PDAR-RSITR: A Progressive Decoupling-Aggregation-Refinement Framework for Remote Sensing Image-Text Retrieval

Shuhuai Wang[1], **Songwei Pei**[1*], **Bingfeng Liu**[1], **Duo Chai**[1],
Yuanzhou Huang[1], **Jia Liu**[1], **Qian Li**[1], **Shangguang Wang**[1]
[1]School of Computer Science (National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications, Beijing, China
{wangshuhuai, peisongwei, bfliu, chaiduo, huangyuanzhou, liujia_bupt, li.qian, sgwang}@bupt.edu.cn

Abstract

Remote Sensing Image-Text Retrieval (RSITR) aims to achieve precise retrieval between remote sensing images and textual descriptions. However, existing methods neglect the multi-dimensional cognitive attributes inherent in remote sensing data and struggle to handle them simultaneously, leading to suboptimal retrieval performance. In this paper, we propose a novel Progressive Decoupling-Aggregation-Refinement framework for RSITR (PDAR-RSITR) to comprehensively capture multi-dimensional cognitive attributes. Specifically, we adapt Multi-dimensional Cognitive Decoupling to learn features across cognitive attributes of different dimensions via multi-process clustering. Subsequently, we utilize Salient Cognitive Aggregation to select and aggregate the most salient attributes via dynamic routing. Furthermore, we propose Expert Collaborative Refinement to enhance critical cross-modal relationships via three complementary expert perspectives. Extensive experimental results demonstrate that PDAR-RSITR significantly outperforms existing state-of-the-art methods across multiple metrics.

1 Introduction

The emergence of massive remote sensing data has made efficient analysis methods a core issue [Zhou *et al.*, 2023; Li *et al.*, 2025], and how to extract key information from large-scale datasets has attracted widespread attention [Zhang *et al.*, 2023; Ji *et al.*, 2023; Zhang *et al.*, 2025; Yuan *et al.*, 2022b]. Remote Sensing Image-Text Retrieval (RSITR) is dedicated to establishing bidirectional mapping relationships between images and texts, achieving cross-modal semantic association, and serves as one of the core technologies for remote sensing information extraction and analysis [Fernandez-Beltran *et al.*, 2021; Mao *et al.*, 2018; Chen *et al.*, 2020].

Traditional methods [Zheng *et al.*, 2023; Zhang *et al.*, 2023] rely on CNNs and RNNs for feature encoding and project the features into a shared embedding space to compute similarity for retrieval. However, these traditional backbones have limited encoding capabilities [Pan *et al.*, 2023; Tang *et al.*, 2024], making it difficult to handle remote sensing data containing multi-dimensional cognitive attributes, resulting in poor performance. Recently, researchers utilized CLIP [Radford *et al.*, 2021] to obtain better features to enhance performance. For instance, PE [Yuan *et al.*, 2023] integrates adapters to enhance encoder performance, while RSCLIP [Cha *et al.*, 2025] and RemoteCLIP [Cha *et al.*, 2025] improve representation capability through enhanced training data quality and larger dataset scales. Nevertheless, these methods lack mechanisms to comprehensively capture multi-dimensional cognitive attributes, thus constraining performance gains in RSITR task.

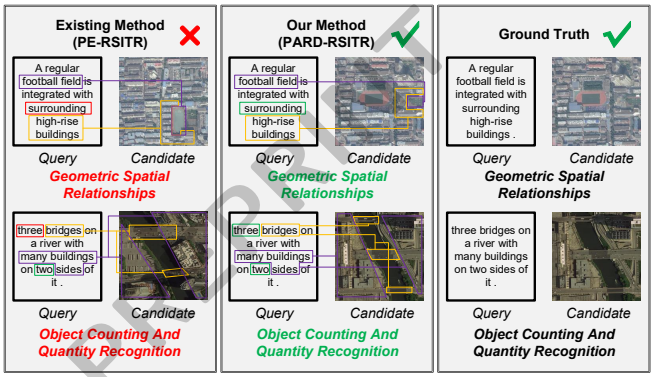

Figure 1: Existing RSITR methods struggle to simultaneously handle multi-dimensional cognitive attributes, leading to retrieval failures, while our PDAR-RSITR comprehensively captures these attributes, enabling accurate retrieval.

Remote sensing data contains multi-dimensional cognitive attributes such as Geometric Spatial Relationships (GSR) and Object Counting and Quantity Recognition (OCQR), whose entanglement with objects severely impedes accurate retrieval. As shown in Figure 1, existing methods struggle to handle these multi-dimensional cognitive attributes simultaneously. For instance, given the query "A regular football field is integrated with surrounding high-rise buildings," existing methods may capture the OCQR attribute (i.e., football field and buildings) but overlook the GSR attribute (i.e., surrounding), resulting in candidates where only some buildings are present around the football field. Similarly, for "three

*Corresponding author.

bridges on a river with many buildings on two sides of it," methods focus on GSR (i.e., bridges and two sides) but miss OCQR (i.e., three), resulting in candidates with only two bridges. This demonstrates that existing methods lack mechanisms to comprehensively capture multi-dimensional cognitive attributes, limiting retrieval accuracy.

To overcome the above challenges, we propose a novel **P**rogressive **D**ecoupling-**A**ggregation-**R**efinement framework for **R**emote **S**ensing **I**mage-**T**ext **R**etrieval, termed PDAR-RSITR. Specifically, inspired by the recent success of CLIP-MoE [Zhang *et al.*, 2025] in general vision-language models, we adapt Multi-dimensional Cognitive Decoupling to the remote sensing domain to learn features of cognitive attributes across different dimensions via multi-process clustering, which learns each attribute separately through a progressive process for in-depth understanding. Subsequently, we leverage Salient Cognitive Aggregation to select and aggregate the most salient attributes via dynamic routing, which efficiently extracts clear feature representations. Furthermore, we propose Expert Collaborative Refinement to enhance critical cross-modal relationships via three complementary expert perspectives, thereby improving retrieval accuracy through better cross-modal relationships. We conducted extensive experiments on four widely-used datasets to validate the effectiveness of our method. The main contributions are summarized as follows:

- We propose PDAR-RSITR, a novel Progressive Decoupling-Aggregation-Refinement framework. This framework effectively addresses the challenge of handling multi-dimensional cognitive attributes in remote sensing data.
- We integrate three synergistic stages: (1) adapting a multi-dimensional cognitive decoupling stage to deeply understand each attribute, (2) leveraging a salient cognitive aggregation stage to aggregate the most salient attributes, and (3) proposing a novel expert collaborative refinement stage to enhance critical cross-modal relationships.
- Extensive experimental results show that the proposed method achieves significant improvements in overall retrieval performance, and attains state-of-the-art results on multiple mainstream performance metrics.

