned in the joint embedding space, we present a detailed visualization in Fig. 11. Fig. 11 shows the detailed change of RS image and text embeddings before and after PETL on three datasets. We utilize the t-distributed stochastic neighbor embedding (t-SNE) method to project the image and text features obtained from the two encoders into a 2-D space. Observing the first column of Fig. 11, the CLIP model of the natural domain before transfer learning exhibits a complete separation of RS image and text modalities due to the domain gap. As shown in the visualization, RSICD has the highest number of sample points. The RSITMD sample points are most scattered and uniformly distributed, indicating the highest fine-grained. The UCM has the lowest number of samples, while different clustering centers indicate different RS scene classes. The unaligned modal representations have semantic inconsistency, thus a large number of errors are sure to occur when Zero-shot CLIP computes the cross-modal similarity.

To tackle this problem, our MRS-Adapter aggregates the embeddings of each modality into a common space, as shown in the second column of Fig. 11. In addition, it attempts to align image-text pairs separately according to remotely sensed scene-level information to form finer and more discriminative clusters. However, we observe some overlap in each modal representation, which is caused by the great intra-class and inter-class similarity. To couple up this problem, we perform the HMMC loss function on top of the proposed MRS-Adapter, as shown in the third column of Fig. 11. The comparison reveals that our method further spars the distribution of multimodal embeddings. The sample points that originally overlapped in a large number of clusters are now clearly visible. In particular, the results of UCM originally had a lot of text clusters without RS image samples next to them, but now there are RS samples near each cluster of text samples. The samples of different modalities are more accurately aligned, alleviating the problem of sample overlap and providing a good representation for cross-modal retrieval and matching.

To qualitatively validate the effectiveness of our PE-RSITR method, we displayed several examples of T2I and I2T retrieval in Fig. 12. Based on these cross-modal retrieval results, our model can accurately distinguish similar samples to retrieve the correct results. Based on these cross-modal retrieval results, our model can accurately distinguish similar samples to retrieve the correct results. The method can understand both abstract phrases and complex long sentences and is robust in the face of both simple and complex images. This is mainly attributed to the fact that our designed PE-RSITR framework not only learns the specific knowledge of RS domain, but also exploits the powerful generalization ability and rich VL knowledge structure of CLIP.

## V. CONCLUSION AND FUTURE WORK

In this paper, we explore a new paradigm for the PETL-based RSITR task, namely PE-RSITR, to bridge the differences between the RS domain and the natural domain. Our proposed MRS-Adapter and the HMMC loss function are

