Parameter-Efficient Remote Sensing Image-Text Retrieval via Hierarchical Gated Multi-Modal Adapters

Xin Pan[1] ,Xiaohui Huang[1] ,Wenhai Li[1] ,and Xuebo Cheng[1]

[1] School of Information and Software Engineering, East China Jiaotong University, Nanchang 330013, Jiangxi, China
hxh016@hotmail.com

Abstract. Remote sensing image-text retrieval faces significant challenges due to the limitations of traditional models and the prohibitive computational costs of full-parameter fine-tuning. To address these issues, this paper proposes a parameter-efficient fine-tuning framework driven by a Hierarchical Gated Multi-modal Adapter. Specifically, we integrate a multi-head self-attention module with a hierarchical gating mechanism to dynamically regulate cross-modal features across varying network depths. Furthermore, to enhance the model's robustness against noise interference and domain shifts, we introduce a novel Cross-modal Semantic Perturbation data augmentation strategy that generates semantically consistent pseudo-samples. Extensive experiments on the RSICD and RSITMD benchmark datasets demonstrate that our proposed HGMA-RSITR framework achieves performance comparable to full-parameter fine-tuning while updating merely 0.6M parameters, offering a highly efficient and robust solution for cross-modal retrieval.

Keywords: Remote Sensing Image-Text Retrieval, Parameter-Efficient Fine-Tuning, Hierarchical Gated Multi-modal Adapter, Data Augmentation.

1 Introduction

The rapid development of aerospace and remote sensing technologies has driven an explosive growth in remote sensing image data[1]. These images contain rich geographical and environmental information, holding significant application value in various fields such as agriculture, forestry, and oceanography. However, efficiently extracting useful knowledge from massive amounts of remote sensing data remains a highly challenging task. To address this challenge, remote sensing image-text retrieval, serving as a key cross-modal technology, provides an effective means for efficient information extraction and knowledge discovery from large-scale remote sensing data by bridging the semantic gap between visual and linguistic modalities.

With the rapid advancement of deep learning, cross-modal remote sensing image-text retrieval has achieved remarkable progress in feature representation learning and semantic alignment. This paradigm primarily extracts high-level semantic features from images and texts separately, map them into a unified semantic embedding space, and ultimately accomplish the retrieval task through feature similarity measurements.

Nevertheless, due to the inherent heterogeneous modality gap between images and texts, achieving accurate and efficient cross-modal feature semantic alignment has become the core key to improving retrieval performance. Based on the feature extraction architectures, existing mainstream methods can be broadly classified into two technical routes. One category, represented by GaLR [2], and SWAN [3], utilizes deep Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs) for text feature encoding. The other category, represented by Transformer-based architectures such as IEFT [4] and PIR [5], relies on the global modeling capabilities of self-attention mechanisms to accomplish feature extraction for both images and texts. The aforementioned methods typically rely on triplet loss and contrastive loss [6] as their optimization core, constraining the model by minimizing the feature distance between positive sample pairs while maximizing the distance between negative sample pairs.

In recent years, the Contrastive Language-Image Pre-training (CLIP) [7] model has demonstrated exceptional zero-shot transfer capabilities in natural image-text understanding tasks. Benefiting from pre-training on massive datasets, CLIP can effectively capture the semantic alignment between images and texts, thereby learning generalized vision-language representations. Its retrieval performance significantly outperforms traditional methods. However, directly applying CLIP to remote sensing image-text retrieval still faces severe challenges. Remote sensing images typically possess high intra-class similarity and complex spatial structures, with scenes often containing dense objects and cluttered backgrounds. These domain specificities make it difficult for the directly transferred CLIP model to fully leverage its cross-modal alignment advantages, thereby limiting further improvements in retrieval performance. To efficiently adapt the general visual-linguistic knowledge of VLP models to the remote sensing domain, existing efforts generally fall into three paradigms:

The first category of methods conducts full fine-tuning of pre-trained models by constructing large-scale remote sensing image-text datasets. For example, Zhang et al. [8] proposed RS5M, a dataset containing over 5 million sample pairs, while Wang et al. [9] constructed the Skyscrip dataset and the corresponding SkyCLIP model. Although such methods can significantly enhance performance, the massive scale of model parameters leads to high training and deployment costs, thus limiting their scalability.
The second category treats CLIP as a frozen feature encoder and strengthens semantic correspondences by introducing cross-modal alignment modules and enhanced loss functions. For instance, GLISA [10] and other approaches have achieved promising results by initializing encoders with pre-trained model parameters and subsequently introducing soft-alignment strategies. PIR-CLIP [11] initializes its encoder with pre-trained parameters based on the PIR [5] architecture, leading to a substantial improvement in model performance. While this category of methods improves retrieval performance and maintains the original semantic representation stability of CLIP, it still requires updating a large number of parameters, resulting in considerable computational overhead.

The third category focuses on Parameter-Efficient Fine-Tuning (PEFT) techniques, aiming to achieve the efficient transfer of pre-trained models to the remote sensing image-text retrieval domain by optimizing only a small fraction of trainable parameters.