2 Related Work

In recent years, the RSITR field has continued to make progress. A variety of methods, such as LW-MCR [Yuan *et al.*, 2021], AMFMN [Yuan *et al.*, 2022a], MSITA [Chen *et al.*, 2023], and MSA [Yang *et al.*, 2024], employ traditional backbones and continuously advance model architectural designs. However, due to the limited capabilities of these backbone networks, breakthroughs in performance metrics have not been achieved despite advances in model innovation. Consequently, researchers have begun adopting CLIP or its variants for this task, achieving superior results. PE [Yuan *et al.*, 2023] integrated adapters into pre-trained CLIP models to achieve parameter-efficient transfer learning, further enhancing performance. RSCLIP [Cha *et al.*, 2025] collected approximately 9.6 million vision-language pairs to train a better model, while RemoteCLIP [Liu *et al.*, 2024] generated a pre-training dataset 12 times larger than all previously available datasets combined to further boost model performance.

Although previous works have achieved impressive performance in RSITR, existing methods lack mechanisms to comprehensively capture multi-dimensional cognitive attributes, effectively select and aggregate the most salient attributes, and strategically enhance critical cross-modal relationships to enable accurate retrieval.

3 Method

The overview of our PDAR-RSITR is illustrated in Figure 2. We first introduce the problem formulation and preliminaries in Sec. 3.1 and the overall framework in Sec. 3.2. Then, we detail our proposed method across three synergistic stages: the Multi-dimensional Cognitive Decoupling Stage (MCDS) in Sec. 3.3 (Figure 2①), the Salient Cognitive Aggregation Stage (SCAS) in Sec. 3.4 (Figure 2②), and the Expert Collaborative Refinement Stage (ECRS) in Sec. 3.5 (Figure 2③).

3.1 Problem Formulation and Preliminaries

For RSITR, given a dataset $\mathcal{D} = (I_i, T_i)_{i=1}^{N}$ containing N image-text pairs, our goal is to learn clear feature representations for retrieval between remote sensing images and text descriptions. Each image I_i and text $T_i = (w_1, w_2, \ldots, w_L)$ are mapped into a shared embedding space, where matched pairs have high similarity scores. RemoteCLIP follows the CLIP architecture: The image encoder first generates patch tokens I_{patch} through patch embedding, then prepends a learnable classification token I_{cls} and adds positional embeddings I_{pos} to obtain $X_I^0 = [I_{cls}; I_{patch}] + I_{pos}$. The text encoder tokenizes captions using byte-pair encoding, adding special tokens [SOS] and [EOS] to generate $T_{token} = ([SOS], w_1, w_2, \ldots, w_L, [EOS])$, then adds positional embeddings T_{pos} to obtain $X_T^0 = T_{token} + T_{pos}$. Subsequently, X_I^0 and X_T^0 are processed through L transformer blocks, with the ℓ-th layer represented as:

$$\begin{aligned} \hat{X}^{\ell} &= X^{\ell-1} + \text{Attn}(\text{LN}(X^{\ell-1})), \\ X^{\ell} &= \hat{X}^{\ell} + \text{FFN}(\text{LN}(\hat{X}^{\ell})), \end{aligned} \tag{1}$$

where X represents image or text features. Finally, the CLS token and EOS token serve as global image feature $\mathbf{v}$ and text feature $\mathbf{u}$, respectively, which are linearly projected to a D-dimensional shared embedding space and L2-normalized. In general, the base model is trained using standard contrastive learning:

$$\mathcal{L}_{CLIP} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathbf{v}_i^T \mathbf{u}_i/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{v}_i^T \mathbf{u}_j/\tau)}, \tag{2}$$

where τ is the temperature parameter.

3.2 Overall Framework

Our PDAR-RSITR framework consists of three synergistic stages. First, we utilize the adapted MCDS and SCAS to