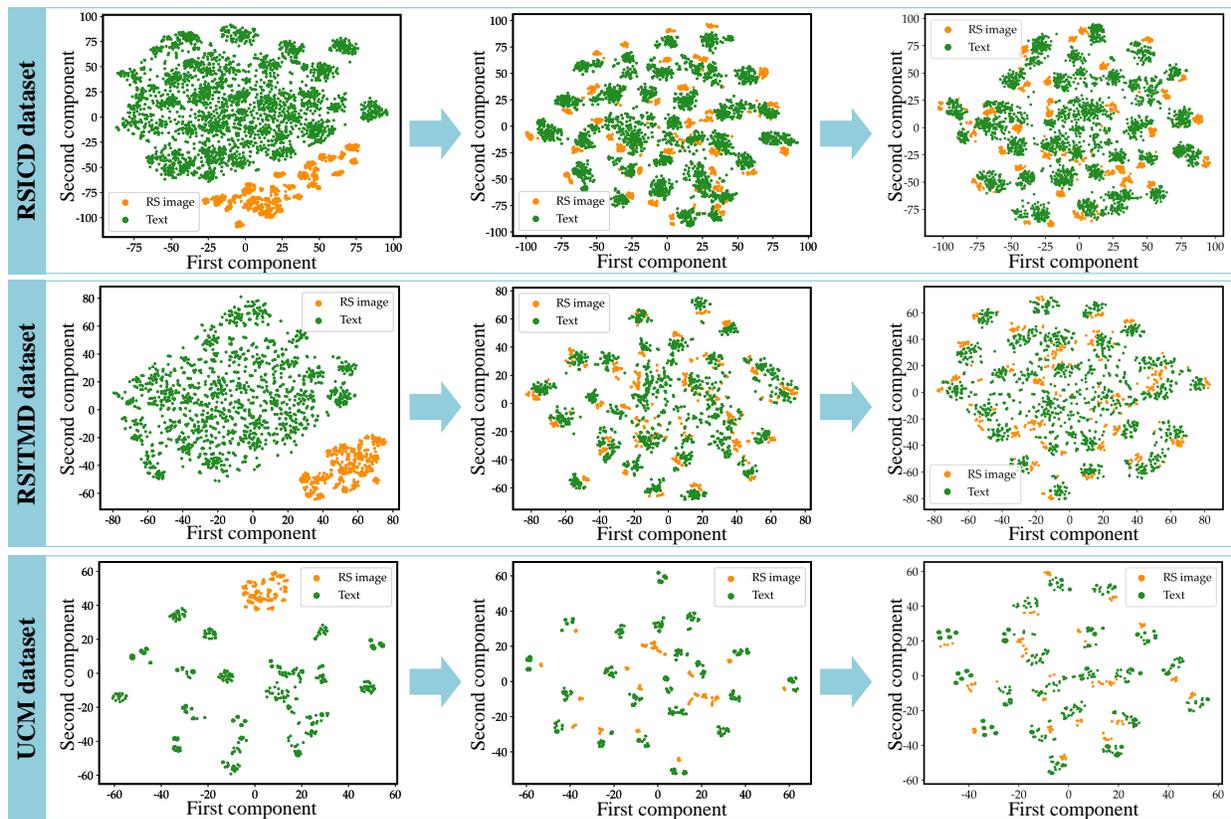


Fig. 11. The detailed change of t-SNE visualizations of RS image and text embeddings before and after transfer learning on three datasets. The first column shows the results of original CLIP before transfer learning, the second shows the results after transfer by MRS-Adapter only, and the third shows the results after adding the framework of inter-modal and intra-modal cooperative constraints.

conceptually simple but can effectively expand CLIP to coarse-grained VL tasks in the RS domain. Inserting a lightweight MRS-Adapter into CLIP without modifying the inherent structure consumes only a little computational cost. It can make the best use of the powerful generalization ability and rich visual-language prior knowledge of the CLIP. Constructing intra-modal positive pairs using random dropout and adding intra-modal constraints can effectively avoid the retrieval interference of similar samples. Extensive experiments on RSITR validate that our method can achieve comparable or even better performance than full fine-tuning.

Manual selection of various dimension parameters is required for the MRS-Adapter. How to adaptively select the parameters of the MRS-Adapter remains a challenging topic. Our proposed method does not account for scale variations, cluttered backgrounds, and sparse or dense distribution of objects in RS imagery. In future work, we will extend our method to other fine-grained perception tasks, such as RS visual grounding (object level) and RS image change captioning (spatiotemporal level). In these tasks, it is essential to extract fine-grained features or similarities from image-text pairs. We hope our work will inspire future research to explore more efficient PETL methods for more fine-grained RS visual-language tasks.

## REFERENCES

- [1] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu, "EarthNets: Empowering AI in Earth Observation," *arXiv:2210.04936*, 2022.
- [2] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and lidar data classification based on structural optimization transmission," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3153–3164, 2023.
- [3] Y. Zhang, M. Zhang, W. Li, S. Wang, and R. Tao, "Language-aware domain generalization network for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.
- [4] L. Dong, X. Lu, G. Liu, and Y. Yuan, "A novel nmf guided for hyperspectral unmixing from incomplete and noisy data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
- [5] X. Zhang, Y. Yuan, and X. Li, "Reweighted low-rank and joint-sparse unmixing with library pruning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
- [6] Y. Yuan, L. Dong, and X. Li, "Hyperspectral unmixing using nonlocal similarity-regularized low-rank tensor factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
- [7] X. Zhang, Y. Yuan, and X. Li, "Sparse unmixing based on adaptive loss minimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
- [8] Z. Yuan, W. Zhang, C. Tian, X. Rong, Z. Zhang, H. Wang, K. Fu, and X. Sun, "Remote sensing cross-modal text-image retrieval based on global and local information," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
- [9] Z. Yuan, W. Zhang, X. Rong, X. Li, J. Chen, H. Wang, K. Fu, and X. Sun, "A lightweight multi-scale crossmodal text-image retrieval method in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.
- [10] Z. Yuan, W. Zhang, K. Fu, X. Li, C. Deng, H. Wang, and X. Sun, "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.
- [11] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, 2018.
- [12] Z. Yuan, L. Mou, Q. Wang, and X. X. Zhu, "From easy to hard: Learning language-guided curriculum for visual question answering on remote