For example, Yuan et al. [12] designed a CLIP-specific adapter for the remote sensing retrieval task, which effectively reduces training costs while improving retrieval performance. However, existing adapters typically employ a monolithic structure across all Transformer layers. This uniform design treats low-level spatial textures and high-level abstract semantics equally, constraining the model's hierarchical reasoning capacity and frequently leading to semantic confusion when matching dense remote sensing scenes with complex textual descriptions.

To overcome the limitations of the aforementioned methods, this paper proposes a parameter-efficient fine-tuning approach based on a Hierarchical Gated Multimodal Adapter. This method introduces a multi-head self-attention module into the adapter structure to enhance the adaptability of the CLIP model to the feature distribution inherent in the remote sensing domain. Simultaneously, a hierarchical gating mechanism is designed to dynamically adjust the contribution weights of the adapters across Transformer layers of different depths. While preserving the original image-text encoding capability of CLIP, this design effectively mitigates semantic confusion, thereby further improving the semantic alignment performance between remote sensing images and texts. Furthermore, to address the issue of insufficient model robustness caused by the limited scale and imbalanced quality of existing remote sensing datasets, this paper proposes a Cross-Modal Semantic Perturbation data augmentation strategy. By implementing aligned combination and perturbation operations within the dataset, this strategy generates pseudo-samples with semantic consistency, effectively enhancing the model's adaptability to noise interference and domain shifts.

The main contributions of this paper can be summarized as follows:

- We propose a parameter-efficient fine-tuning framework oriented towards remote sensing image-text retrieval. By combining CLIP with a hierarchical gated multi-modal adapter, it achieves efficient cross-modal semantic alignment while updating only a minimal number of parameters.
- We design a cross-modal semantic perturbation data augmentation strategy, which effectively enriches sample diversity and alleviates training challenges brought by data scarcity and high intra-class similarity.
- Extensive experiments on multiple mainstream benchmark datasets demonstrate that our method requires training only about 0.6M parameters to achieve performance comparable to, or even better than, full-parameter fine-tuning.

2 Methodology

This paper proposes a Remote Sensing Image-Text Retrieval model based on a Hierarchical Gated Multimodal Adapter (HGMA-RSITR), and its overall architecture is illustrated in Figure 1. Built upon the pre-trained CLIP as the foundational backbone, the model primarily consists of three core modules: a dual-stream feature encoder with parallel HGMA adapters, a cross-modal semantic alignment module, and a Cross-modal Semantic Perturbation (CMSP) data augmentation module. By efficiently fine-tuning a small number of HGMA parameters within the frozen backbone network,

complemented by the data augmentation strategy, the model achieves accurate matching of remote sensing images and texts while significantly reducing training costs.

2.1 Overall Model Architecture

The proposed model is constructed based on a dual-stream architecture. It extracts high-level semantic features through independent image and text encoders, maps them into a shared image-text joint embedding space, and ultimately achieves cross-modal retrieval by measuring the distance between the features.

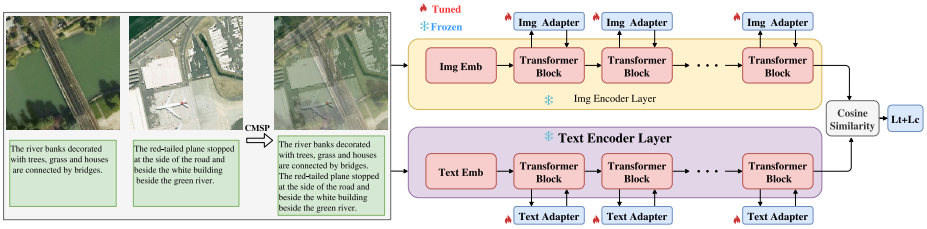

Fig. 1. Overall architecture of the Hierarchical Gated Multimodal Adapter Remote Sensing Image-Text Retrieval (HGMA-RSITR) model.

Image Encoder: The image encoder is based on the ViT-B/32 of CLIP. Given an input remote sensing image $I \in \mathbb{R}^{H\times W\times C}$, it is first divided into $N = HW / P^2$ non-overlapping image patches of size $P\times P$ (where C is the number of channels). These image patches are then converted into a sequence of D_v-dimensional feature vectors $I_{patch} \in \mathbb{R}^{N\times D_v}$ through a linear projection. Subsequently, a learnable global class token $I_{cls} \in \mathbb{R}^{1\times D_v}$ is prepended to the sequence, and a learnable position embedding $I_{pos} \in \mathbb{R}^{(N+1)\times D_v}$ of the same dimension is added to form the initial input I_0 for the image encoder:

$$I_0 = [I_{cls}; I_{patch}] + I_{pos} \tag{1}$$

The image encoder consists of an L-layer Transformer architecture, where each layer contains a Multi-Head self-Attention (MHA) module and a Multi-Layer Perceptron (MLP). The overall processing of the l_{th} layer can be expressed as:

$$\hat{I}_l = \mathrm{MHA}(\mathrm{LN}(I_{l-1})) + I_{l-1}, \quad l = 1,\ldots,L, \tag{2}$$

$$I_l = \mathrm{MLP}(\mathrm{LN}(\hat{I}_l)) + \hat{I}_l, \quad l = 1,\ldots,L, \tag{3}$$

Finally, the output of the encoder is mapped to a D-dimensional image-text joint embedding space through a linear projection layer, yielding the final image feature sequence $I_f = \{v, I_1,\ldots,I_N\} \in \mathbb{R}^{(N+1)\times D}$, where v represents the global image feature mapped from I_{cls}.