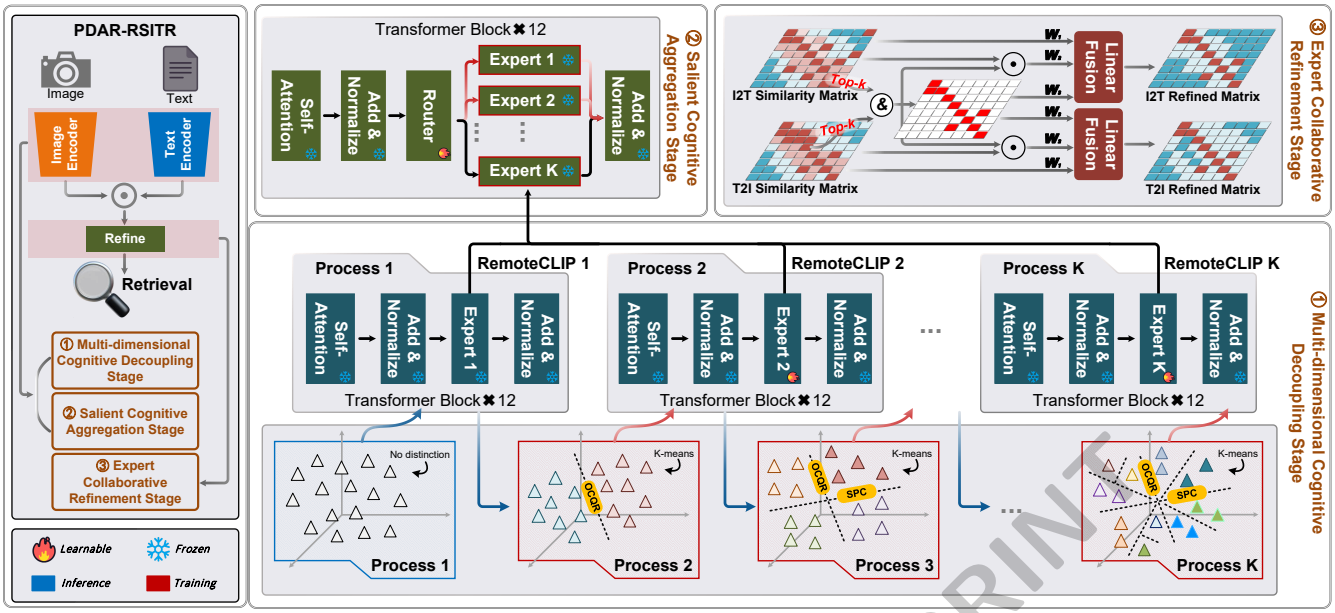

Figure 2: The PDAR-RSITR framework. Left: main retrieval pipeline. Right: ① Multi-dimensional Cognitive Decoupling Stage (MCDS), ② Salient Cognitive Aggregation Stage (SCAS), ③ Expert Collaborative Refinement Stage (ECRS).

train the image and text encoders. The encoders comprehensively capture multi-dimensional cognitive attributes, select and aggregate the most salient features, and efficiently extract clear feature representations. During the inference phase, the trained encoders are employed to extract features for images and texts, and compute their initial similarity scores in the shared embedding space. Then enhance critical cross-modal relationships through ECRS, improving retrieval accuracy.

3.3 Multi-dimensional Cognitive Decoupling Stage

Inspired by multi-stage contrastive learning [Zhang *et al.*, 2024], we progressively perform an inference-clustering-training process. Through iterative inference and clustering, we gradually disentangle the entangled attribute space to generate joint attribute labels. Different from prior methods, to adapt to the remote sensing domain, we utilize RemoteCLIP [Liu *et al.*, 2024] instead of generic CLIP as the base model. Meanwhile, by fine-tuning during training, we construct a complementary expert ensemble (i.e., a set of RemoteCLIP models), enabling each expert to focus on different dimensional cognitive attributes. Specifically, we perform fine-tuning over K processes on the FFN layers of K RemoteCLIP models, where K denotes the total number of progressive processes and k represents the current process index ($k \in \{1, 2, ..., K\}$). Similarly, our consideration of the K value follows the number of salient cognitive attributes in remote sensing, as analyzed in [Adejumo *et al.*, 2025]. These FFN layers are complementary and are finally integrated into the MoE architecture as experts. The progressive process of MCDS is as follows:

Progressive process k ($k = 1$): In this process, we rely solely on the inference of the base RemoteCLIP model and obtain initial attribute labels through clustering. Specifically, we extract image features $\mathbf{v}_i^{(1)}$ and text features $\mathbf{u}_i^{(1)}$, respectively, and obtain image attribute labels $y_i^{(1,I)}$ and text attribute labels $y_i^{(1,T)}$ via K-means clustering. The joint attribute label is denoted as $\hat{y}_i^{(1)} = (y_i^{(1,I)}, y_i^{(1,T)})$.

Progressive process k ($k > 1$): In these processes, we fine-tune only the FFN layers of the base RemoteCLIP model. First, we employ the RemoteCLIP model fine-tuned in process $k-1$ to extract image and text features $\mathbf{v}_i^{(k)}$ and $\mathbf{u}_i^{(k)}$, respectively, and perform clustering to obtain new attribute labels $\hat{y}_i^{(k)}$. To fully exploit the attribute labels of samples across different processes, we accumulate the joint attribute labels obtained at each process to form cumulative joint attribute labels $\tilde{y}_i^{(k)} = [\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \ldots, \hat{y}_i^{(k)}]$. Next, during the fine-tuning of the RemoteCLIP model in process k, we select negative samples only from the set of samples whose cumulative joint attribute labels are completely identical. For example, when image I_i serves as the anchor, the negative sample set is defined as

$$N_i^{(k)} = \{T_j \mid j \neq i, \tilde{y}_j^{(k-1)} = \tilde{y}_i^{(k-1)}\}, \tag{3}$$

that is, texts whose attribute labels in the previous $k-1$ processes are all consistent with the anchor but do not match the anchor itself. This design encourages the model to focus on attributes that were not distinguished in the previous processes.

The training loss function for each progressive process is as follows:

$$L^{(k)} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s_{ii}^{(k)}}}{e^{s_{ii}^{(k)}} + \sum_{j \in N_i^{(k)}} e^{s_{ij}^{(k)}}}, \tag{4}$$

where $s_{ij}^{(k)} = \mathbf{v}_i^{(k)\top} \mathbf{u}_j^{(k)} / \tau$ and τ is the temperature parameter.

3.4 Salient Cognitive Aggregation Stage

Following the upcycling strategy in CLIP-MoE [Zhang *et al.*, 2025], after all stages fine-tuning, we obtain a set of RemoteCLIP models to construct a complementary expert ensemble, and integrate these specially fine-tuned FFN layers into a MoE architecture:

$$\{E_1, E_2, ..., E_K\} = \{\text{FFN}_{\text{stage-}j}\}_{j=1}^{K}, \tag{5}$$

After obtaining these complementary experts, the key challenge is how to dynamically select the most relevant expert subsets, extracting salient attributes to efficiently extract clear feature representations. Therefore, we employ an adaptive routing mechanism to address this issue. For each transformer layer ℓ, given input $\mathbf{h}$, we compute expert selection probabilities through a learnable router:

$$\mathbf{g}^{(\ell)} = \text{Softmax}(\mathbf{W}_r^{(\ell)}\mathbf{h}), \tag{6}$$

where $\mathbf{W}_r^{(\ell)}$ are the router parameters. To control computational cost, we select top-r experts: $\mathcal{S} = \text{TopR}(\mathbf{g}^{(\ell)}, r)$ and renormalize the routing weights:

$$\hat{g}_j^{(\ell)} = \begin{cases} \frac{g_j^{(\ell)}}{\sum_{i\in\mathcal{S}} g_i^{(\ell)}} & \text{if } j \in \mathcal{S} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