Query	Top-5 Retrieval Result	Query	Top-5 Retrieval Result
Two large ships loaded with cargo were moored on both sides of the gray port.			Several buildings and bare land are around an oval pond. Several buildings and bare land surround an oval pond. Several buildings and barley are to the REDOR of an oval lake. Some buildings and bars Landes are around the elliptical pond. A pond between green and yellow wasteland
There are many white planes parked beside the blue house.			The park is built on bare ground with two ponds . A circular square park is surrounded by the road. The rectangular park with circular square is surrounded by streets. The park with a blue pool surrounding the circular square is in the desert. A rectangular park with a circular square is surrounded by roads.
The red rectangular central building is located between the railway and the highway with cars.			The resort consists of pools, plants and homes. The resort has buildings, lakes and swimming pools, close to the beach. The resort with buildings, lakes and swimming pools is close to the beach. There is a resort near the beach with buildings, ponds and swimming pools. The resort has a swimming pool and a swimming pool.
it is a large stadium with numerous bleachers and a soccer field where players are playing on half of it.			There is a dark green vegetation near the school building. The school is on the forest and lawn. We can see the playground. The school is located in the forest and on the lawn you can see some sports fields. The school is located in the forest. We can see some sports grounds on the lawn. Near the schoolhouse is a dark green plant.
Many roads around are connected with many green vegetation.			In some compact houses, a road runs vertically through a dry river. This residential area where roofs are mostly dark is separated by roads. This area is a luxurious block. Two roads stretches through this dense residential area. A medium residential area with a narrow road goes through this area.

Fig. 12. Qualitative top-5 text-to-image and image-to-text retrieval results of our PE-RSITR method on the RSITMD dataset. Images with orange bounding boxes are the ground-truth and the ground-truth texts are marked with green.

- sensing data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
- [13] Y. Zhan, Z. Xiong, and Y. Yuan, “Rsvg: Exploring data and models for visual grounding on remote sensing data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023.
- [14] Y. Sun, S. Feng, X. Li, Y. Ye, J. Kang, and X. Huang, “Visual grounding in remote sensing images,” in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 404–412.
- [15] Z. Yuan, W. Zhang, C. Li, Z. Pan, Y. Mao, J. Chen, S. Li, H. Wang, and X. Sun, “Learning to evaluate performance of multimodal semantic localization,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.
- [16] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, “Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, 2022.
- [17] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, “Change detection meets visual question answering,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
- [18] H. Ning, B. Zhao, and Y. Yuan, “Semantics-consistent representation learning for remote sensing image–voice retrieval,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
- [19] S. Lobry, D. Marcos, J. Murray, and D. Tuia, “Rsvqa: Visual question answering for remote sensing data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8555–8566, 2020.
- [20] M. Zhang, X. Zhao, W. Li, Y. Zhang, R. Tao, and Q. Du, “Cross-scene joint classification of multisource data with multilevel domain adaption network,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, 2023.
- [21] M. Liang, R. W. Liu, Y. Zhan, H. Li, F. Zhu, and F.-Y. Wang, “Fine-grained vessel traffic flow prediction with a spatio-temporal multigraph convolutional network,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23 694–23 707, 2022.
- [22] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, “Single-source domain expansion network for cross-scene hyperspectral image classification,” *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, 2023.
- [23] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, “Topological structure and semantic information transfer network for cross-scene hyperspectral image classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 2817–2830, 2023.
- [24] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu, “Vlp: A survey on vision-language pre-training,” *Mach. Intell. Res.*, vol. 20, no. 1, pp. 38–56, 2023.
- [25] Y. Du, Z. Liu, J. Li, and W. X. Zhao, “A survey of vision-language pre-trained models,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 5436–5443.
- [26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.
- [27] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, “Parameter-efficient transfer learning for NLP,” in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 2790–2799.
- [28] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4582–4597.
- [29] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 3045–3059.
- [30] H. Jiang, J. Zhang, R. Huang, C. Ge, Z. Ni, J. Lu, J. Zhou, S. Song, and G. Huang, “Cross-modal adapter for text-video retrieval,” *arxiv.2211.09623*, 2022.
- [31] H. Lu, M. Ding, Y. Huo, G. Yang, Z. Lu, M. Tomizuka, and W. Zhan, “Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling,” *arxiv.2302.06605*, 2023.
- [32] B. Qu, X. Li, D. Tao, and X. Lu, “Deep semantic understanding of high resolution remote sensing image,” in *Proc. Int. Conf. Comput., Inf. Telecommunication Syst.*, 2016, pp. 1–5.
- [33] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. Zuair, “Textrs: Deep bidirectional triplet network for matching text to remote sensing images,” *Remote Sens.*, vol. 12, no. 3, p. 405, 2020.
- [34] G. Hoxha, F. Melgani, and B. Demir, “Toward remote sensing image retrieval under a deep image captioning perspective,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4462–4475, 2020.
- [35] M. M. A. Rahhal, Y. Bazi, T. Abdullah, M. L. Mekhalfi, and M. Zuair, “Deep unsupervised embedding for remote sensing image retrieval using textual cues,” *Appl. Sci.*, vol. 10, no. 24, p. 8931, 2020.
- [36] Q. Cheng, Y. Zhou, P. Fu, Y. Xu, and L. Zhang, “A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4284–4297, 2021.
- [37] M. M. A. Rahhal, Y. Bazi, N. A. Alsharif, L. Bashmal, N. Alajlan, and F. Melgani, “Multilanguage transformer for improved text to remote sensing image retrieval,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9115–9126, 2022.
- [38] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, “Oscar: Object-semantics aligned pre-training for vision-language tasks,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 121–137.
- [39] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, and M. Zhou, “Xgpt: Cross-modal generative pre-training for image captioning,” in *Proc. 10th CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, 2021, pp. 786–797.
- [40] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, “SimVLM: Simple visual language model pretraining with weak supervision,” in *Proc. Int. Conf. Learn. Represent.*, 2022.