Text Encoder: Given a descriptive text $T = \{w_m\}_{m=1}^{M}$, it is first mapped into a feature sequence $T_{token} \in \mathbb{R}^{M \times D_t}$ through tokenization and a word embedding layer. To identify the start and end of the sequence, special tokens $T_{sos} \in \mathbb{R}^{1 \times D_t}$ and $T_{eos} \in \mathbb{R}^{1 \times D_t}$ are concatenated to the beginning and end of the sequence, respectively, and a text position embedding $T_{pos} \in \mathbb{R}^{(M+2) \times D_t}$ is added to obtain the initial input T_0 for the text encoder:

$$T_0 = [T_{sos}; T_{token}; T_{eos}] + T_{pos}, \tag{4}$$

Similar to the image branch, after T_0 is processed by the text encoder and a linear projection layer, the output is mapped to the same joint embedding space to form the text feature sequence $T_f = \{T_{sos}, T_1, \ldots, T_M, t\} \in \mathbb{R}^{(M+2) \times D}$, where t represents the global text feature mapped from the end token T_{eos}.

Loss Function Design: To achieve effective alignment of cross-modal features, this paper employs cosine similarity to measure the global feature similarity between images and texts:

$$\mathrm{Sim}(v, t) = \frac{v \cdot t}{\| v \| \| t \|}, \tag{5}$$

where v and t are the aforementioned extracted and L2-normalized global features. Building upon this metric, we formulate a contrastive loss function utilizing the similarity distribution:

$$\mathcal{L}_{con} = -\frac{1}{B} \sum_{i=1}^{B} \left[\log \frac{\exp(\mathrm{Sim}(v_i, t_i) / \tau)}{\sum_{j=1}^{B} \exp(\mathrm{Sim}(v_i, t_j) / \tau)} + \log \frac{\exp(\mathrm{Sim}(t_i, v_i) / \tau)}{\sum_{j=1}^{B} \exp(\mathrm{Sim}(t_i, v_j) / \tau)} \right] \tag{6}$$

where B is the batch size, and τ is a learnable temperature scaling parameter. To further strengthen the model's ability to distinguish hard negative samples in complex remote sensing scenarios, a bidirectional Triplet Loss is introduced, which widens the similarity gap between positive and negative sample pairs through an explicit margin constraint:

$$\mathcal{L}_{triplet} = \frac{1}{B} \sum_{i=1}^{B} \begin{bmatrix} \max(0, m + \mathrm{Sim}(v_i, \overline{t_i}) - \mathrm{Sim}(v_i, t_i)) + \\ \max(0, m + \mathrm{Sim}(t_i, \overline{v_i}) - \mathrm{Sim}(t_i, v_i)) \end{bmatrix} \tag{7}$$

where m is the predefined margin parameter, and $\overline{t_i}$ and $\overline{v_i}$ denote the hardest negative samples corresponding to v_i and t_i within the batch, respectively. Synthesizing the above two terms, the final optimization objective of this paper is defined as a weighted sum:

$$\mathcal{L} = \mathcal{L}_{con} + \lambda \mathcal{L}_{triplet} \tag{8}$$

where λ is the weight coefficient balancing the relative contributions of the two losses.

$$T_0 = [T_{sos}; T_{token}; T_{eos}] + T_{pos}, \tag{9}$$

2.2 Hierarchical Gated Multimodal Adapter

To effectively enhance the model's cross-modal representation learning capability for remote sensing features with minimal parameter overhead, we propose a Hierarchical Gated Multimodal Adapter (HGMA). As illustrated in Fig. 2, the HGMA is embedded into the frozen image and text encoders in a parallel manner. To avoid semantic interference during the cross-modal alignment process, the HGMA adopts an independent dual-branch structure for images and texts. Each branch sequentially comprises a down-projection layer, an independent Multi-Head self-Attention (MHA) module, and an up-projection layer. Furthermore, a dual gating mechanism is introduced at both the internal feature fusion point and the output of the adapter to achieve adaptive dynamic regulation of cross-modal information at different hierarchical levels.