The final output is obtained through weighted combination of selected experts' outputs:

$$\mathbf{h}_{out} = \sum_{j=1}^{K} \hat{g}_j^{(\ell)} \cdot E_j^{(\ell)}(\mathbf{h}), \tag{8}$$

To prevent expert collapse where only a few experts are utilized, we use load balancing constraints. The training objective combines standard contrastive loss with a load balancing term:

$$\mathcal{L}_{total} = \mathcal{L}_{CLIP} + \lambda \sum_{\ell=1}^{L} \mathcal{L}_{balance}^{(\ell)}, \tag{9}$$

where $\lambda = 0.01$ is the balancing weight. The load balancing loss encourages uniform expert usage across batches:

$$\mathcal{L}_{balance}^{(\ell)} = K \sum_{j=1}^{K} f_j^{(\ell)} \cdot p_j^{(\ell)}, \tag{10}$$

where $f_j^{(\ell)} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}[\text{argmax}(\mathbf{g}_b^{(\ell)}) = j]$ represents the selection frequency of expert j in layer ℓ, and $p_j^{(\ell)} = \frac{1}{B}\sum_{b=1}^{B} g_{b,j}^{(\ell)}$ represents the average gate probability of expert j for batch size B. Finally, an end-to-end global fine-tuning is applied to ensure global consistency across the framework.

3.5 Expert Collaborative Refinement Stage

To further improve retrieval accuracy, we recognize that initial rankings may overlook important cross-modal relationships. Inspired by the expert collaboration mechanism in MoE architectures, we introduce an ECRS that optimizes matching results from multiple perspectives by leveraging the ranking confidence expert, symmetry verification expert, and saliency assessment expert working collaboratively, thereby significantly enhancing retrieval accuracy.

Specifically, given the initial similarity matrix $\mathbf{S}$. The ranking confidence expert directly utilizes the initial similarity scores to preserve the original cross-modal matching information:

$$E_{\text{rank}}(\mathbf{S}) = w_1 \cdot \mathbf{S} \tag{11}$$

where w_1 is the confidence weight. The symmetry verification expert identifies high-quality matches through bidirectional retrieval consistency. This expert computes the intersection of the image-to-text top-E mask $\mathbf{M}_{i2t}^{(E)}$ and the text-to-image top-L mask $\mathbf{M}_{t2i}^{(L)}$:

$$\mathbf{M}_{\text{sym}} = \mathbf{M}_{i2t}^{(E)} \odot \mathbf{M}_{t2i}^{(L)} \tag{12}$$

where $\odot$ denotes element-wise multiplication. This mask only retains matching pairs that rank highly in both bidirectional retrievals, ensuring bidirectional consistency of matches. The output of this expert is:

$$E_{\text{sym}}(\mathbf{S}, \mathbf{M}_{\text{sym}}) = w_2 \cdot \mathbf{S} \odot \mathbf{M}_{\text{sym}} \tag{13}$$

where w_2 is the confidence weight. The saliency assessment expert provides additional saliency enhancement to matching pairs that pass bidirectional verification, reinforcing distinctive matches with high discriminability:

$$E_{\text{sal}}(\mathbf{M}_{\text{sym}}) = w_3 \cdot \mathbf{M}_{\text{sym}} \tag{14}$$

where w_3 is the confidence weight. This expert provides additional confidence boost to matching pairs that pass symmetry verification, effectively avoiding ambiguous matching results. Finally, we integrate the outputs of the three experts through a weighted fusion strategy to obtain the refined similarity matrix:

$$\mathbf{S}_{\text{refine}} = \frac{E_{\text{rank}}(\mathbf{S}) + E_{\text{sym}}(\mathbf{S}, \mathbf{M}_{\text{sym}}) + E_{\text{sal}}(\mathbf{M}_{\text{sym}})}{1 + \xi} \tag{15}$$

where ξ is the smoothing factor. Through the collaborative work of the three experts, this mechanism can effectively improve the accuracy of retrieval results.

4 Experiments

4.1 Datasets and Evaluation Metrics

To evaluate the effectiveness of the proposed method, we select four widely-used datasets: RSICD [Lu *et al.*, 2017], RSITMD [Yuan *et al.*, 2022a] and UCM [Qu *et al.*, 2016] for retrieval performance evaluation, and RSMMVP [Adejumo *et al.*, 2025] for expert complementarity assessment. For retrieval performance evaluation, we adopt Recall at K ($R@K$, where $K = 1, 5, 10$) [Lu *et al.*, 2017] as the evaluation metric, and calculate $R@K$ for both Image-to-Text (I2T) and Text-to-Image (T2I) retrieval tasks to measure the model's performance. In addition, we report the mean Recall (mR) as an overall performance indicator. For expert complementarity assessment, we use Blind Pair Accuracy (BPA) [Adejumo *et al.*, 2025] as the evaluation metric. A prediction is considered successful only if the model makes correct judgments on both pairs of images.