- [41] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13–23.
- [42] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 9694–9705, 2021.
- [43] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 4904–4916.
- [44] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–1, 2022.
- [45] T. Zhang, P. Gao, H. Dong, Y. Zhuang, G. Wang, W. Zhang, and H. Chen, "Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain," *Remote Sens.*, vol. 14, no. 22, p. 5675, 2022.
- [46] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 709–727.
- [47] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [48] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16 795–16 804.
- [49] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Unified vision and language prompt learning," *arXiv preprint arXiv:2210.07225*, 2022.
- [50] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19 113–19 122.
- [51] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T.-S. Chua, and M. Sun, "Cpt: Colorful prompt tuning for pre-trained vision-language models," *arXiv preprint arXiv:2109.11797*, 2021.
- [52] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "AdapterFusion: Non-destructive task composition for transfer learning," in *Proc. 16th Conf. Eur. Chapter Assoc. Computat. Linguistics*, 2021, pp. 487–503.
- [53] R. Karimi Mahabadi, J. Henderson, and S. Ruder, "Compacter: Efficient low-rank hypercomplex adapter layers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 1022–1035, 2021.
- [54] R. Karimi Mahabadi, S. Ruder, M. Dehghani, and J. Henderson, "Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 565–576.
- [55] A. Rücklé, G. Geigle, M. Glockner, T. Beck, J. Pfeiffer, N. Reimers, and I. Gurevych, "AdapterDrop: On the efficiency of adapters in transformers," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 7930–7946.
- [56] H. Chen, R. Tao, H. Zhang, Y. Wang, W. Ye, J. Wang, G. Hu, and M. Savvides, "Conv-adapter: Exploring parameter efficient transfer learning for convnets," *arXiv preprint arXiv:2208.07463*, 2022.
- [57] S. Jie and Z.-H. Deng, "Convolutional bypasses are better vision transformer adapters," *arXiv preprint arXiv:2207.07039*, 2022.
- [58] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adapterformer: Adapting vision transformers for scalable visual recognition," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 16 664–16 678, 2022.
- [59] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li, "ST-adapter: Parameter-efficient image-to-video transfer learning," *Adv. Neural Inf. Process. Syst.*, 2022.
- [60] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *arXiv:2110.04544*, 2021.
- [61] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 493–510.
- [62] O. Pantazis, G. Brostow, K. E. Jones, and O. Mac Aodha, "Svl-adapter: Self-supervised adapter for vision-language pretrained models," in *Proc. 33rd Brit. Mach. Vis. Conf.*, 2022.
- [63] Y.-L. Sung, J. Cho, and M. Bansal, "VI-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5227–5237.
- [64] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 935–943.
- [65] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 6894–6910.
- [66] Z. Yuan, W. Zhang, C. Tian, Y. Mao, R. Zhou, H. Wang, K. Fu, and X. Sun, "Mcrn: A multi-source cross-modal retrieval network for remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 115, p. 103071, 2022.
- [67] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.
- [68] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5764–5773.
- [69] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, "Matching images and text with multi-modal tensor fusion and re-ranking," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 12–20.
- [70] F. Zheng, W. Li, X. Wang, L. Wang, X. Zhang, and H. Zhang, "A cross-attention mechanism based on regional-level semantic features of images for cross-modal text-image retrieval in remote sensing," *Appl. Sci.*, vol. 12, no. 23, 2022.