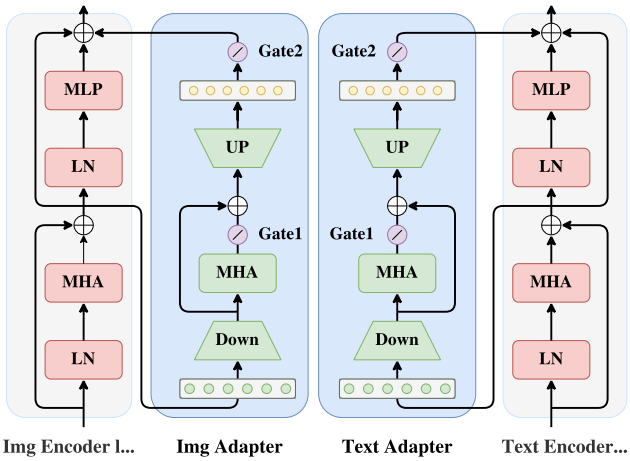

Fig. 2. Architecture of the HGMA Adapter module.

Formally, in the l-th Transformer encoder layer, let the intermediate features processed by the original MHA module (which serve as the common input for both the MLP module and the adapter) be denoted as $\hat{\mathbf{I}}_l \in \mathbb{R}^{N_v \times D_v}$ (image branch) and $\hat{\mathbf{T}}_l \in \mathbb{R}^{N_t \times D_t}$ (text branch). For the image branch, the internal computation and the output features of the HGMA can be formulated as:

$$\hat{\mathbf{I}}_l^{\downarrow} = \mathrm{GELU}(\hat{\mathbf{I}}_l \mathbf{W}_l^{\mathrm{down}}) \tag{10}$$

$$\hat{\mathbf{I}}_l^A = \left(g_I^1 \cdot \mathrm{MHA}(\hat{\mathbf{I}}_l^{\downarrow}) + (1 - g_I^1) \cdot \hat{\mathbf{I}}_l^{\downarrow} \right) \mathbf{W}_I^{\mathrm{up}} \tag{11}$$

Similarly, the output features of the HGMA in the text branch are expressed as:

$$\hat{\mathbf{T}}_l^{\downarrow} = \mathrm{GELU}(\hat{\mathbf{T}}_l \mathbf{W}_T^{\mathrm{down}}) \tag{12}$$

$$\hat{\mathbf{T}}_l^A = \left(g_T^1 \cdot \mathrm{MHA}(\hat{\mathbf{T}}_l^{\downarrow}) + (1 - g_T^1) \cdot \hat{\mathbf{T}}_l^{\downarrow} \right) \mathbf{W}_T^{\mathrm{up}} \tag{13}$$

where $\mathbf{W}_I^{\mathrm{down}} \in \mathbb{R}^{D_v \times r}$ and $\mathbf{W}_I^{\mathrm{up}} \in \mathbb{R}^{r \times D_v}$ represent the learnable down-projection and up-projection weight matrices in the image branch, respectively (r is the bottleneck dimension of the adapter, and $r \ll D_v$). $\mathbf{W}_T^{\mathrm{down}}$ and $\mathbf{W}_T^{\mathrm{up}}$ are the corresponding weight matrices for the text branch. g_I^1 and g_T^1 are learnable gating parameters utilized to adaptively balance the deep contextual representations extracted by the MHA module and the original features preserved by the down-projection within the adapter.

Finally, the HGMA is integrated with the original Transformer layer in a parallel fashion. The intermediate features obtained from the previous layer after passing through the original MHA module are:

$$\hat{\mathbf{I}}_l = \mathrm{MHA}(\mathrm{LN}(\mathbf{I}_{l-1})) + \mathbf{I}_{l-1} \tag{14}$$

$$\hat{\mathbf{T}}_l = \mathrm{MHA}(\mathrm{LN}(\mathbf{T}_{l-1})) + \mathbf{T}_{l-1} \tag{15}$$

Subsequently, the final output of the current layer is jointly determined by the output of the original MLP module, the residual connection, and the output of the adapter. Modulated by the second-stage gating mechanism, the updated features are represented as:

$$\mathbf{I}_l = \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{I}}_l)) + g_I^2 \cdot \hat{\mathbf{I}}_l + (1 - g_I^2) \cdot \hat{\mathbf{I}}_l^A \tag{16}$$

$$\mathbf{T}_l = \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{T}}_l)) + g_T^2 \cdot \hat{\mathbf{T}}_l + (1 - g_T^2) \cdot \hat{\mathbf{T}}_l^A \tag{17}$$

where $\hat{\mathbf{I}}_l^A$ and $\hat{\mathbf{T}}_l^A$ denote the outputs of the image and text adapters at the current layer, respectively. g_I^2 and g_T^2 are learnable gating parameters designed to adaptively adjust the contribution ratio between the domain-specific features from the adapter and the original pre-trained features of the model, thereby maximizing the effect of feature transfer.