Methods	RSICD				RSITMD				UCM			
	R@1	R@5	R@10	mR_{I2T}	R@1	R@5	R@10	mR_{I2T}	R@1	R@5	R@10	mR_{I2T}
LW-MCR$_{\text{TGRS'21}}$	4.57	13.71	20.11	12.80	9.07	22.79	38.05	23.30	12.38	43.81	59.52	38.57
AMFMN$_{\text{TGRS'22}}$	5.05	14.53	21.57	13.72	11.06	25.88	39.82	25.59	12.86	51.90	66.67	43.81
MSITA$_{\text{TGRS'23}}$	8.67	22.71	33.91	21.76	15.22	34.20	47.65	32.36	16.86	49.33	73.33	46.51
MSA$_{\text{TGRS'24}}$	9.52	25.25	35.32	23.36	15.93	38.50	50.88	35.10	13.33	59.05	77.14	49.84
CLIP-zero-shot$_{\text{ICML'21}}$	6.77	15.37	23.15	15.10	9.29	26.33	37.39	24.34	10.95	34.76	55.71	33.81
Linear probe$_{\text{ICML'21}}$	8.46	24.41	37.72	23.53	17.02	33.12	48.35	32.83	13.33	52.71	77.62	47.89
CoOp$_{\text{IJCV'22}}$	6.32	15.89	27.60	16.60	12.19	30.69	42.82	28.57	7.19	42.76	74.19	41.38
VPT$_{\text{ECCV'22}}$	7.23	19.40	30.77	19.13	14.98	32.05	40.15	29.06	8.86	47.81	78.62	45.10
AdaptFormer$_{\text{NIPS'22}}$	12.46	28.49	41.86	27.60	16.71	30.16	42.91	29.93	16.92	54.39	77.46	49.59
RSCLIP$_{\text{JAG'23}}$	10.43	25.34	39.34	25.04	19.25	36.06	46.68	33.99	19.05	56.19	-	-
PE$_{\text{TGRS'23}}$	14.13	31.51	44.78	30.14	23.67	44.07	60.36	42.70	*22.71*	55.81	80.33	52.95
LAPS$_{\text{CVPR'24}}$	14.42	31.85	43.29	29.85	19.87	36.72	50.23	35.61	18.62	52.41	74.36	48.46
CLIP-Adapter$_{\text{IJCV'24}}$	7.11	19.48	31.01	19.20	12.83	28.84	39.05	26.91	9.83	42.52	66.79	39.71
RemoteCLIP$_{\text{TGRS'24}}$	*17.02*	*37.97*	*51.51*	*35.50*	*27.88*	*50.66*	*65.71*	*48.08*	20.48	*59.85*	*83.33*	*54.55*
CLGSA$_{\text{TGRS'25}}$	14.38	36.73	45.35	32.15	26.77	48.23	60.39	45.13	19.05	53.81	81.90	51.59
PDAR-RSITR (ours)	**18.76**	**39.79**	**54.24**	**37.60**	**28.54**	**52.34**	**66.05**	**48.98**	**22.94**	**60.01**	**84.39**	**55.78**

Table 1: Comparison of I2T retrieval tasks performance between PDAR-RSITR and other methods. **Bold**: best; *italic*: second best.

Methods	RSICD				RSITMD				UCM			
	R@1	R@5	R@10	mR_{T2I}	R@1	R@5	R@10	mR_{T2I}	R@1	R@5	R@10	mR_{T2I}
LW-MCR$_{\text{TGRS'21}}$	4.02	16.47	28.23	16.24	6.11	27.74	49.56	27.80	12.00	46.38	72.48	43.62
AMFMN$_{\text{TGRS'22}}$	5.05	19.74	31.04	18.61	9.82	33.94	51.90	31.89	14.19	51.71	78.48	48.13
MSITA$_{\text{TGRS'23}}$	6.13	21.98	35.39	21.17	12.15	39.92	57.72	36.60	14.29	57.16	91.58	54.34
MSA$_{\text{TGRS'24}}$	6.55	24.43	39.63	23.54	14.96	45.22	61.86	40.68	14.19	57.33	87.05	52.86
CLIP-zero-shot$_{\text{ICML'21}}$	5.01	15.75	24.21	15.00	7.79	23.67	38.89	23.45	7.33	35.14	54.57	32.35
Linear probe$_{\text{ICML'21}}$	7.81	25.89	42.47	25.39	13.33	41.80	63.89	39.67	14.43	59.42	90.28	54.71
CoOp$_{\text{IJCV'22}}$	4.31	17.89	31.73	17.98	9.16	33.85	54.35	32.45	9.78	48.34	87.41	48.51
VPT$_{\text{ECCV'22}}$	6.94	25.26	40.73	24.31	15.97	41.35	60.35	39.22	13.64	55.61	94.05	54.43
AdaptFormer$_{\text{NIPS'22}}$	9.09	29.89	46.81	28.60	14.27	41.53	61.46	39.09	18.74	*63.49*	91.16	57.80
RSCLIP$_{\text{JAG'23}}$	9.90	30.52	45.03	28.48	12.92	42.04	63.14	39.37	18.67	61.52	-	-
PE$_{\text{TGRS'23}}$	11.63	33.92	50.73	32.09	20.10	50.63	67.97	46.23	*18.82*	62.84	93.72	58.46
LAPS$_{\text{CVPR'24}}$	9.39	27.84	40.50	25.91	16.70	46.27	62.73	41.90	15.07	57.14	90.67	54.29
CLIP-Adapter$_{\text{IJCV'24}}$	7.67	24.87	39.73	24.09	13.30	40.20	60.06	37.85	14.61	51.03	84.14	49.93
RemoteCLIP$_{\text{TGRS'24}}$	*13.71*	*37.11*	*54.25*	*35.02*	*22.17*	*56.46*	*73.41*	*50.68*	18.67	61.52	*94.29*	*58.16*
CLGSA$_{\text{TGRS'25}}$	11.75	34.07	50.12	31.98	22.52	52.39	69.46	48.12	18.57	62.38	94.10	58.35
PDAR-RSITR (ours)	**15.54**	**42.46**	**58.57**	**38.85**	**26.16**	**59.07**	**75.78**	**53.67**	**19.56**	**63.62**	**95.61**	**59.60**

Table 2: Comparison of T2I retrieval tasks performance between PDAR-RSITR and other methods. **Bold**: best; *italic*: second best.

4.2 Implementation Details

All experiments are conducted on four NVIDIA A100 80GB GPUs. Following the protocol in [Yuan *et al.*, 2022a], the RSICD, RSITMD and UCM datasets are divided into training (80%), validation (10%), and test (10%) sets. The RSMMVP dataset is used exclusively as the test set. To ensure statistical validity, we exclude categories with insufficient samples (i.e., CRC and SPC, which contain only 1–3 pairs). It is worth noting that the original RSMMVP dataset was designed for visual question answering tasks; we adapt it for image-text retrieval by concatenating the relevant information.