2.3 Cross-modal Semantic Perturbation

To address the challenges of limited scale and imbalanced annotation quality in remote sensing image-text datasets, we propose a Cross-modal Semantic Perturbation (CMSP) data augmentation mechanism. Unlike natural images, remote sensing scenes

are typically composed of multiple geospatial entities (e.g., buildings, roads, water bodies, and vegetation), which endows them with inherent compositionality in the semantic space. Consequently, the linear mixing of different remote sensing images can be interpreted as a smooth transition of land-cover distributions rather than a semantic conflict.

Based on this observation, CMSP generates pseudo-paired samples by synchronously combining images and texts within the dataset. Given two randomly sampled image-text pairs (I_i, T_i) and (I_j, T_j) from the same batch, a new pseudo-sample $(\tilde{I}_{ij}, \tilde{T}_{ij})$ is constructed as follows:

$$\tilde{I}_{ij} = \lambda I_i + (1-\lambda) I_j \tag{18}$$

$$\tilde{T}_{ij} = [T_i; T_j] \tag{19}$$

where $[T_i; T_j]$ denotes the simple concatenation of the two text sequences, effectively constructing a joint scene representation containing multi-entity descriptions at the semantic level. During the training phase, each image-text pair triggers this combination operation with a probability $\alpha \in [0,1]$.

Notably, distinct from the conventional Mixup setting where λ follows a random Beta distribution, we strictly fix the image mixing ratio λ to 0.5 in CMSP. The theoretical justification is that the direct concatenation of text essentially performs a 1:1 lossless semantic combination of two scenes. To maintain strict alignment in the cross-modal mapping, the mixing ratio of visual features must remain highly consistent with the semantic combination ratio of the text. Introducing a random λ would result in a significant semantic mismatch between the mixed image (e.g., 10% of scene A and 90% of scene B) and the equally weighted joint text, thereby introducing modal noise.

Through this controlled cross-modal synchronous perturbation, CMSP effectively expands the support domain of the joint embedding space without introducing external data. It forces the model to maintain stable matching relationships even when faced with complex joint semantics, thereby significantly enhancing the model's robustness against domain shifts and annotation noise.

3 Experiments

This section details the experimental settings and compares the proposed model with nine baseline methods on two remote sensing image-text retrieval benchmark datasets: RSICD[13]and RSITMD[14]. The baselines include three state-of-the-art traditional deep learning methods, one CLIP zero-shot transfer method, one fully fine-tuned CLIP method, and five state-of-the-art parameter-efficient adapter fine-tuning methods. Furthermore, ablation studies are conducted to verify the effectiveness of each core module, and hyperparameter analysis is performed to determine the optimal mixing ratio in data augmentation.

3.1 Datasets and Experimental Settings

Datasets: The proposed model is evaluated on two datasets: RSICD and RSITMD. The RSICD dataset contains 10,921 images with 54,605 descriptions, covering 30 categories. The RSITMD dataset consists of 4,743 images with 23,715 descriptions across 32 categories. To ensure a fair comparison with the baseline methods, we adopt the widely used public standard dataset splits, partitioning the samples into training, validation, and testing sets with a ratio of 80%, 10%, and 10%, respectively.

Evaluation Metrics: We employ the standard Recall at K (R@K, K=1, 5, 10) and mean Recall (mR) for cross-modal retrieval. R@K measures the proportion of queries for which at least one correct matching item is found within the top K retrieved results. The mR is the arithmetic mean of all six R@K values across both image-to-text and text-to-image retrieval tasks, comprehensively reflecting the model's retrieval performance.

Experimental Settings: The visual encoder adopts ViT-B/32, and the text encoder utilizes its corresponding Transformer architecture. Both are initialized with the pre-trained CLIP model weights provided by Hugging Face, and their backbone parameters are kept frozen during the fine-tuning process. The temperature parameter τ for the contrastive loss is fixed at 0.07.

Implementation Details: The model is implemented using the PyTorch framework, and all experiments are conducted on a single NVIDIA RTX 4090 GPU. During training, the Adam optimizer is employed with an initial learning rate of 4×10^{-4} and a weight decay of 0.04. A linear learning rate scheduler is also applied. The training batch size is set to 280, and the model is trained for 80 epochs. The model is evaluated on the validation set, and the weights that achieve the highest mR are saved for the final testing phase.

3.2 Comparative Experimental Results and Analysis

Tables 1 and 2 compare different methods on the RSICD and RSITMD datasets. By introducing merely 0.6M trainable parameters, the proposed HGMA-RSITR achieves mean Recall (mR) scores of 34.44% and 48.47% in bidirectional cross-modal retrieval tasks, respectively.

Traditional deep learning methods, represented by GaLR [2] and PIR [5], yield mR scores that are mostly confined between 18% and 38%. In contrast, the introduction of large pre-trained models presents significant advantages. Even without prior training on remote sensing data, the zero-shot CLIP [7] reaches mR scores of 18.86% and 27.39%. The fully fine-tuned CLIP-CL further elevates the mR to 33.96% and 47.63%. However, full-parameter fine-tuning incurs high computational costs. By updating only a minimal number of parameters, our method achieves a retrieval accuracy marginally superior to that of CLIP-CL. This demonstrates that the adapter architecture can effectively unlock the remote sensing domain adaptability of large model, realizing a balance between parameter efficiency and retrieval performance.