4.3 Comparison Methods

We compare our method against 15 state-of-the-art methods, including traditional methods (LW-MCR [Yuan *et al.*, 2021], AMFMN [Yuan *et al.*, 2022a], MSITA [Chen *et al.*, 2023], and MSA [Yang *et al.*, 2024]), as well as recent CLIP-based methods such as CLIP [Radford *et al.*, 2021], Linear Probing [Radford *et al.*, 2021], CoOp [Zhou *et al.*, 2022], VPT [Jia *et al.*, 2022], AdaptFormer [Chen *et al.*, 2022], RSCLIP [Cha *et al.*, 2025], PE [Yuan *et al.*, 2023], LAPS [Fu *et al.*, 2024], CLIP-Adapter [Gao *et al.*, 2024], RemoteCLIP [Liu *et al.*, 2024], CLGSA [Chen *et al.*, 2025].

Traditional methods (LW-MCR, AMFMN, MSITA, MSA) primarily focus on multi-scale feature encoding and fine-grained alignment for cross-modal retrieval. For CLIP-based methods, Linear Probing adds trainable classifiers; CoOp and VPT introduce learnable prompts; AdaptFormer, CLIP-Adapter and PE insert lightweight adaptation modules; RSCLIP and RemoteCLIP are vision-language foundation models specifically designed for remote sensing; LAPS and CLGSA introduce new modules to enhance cross-modal interaction mechanisms.

4.4 Main Results

As shown in Tables 1 and 2, we compare the performance of our proposed method with existing methods on I2T and T2I retrieval tasks across three benchmark datasets. The results demonstrate that PDAR-RSITR consistently outperforms all baseline methods, achieving state-of-the-art performance in both retrieval directions. Compared to traditional methods, pre-trained vision-language models inherently exhibit significant performance improvements. Among CLIP-based methods, PDAR-RSITR achieves better results than other methods. This advantage is attributed to our method's ability

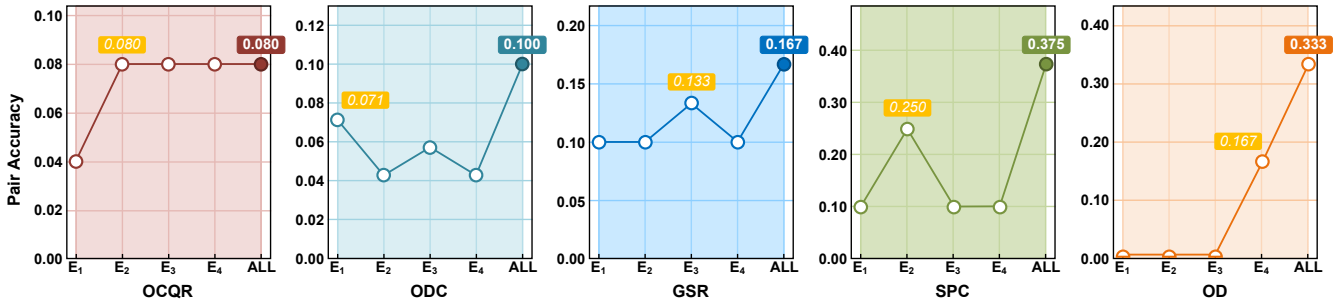

Figure 3: Individual expert performance across visual reasoning patterns on RSMMVP.

to comprehensively capture multi-dimensional cognitive attributes, select and aggregate the most salient features, and efficiently extract clear feature representations. Moreover, it can enhance critical cross-modal relationships, thereby improving retrieval accuracy.

Method			Performance Metrics		
MCDS	SCAS	ECRS	mR_{I2T}	mR_{T2I}	mR
			35.50	35.02	35.26
	✓		35.33	35.22	35.28
	✓	✓	35.67	35.83	35.75
✓			35.89	37.77	36.83
✓		✓	36.31	37.95	37.13
✓	✓		36.94	38.52	37.73
✓	✓	✓	**37.60**	**38.86**	**38.23**

Table 3: Ablation study results of PDAR-RSITR on the RSICD.

4.5 Ablation Study

To validate each component's effectiveness, we conduct ablation studies on RSICD, with results shown in Table 3. Without all components, the model degrades to the RemoteCLIP baseline. Without MCDS, four randomly fine-tuned experts yield negligible gains. Without SCAS, we use uniform weighting to integrate these specially fine-tuned layers into a MoE architecture. Although this approach achieves certain improvements, the performance is far inferior to that with SCAS. ECRS further improves performance through ranking confidence assessment, symmetry verification, and saliency evaluation. The complete PDAR-RSITR model achieves optimal performance, demonstrating that MCDS, SCAS, and ECRS synergistically contribute to superior retrieval accuracy.

4.6 Analysis of Expert Complementarity

We evaluated each expert individually on RSMMVP, as illustrated in Figure 3. The results reveal that Expert 1 excels at ODC, Expert 2 at SPC, Expert 3 at GSR, and Expert 4 at OD. OCQR is mastered by Experts 2, 3, and 4. When these experts are combined within PDAR-RSITR, the model achieves the highest accuracy across all attributes.

To verify PDAR-RSITR's ability to comprehensively consider multi-dimensional cognitive attributes, we conducted experiments on the RSMMVP dataset, encompassing Object Counting and Quantity Recognition (OCQR), Object Detection and Classification (ODC), Geometric Spatial Relationships (GSR), and Orientation Detection (OD), as shown in Figure 4. We compared PDAR-RSITR with MSA, PE, LAPS, and RemoteCLIP on the challenging BPA task. The results show that PDAR-RSITR achieves improvements across all attributes, while other methods fail with some reaching 0% accuracy.

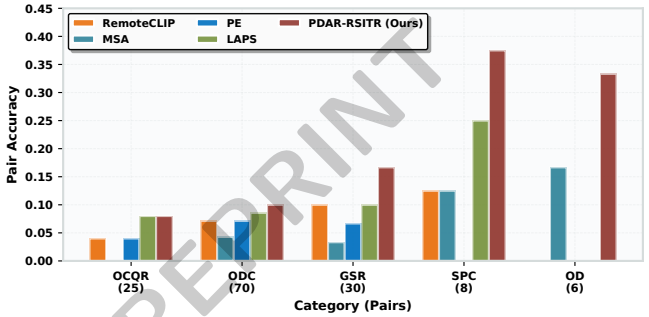

Figure 4: Expert complementarity analysis on BPA task accuracy across all attributes of RSMMVP.

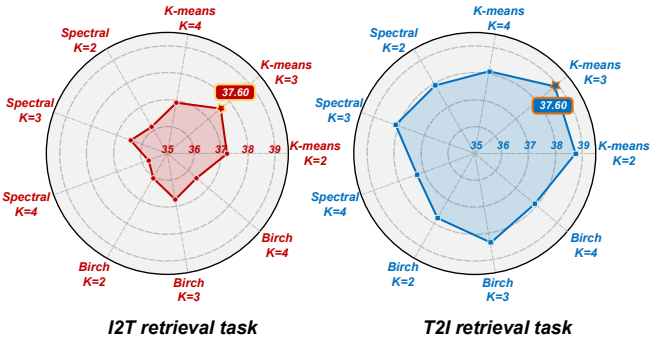

Figure 5: Comparison of clustering strategies for attribute label generation on RSICD

4.7 Clustering Strategy Analysis

To evaluate clustering methods for attribute label generation, we conducted experiments on RSICD with $\mathcal{K}$-means, Spectral Clustering, and Birch using $\mathcal{K} \in \{2, 3, 4\}$, as shown in Figure 5. The experimental results demonstrate that all

methods achieve peak performance at $\mathcal{K} = 3$. Specifically, $\mathcal{K}$-means lacks sufficient discriminability at $\mathcal{K} = 2$, reaches optimal granularity at $\mathcal{K} = 3$, and suffers from performance degradation due to over-fragmentation at $\mathcal{K} = 4$. Spectral Clustering exhibits higher sensitivity to the $\mathcal{K}$ value, showing significant performance deterioration at $\mathcal{K} = 4$ due to instability in eigenvalue decomposition. While Birch demonstrates stable performance, it consistently underperforms, constrained by the limitations of its tree-based strategy in high-dimensional spaces. $\mathcal{K}$-means achieves the best performance, benefiting from its effective convex space partitioning capability and computational efficiency of $O(n\mathcal{K}d)$, which is significantly superior to the $O(n^3)$ complexity of Spectral Clustering.

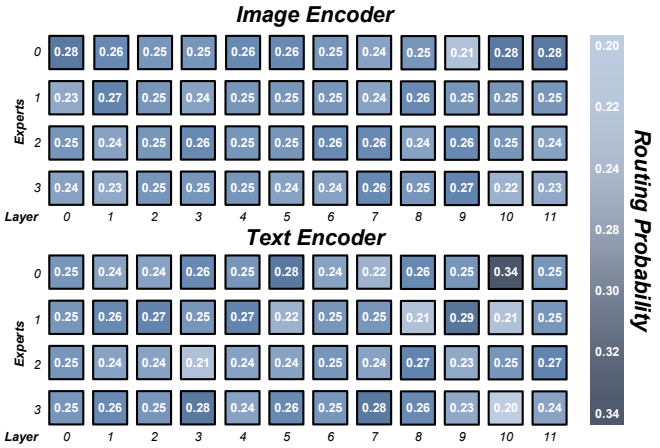

Figure 6: Routing mechanism analysis showing expert utilization patterns on RSICD.

4.8 Routing Mechanism Analysis

To evaluate whether PDAR-RSITR sufficiently utilizes all experts learned through MCDS, we analyzed the adaptive routing mechanism learned in SCAS. We used a PDAR-RSITR model with $\mathcal{K} = 4$ experts and top-2 routing trained on RSICD, and calculated the proportion of tokens assigned to each expert, as shown in Figure 6. As can be seen, each expert from the MCDS stages (each column in the heatmap) is relatively balanced, indicating that these experts are heavily utilized, which demonstrates that all MCDS stages provide useful experts.

4.9 Training and Inference Speed Comparison

To evaluate computational efficiency, Table 4 compares the training time, inference time, and accuracy of different methods on RSICD. While PE and RemoteCLIP offer fast inference, PE suffers from lower accuracy, and RemoteCLIP requires excessive training time. Conversely, MSA and LAPS are inefficient in both speed and performance. PDAR-RSITR achieves the highest accuracy with moderate training time. Its slight inference overhead compared to RemoteCLIP is due to MoE routing, which is a worthwhile trade-off for the significant performance gain.

Model	Time (s)		mR (%)
	Training	Inference	
MSA	16526	259	23.45
LAPS	15020	324	27.88
PE	2892	52	31.12
RemoteCLIP	140040	47	35.26
PDAR-RSITR	19436	83	38.23

Table 4: Comparison of training time, inference time, and accuracy across different methods on RSICD.

4.10 Sensitivity Analysis of ECRS Parameters

To investigate the impact of ECRS expert weights (w_1, w_2, w_3), we conducted a parameter sensitivity analysis on the RSICD as shown in Figure 7. Results show that w_1 (ranking confidence) has the most significant impact: performance steadily improves from 0.5 to 1.5 but declines at 2.0, indicating over-reliance on initial scores weakens multi-expert collaboration. The weights w_2 (symmetry verification) and w_3 (saliency assessment) show moderate and minimal effects respectively. The optimal configuration $w_1 : w_2 : w_3 = 1.5 : 1.0 : 1.0$ balances all three experts and serves as a robust default for different scenarios.

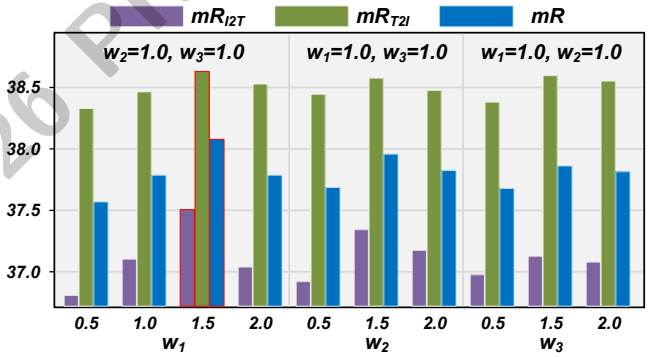

Figure 7: Sensitivity analysis of ECRS weight ratios on RSICD.