Table 1. Quantitative experimental results of various methods on the RSICD datasets.

Method	Params	Text retrieval R@1/ R@5/R@10	Image retrieval R@1/R@5/R@10	mR
GaLR[2]	46M	6.59 / 19.85 /31.04	4.69/19.48 / 32.13	18.96
SWAN[3]	40M	7.41 / 20.13 /30.86	5.56/22.26 / 37.41	20.61
PIR[5]	161M	9.88 / 27.26 /39.16	6.97/24.56 / 38.92	24.46
Zero-shot CLIP[7]	0M	6.31 / 17.2 / 27.17	6.55/21.76 / 34.16	18.86
CLIP-CL[8]	151M	17.84 / 35.96 /50.14	13.89/35.15/50.08	33.96
CLIP-Adapter[15]	0.52M	7.11 / 19.48 / 31.01	7.67/24.87 / 39.73	21.65
AdapterFormer[16]	0.17M	12.46 / 28.49 /41.86	9.09/29.89 / 46.81	28.10
UniAdapter[17]	0.55M	12.65 / 30.81 /42.74	9.61/30.06 / 47.16	28.84
PE-RSITR[12]	0.16M	14.13 / 31.51 /44.77	11.63/33.92/50.73	31.12
HarMA[18]	0.50M	16.36 / 34.48 /47.74	12.92/37.17/53.07	33.62
Ours	0.60M	16.38 / 35.86 /51.24	12.57/37.11/53.50	34.44

Table 2. Quantitative experimental results of various methods on the RSITMD datasets.

Method	Params	Text retrieval R@1 / R@5 / R@10	Image retrieval R@1 / R@5 / R@10	mR
GaLR[2]	46M	14.82 / 31.64 / 42.48	11.15 / 36.68 / 51.68	31.41
SWAN[3]	40M	13.35 / 32.15 / 46.90	11.24 / 40.40 / 60.60	34.11
PIR[5]	161M	18.14 / 41.15 / 52.88	12.17 / 41.68 / 63.41	38.24
Zero-shot CLIP[7]	0M	10.18 / 24.12 / 35.84	10.00 / 33.63 / 50.58	27.39
CLIP-CL[8]	151M	24.78 / 50.00 / 63.27	22.61 / 55.27 / 69.87	47.63
CLIP-Adapter[15]	0.52M	12.83 / 28.84 / 39.05	13.30 / 40.20 / 60.06	32.38
AdapterFormer[16]	0.17M	16.71 / 30.16 / 42.91	14.27 / 41.53 / 61.46	34.81
UniAdapter[17]	0.55M	19.86 / 36.32 / 51.28	7.54 / 44.89 / 65.46	39.23
PE-RSITR[12]	0.16M	23.67 / 44.07 / 60.36	20.10 / 50.63 / 67.97	44.47
HarMA[18]	0.50M	25.81 / 48.37 / 60.01	19.92 / 53.27 / 71.21	46.53
Ours	0.60M	26.77 / 50.22 / 63.5	22.3 / 55.93 / 72.08	48.47

To further validate the efficacy of our architecture, we benchmark HGMA-RSITR against five mainstream parameter-efficient fine-tuning (PEFT) methods. As indicated in Table 1, model performance progressively improves with the sophistication of the adapter design. The early CLIP-Adapter [15], utilizing simple multi-layer perceptrons, achieves an mR of 32.38% on RSITMD. Subsequently, Adapter-Former [16] and Uni-Adapter [17], integrated deeper within the Transformer layers, elevate the performance to the 34%–39% range. Notably, while domain-tailored methods like PE-RSITR [12] and HarMA [18] exhibit strong capabilities, our HGMA-RSITR establishes highly competitive results across all fine-grained recall metrics (R@1, R@5, and R@10) while requiring exceptionally minimal trainable parameters .

We attribute this highly favorable performance-efficiency trade-off to two synergistic structural designs:

Internal Self-Attention for Spatial Reasoning: To address the cluttered backgrounds inherent in remote sensing imagery, we introduce a Multi-Head Attention (MHA) module to replace traditional linear bottlenecks. This design effectively captures complex contextual correlations among local geographic objects within a reduced-dimensional subspace, thereby encouraging the extraction of more discriminative, domain-specific features.

Hierarchical Gating for Prior Preservation: To mitigate the risk of catastrophic forgetting during cross-domain transfer, we replace fixed residual connections with learnable scalar gates (e.g., γ). Acting as adaptive valves, these gates dynamically regulate the injection of downstream domain features into the frozen backbone. This ensures an optimal balance between adapting to complex remote sensing semantics and preserving CLIP's robust, generalizable priors.

3.3 Ablation Study

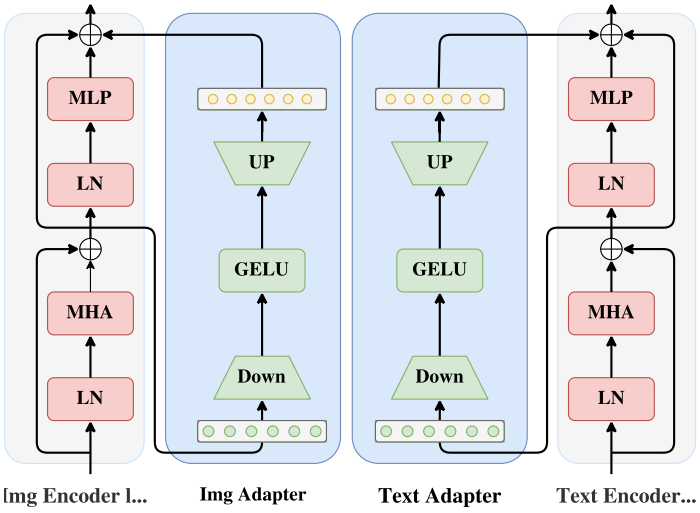

Fig. 3. Architecture of the baseline adapter.

Table 3. Module ablation experimental results.

Method					Text retrieval R@1 / R@5 / R@10	Image retrieval R@1 / R@5 /R@10	mR
Baseline	MHA	Gate1	Gate2	Gate3			
√					24.34 / 48.23 / 60.18	22.48 / 54.91 / 71.06	46.87
√			√		24.12 / 48.67 / 59.07	22.74 / 53.89 / 70.97	46.58
√	√				22.57 / 42.04 / 53.76	17.26 / 48.14 / 67.74	41.92
√	√	√			26.77 / 48.23 / 62.17	20.53 / 53.50 / 71.24	47.07
√	√	√	√		27.88 / 49.34 / 61.28	21.55 / 55.44 / 71.37	47.81
√	√	√	√	√	26.77 / 50.22 / 63.5	22.3 / 55.93 / 72.08	48.47

To comprehensively evaluate the individual contributions of the proposed components in the cross-modal remote sensing retrieval task, we conduct detailed ablation studies on the RSITMD dataset. The evaluated modules include the Multi-Head Attention (MHA) module, the hierarchical gating mechanism (Gate1 and Gate2), and the Cross-Modal Semantic Perturbation (CMSP) strategy. The architecture of the baseline adapter is illustrated in Figure 3, and the quantitative results are summarized in Table 2.

As shown in Table 3, the baseline model achieves a mean Recall (mR) of 46.87%. Interestingly, when solely introducing the MHA module into the baseline network, the performance experiences a slight degradation, with the mR dropping to 46.58%. We attribute this phenomenon to the fact that while MHA can explicitly model global dependencies among features, its unconstrained, dense attention weights are highly susceptible to noise and irrelevant background clutter inherent in complex remote sensing scenes. This observation strongly validates the necessity of designing a hierarchical gating mechanism. By progressively integrating Gate1 and Gate2, the model gains the capacity to adaptively regulate feature representations. Specifically, Gate1 stabilizes the shallow features to enhance the capture of local semantics and fine-grained information, whereas Gate2 dynamically filters cross-modal features in a higher-level semantic space to mitigate distribution shifts. Through this synergistic interaction, the gating mechanism successfully filters the redundant noise potentially introduced by the raw MHA, thereby elevating the overall mR to a robust 47.81%.

Finally, deploying the CMSP strategy on top of the complete architecture yields optimal performance across both image-to-text and text-to-image retrieval tasks. By introducing controlled cross-modal semantic perturbations during the training phase, CMSP encourages the model to maintain stable matching relationships despite semantic variations. This effectively enhances the model's robustness against semantic shifts and annotation noise. Experimental results indicate that this strategy is particularly effective in improving mid-to-high recall metrics (R@5 and R@10), ultimately pushing the final mR to 48.47%.

The ablation results suggest a clear division of labor and strong structural synergy among our proposed designs. The MHA provides the foundational capacity for global feature interaction, the hierarchical gating mechanism acts as a multi-level semantic regulator to prevent the risk of noise amplification from unconstrained attention, and the CMSP strategy maximizes the generalization and anti-interference capabilities of the model at the training level.

3.4 Hyperparameter Analysis

To systematically evaluate the sensitivity of the retrieval performance to the perturbation probability parameter α within the cross-modal semantic perturbation mechanism, corresponding experiments are conducted on the RSITMD dataset, with quantitative results detailed in Table 4. The experimental data indicate that as the value of α increments, the mean recall exhibits a typical inverted-U trajectory, achieving an empirical peak of 48.47% when α equals 0.3. Considering the formulation of this data augmentation mechanism, this empirical behavior might be explained from the perspective of data distribution shifts and model fitting tendencies. When α is assigned relatively small values, the triggering frequency of image mixing and text concatenation remains limited. This limitation may prevent the model from observing sufficiently diverse composite semantic samples, potentially underutilizing the generalization potential of the mechanism. Conversely, when α reaches higher levels, an excessive proportion of composite scene pseudo-samples is generated during the training process. Because these pseudo-samples exhibit artificially superimposed joint semantics in both visual and textual modalities, their over-introduction might cause a noticeable shift between the training data distribution and the true distribution of discrete scenes in the test set. Under such circumstances, the model may tend to overfit to parsing these complex concatenated semantics, which could paradoxically impair its precise alignment capability when processing standard unmixed remote sensing scenes. Therefore, setting α to 0.3 appears to strike a reasonable balance between enriching the training data manifold and preserving the original data distribution characteristics, thereby improving the overall robustness of the model without compromising the fundamental retrieval precision.