5 Conclusion

This paper presents PDAR-RSITR, a novel framework that addresses the challenge of comprehensively capturing multi-dimensional cognitive attributes in RSITR. We integrate three synergistic stages to comprehensively capture multi-dimensional cognitive attributes, select and aggregate the most salient features, and efficiently extract clear feature representations. Moreover, our method can enhance critical cross-modal relationships, thereby improving retrieval accuracy. Extensive experiments on four widely-used benchmark datasets demonstrate that PDAR-RSITR achieves state-of-the-art performance across multiple mainstream metrics.

Ethical Statement

There are no ethical issues.

Acknowledgements

This work was supported by the Beijing Municipal Science and Technology Program (Z251100003625011), the Beijing Natural Science Foundation (L252018), the NSFC through grant No.62402054, the China Postdoctoral Science Foundation through grant 2024M760279, and the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under grant BX20250390.

References

[Adejumo *et al.*, 2025] Abduljaleel Adejumo, Faegheh Yeganli, Clifford Broni-bediako, Aoran Xiao, Naoto Yokoya, and Mennatullah Siam. A vision centric remote sensing benchmark. arXiv preprint arXiv:2503.15816, 2025.

[Cha *et al.*, 2025] Keumgang Cha, Donggeun Yu, Junghoon Seo, Hyunguk Choi, and Taegyun Jeon. Pushing the limits of vision-language models in remote sensing without human annotations. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 249–254, 2025.

[Chen *et al.*, 2020] Yaxiong Chen, Xiaoqiang Lu, and Shuai Wang. Deep cross-modal image–voice retrieval in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, pages 7049–7061, 2020.

[Chen *et al.*, 2022] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, pages 16664–16678, 2022.

[Chen *et al.*, 2023] Yaxiong Chen, Jinghao Huang, Xiaoyu Li, Shengwu Xiong, and Xiaoqiang Lu. Multiscale salient alignment learning for remote-sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–13, 2023.

[Chen *et al.*, 2025] Xiumei Chen, Xiangtao Zheng, and Xiaoqiang Lu. Context-aware local–global semantic alignment for remote sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–12, 2025.

[Fernandez-Beltran *et al.*, 2021] Ruben Fernandez-Beltran, Begüm Demir, Filiberto Pla, and Antonio Plaza. Unsupervised remote sensing image retrieval using probabilistic latent semantic hashing. *IEEE Geoscience and Remote Sensing Letters*, pages 256–260, 2021.

[Fu *et al.*, 2024] Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *Conference on Computer Vision and Pattern Recognition*, pages 26307–26316, 2024.

[Gao *et al.*, 2024] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 581–595, 2024.

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

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727, 2022.

[Li *et al.*, 2025] Liangzhi Li, Ling Han, Yuanxin Ye, Yuming Xiang, and Tingyu Zhang. Deep learning in remote sensing image matching: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, pages 88–112, 2025.

[Liu *et al.*, 2024] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–16, 2024.

[Lu *et al.*, 2017] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, pages 2183–2195, 2017.

[Mao *et al.*, 2018] Guo Mao, Yuan Yuan, and Lu Xiaoqiang. Deep cross-modal retrieval for remote sensing image and audio. In *IAPR Workshop on Pattern Recognition in Remote Sensing*, pages 1–7, 2018.

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

[Qu *et al.*, 2016] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. Deep semantic understanding of high resolution remote sensing image. In *International Conference on Computer, Information and Telecommunication Systems*, pages 1–5, 2016.

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

[Tang *et al.*, 2024] Xu Tang, Dabiao Huang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Prior-experience-based vision-language model for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–13, 2024.

[Yang *et al.*, 2024] Rui Yang, Shuang Wang, Yingping Han, Yuanheng Li, Dong Zhao, Dou Quan, Yanhe Guo, Licheng Jiao, and Zhi Yang. Transcending fusion: A multi-scale alignment method for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–17, 2024.

[Yuan *et al.*, 2021] Zhiqiang Yuan, Wenkai Zhang, Xuee Rong, Xuan Li, Jialiang Chen, Hongqi Wang, Kun Fu,

and Xian Sun. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–19, 2021.

[Yuan *et al.*, 2022a] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–19, 2022.

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

[Yuan *et al.*, 2023] Yuan Yuan, Yang Zhan, and Zhitong Xiong. Parameter-efficient transfer learning for remote sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–14, 2023.

[Zhang *et al.*, 2023] Weihang Zhang, Jihao Li, Shuoke Li, Jialiang Chen, Wenkai Zhang, Xin Gao, and Xian Sun. Hypersphere-based remote sensing cross-modal text–image retrieval via curriculum learning. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–15, 2023.

[Zhang *et al.*, 2024] Jihai Zhang, Xiang Lan, Xiaoye Qu, Yu Cheng, Mengling Feng, and Bryan Hooi. Learning the unlearned: Mitigating feature suppression in contrastive learning. In *European Conference on Computer Vision*, pages 35–52, 2024.

[Zhang *et al.*, 2025] Jihai Zhang, Xiaoye Qu, Tong Zhu, and Yu Cheng. CLIP-MoE: Towards building mixture of experts for clip with diversified multiplet upcycling. In *Conference on Empirical Methods in Natural Language Processing*, pages 5406–5419, 2025.

[Zheng *et al.*, 2023] Chengyu Zheng, Ning Song, Ruoyu Zhang, Lei Huang, Zhiqiang Wei, and Jie Nie. Scale-semantic joint decoupling network for image-text retrieval in remote sensing. *ACM Transactions on Multimedia Computing, Communications and Applications*, pages 1–20, 2023.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pages 2337–2348, 2022.

[Zhou *et al.*, 2023] Weixun Zhou, Haiyan Guan, Ziyu Li, Zhenfeng Shao, and Mahmoud R Delavar. Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1447–1473, 2023.