Table 4. Retrieval results with different perturbation probabilities α on the RSITMD dataset.

α	Text retrieval			Image retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	mR
0.1	25.22	50.88	61.73	22.12	54.51	70.58	47.51
0.2	24.78	50.00	64.16	22.35	53.27	70.22	47.46
0.3	26.77	50.22	63.50	22.30	55.93	72.08	48.47
0.4	25.00	47.57	61.06	21.86	53.89	71.99	46.90
0.5	23.23	47.35	61.73	20.31	53.58	72.52	46.45

4 Conclusions and Future Work

To address the issues of domain differences, limited data, and high fine-tuning costs in remote sensing image-text retrieval, this paper proposes a parameter-efficient fine-tuning framework. The framework utilizes a hierarchical gating multimodal adapter to achieve the dynamic regulation and fusion of cross-modal features, and incorporates a cross-modal semantic perturbation strategy to mitigate data distribution bias and improve model generalization. Experimental results indicate that this method effectively improves retrieval accuracy with minimal additional parameters, achieving a reasonable balance between parameter efficiency and performance. Future work will explore more refined feature selection strategies to enhance the recognition of fine-grained targets in complex scenes, and consider integrating self-supervised learning or visual generative priors to reduce reliance on high-quality paired annotations, thereby improving applicability under low-resource conditions.

References

1. Sudmanns, M., Tiede, D., Lang, S., et al.: Big earth data: Disruptive changes in earth observation data management and analysis? International Journal of Digital Earth 13(7), 832-850 (2020)
2. Yuan, Z., Zhang, W., Tian, C., et al.: Remote sensing cross-modal text-image retrieval based on global and local information. IEEE Transactions on Geoscience and Remote Sensing 60, 1-16 (2022)
3. Pan, J., Ma, Q., Bai, C.: Reducing semantic confusion: Scene-aware aggregation network for remote sensing cross-modal retrieval. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, pp. 398-406 (2023)
4. Tang, X., Wang, Y., Ma, J., et al.: Interacting-enhancing feature transformer for cross-modal remote-sensing image and text retrieval. IEEE Transactions on Geoscience and Remote Sensing 61, 1-15 (2023)
5. Pan, J., Ma, Q., Bai, C.: A prior instruction representation framework for remote sensing image-text retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 611-620 (2023)
6. Faghri, F., Fleet, D.J., Kiros, J.R., et al.: Vse++: Improving visual-semantic embeddings. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 935-943 (2018)
7. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, pp. 8748-8763 (2021)
8. Zhang, Z., Zhao, T., Guo, Y., et al.: RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing 62, 1-23 (2024)
9. Wang, Z., Prabha, R., Huang, T., et al.: Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 5805-5813 (2024)

10. Hu, G., Wen, Z., Lv, Y., et al.: Global–local information soft-alignment for cross-modal remote-sensing image–text retrieval. IEEE Transactions on Geoscience and Remote Sensing 62, 1-15 (2024)
11. Pan, J., Ma, M., Ma, Q., et al.: PIR: Remote sensing image-text retrieval with prior instruction representation learning. IEEE Transactions on Geoscience and Remote Sensing 62, 1-13 (2024)
12. Yuan, Y., Zhan, Y., Xiong, Z.: Parameter-efficient transfer learning for remote sensing image–text retrieval. IEEE Transactions on Geoscience and Remote Sensing 61, 1-14 (2023)
13. Lu, X., Wang, B., Zheng, X., et al.: Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing 56(4), 2183-2195 (2018)
14. Yuan, Z., Zhang, W., Fu, K., et al.: Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. IEEE Transactions on Geoscience and Remote Sensing 60, 1-19 (2022)
15. Gao, P., Geng, S., Zhang, R., et al.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision 132(2), 581-595 (2024)
16. Chen, S., Ge, C., Tong, Z., et al.: Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems 35, 16664-16678 (2022)
17. Lu, H., Huo, Y., Yang, G., et al.: UniAdapter: Unified parameter-efficient transfer learning for cross-modal modeling. In: The Twelfth International Conference on Learning Representations (2024)
18. Huang, T.: Efficient remote sensing with harmonized transfer learning and modality alignment. In: ICLR Workshop on Machine Learning for Remote Sensing (ML4RS) (2024).