PMPGuard: Catching Pseudo-Matched Pairs in Remote Sensing Image–Text Retrieval

Pengxiang Ouyang[1], Qing Ma[2*], Zheng Wang[1], Cong Bai[1,3],

[1]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[2]School of mathematical sciences, Zhejiang University of Technology, Hangzhou 310023, China
[3]Zhejiang Key Laboratory of Visual Information Intelligent Processing, Hangzhou 310023
oypx@zjut.edu.cn, maqing@zjut.edu.cn, zhengwang@zjut.edu.cn, congbai@zjut.edu.cn

Abstract

Remote sensing (RS) image–text retrieval faces significant challenges in real-world datasets due to the presence of Pseudo-Matched Pairs (PMPs), semantically mismatched or weakly aligned image–text pairs, which hinder the learning of reliable cross-modal alignments. To address this issue, we propose a novel retrieval framework that leverages Cross-Modal Gated Attention and a Positive–Negative Awareness Attention mechanism to mitigate the impact of such noisy associations. The gated module dynamically regulates cross-modal information flow, while the awareness mechanism explicitly distinguishes informative (positive) cues from misleading (negative) ones during alignment learning. Extensive experiments on three benchmark RS datasets, i.e., RSICD, RSITMD, and RS5M, demonstrate that our method consistently achieves state-of-the-art performance, highlighting its robustness and effectiveness in handling real-world mismatches and PMPs in RS image–text retrieval tasks.

Introduction

Remote-sensing (RS) image–text retrieval, which aims to establish accurate semantic correspondence between aerial imagery and natural-language descriptions (Li, Ma, and Zhang 2021; Qu et al. 2016), underpins a broad range of Earth-observation applications—from disaster monitoring and land-use classification to urban planning. The task is inherently cross-modal: given a textual query, one must retrieve the most relevant RS images, and vice versa. Yet several challenges impede progress. First, the visual and textual domains exhibit a pronounced modality gap, exacerbated by the rich, fine-grained semantics typical of overhead scenes (Ouyang et al. 2024; Lin et al. 2025; Ouyang, Ma, and Bai 2025). Second, existing RS datasets are often constructed via automated captioning or legacy metadata pipelines, resulting in incomplete, error-prone, or semantically extraneous descriptions. Consequently, a substantial fraction of image-text pairs are only partially aligned-or outright mismatched-introducing noisy supervision that misleads models trained under the assumption of perfect correspondence. To sustain robust retrieval in the face of such imperfect annotations, it is essential not only to suppress the deleterious effects of pseudo-matched pairs while distilling potentially informative cues from them (Zhang et al. 2024).

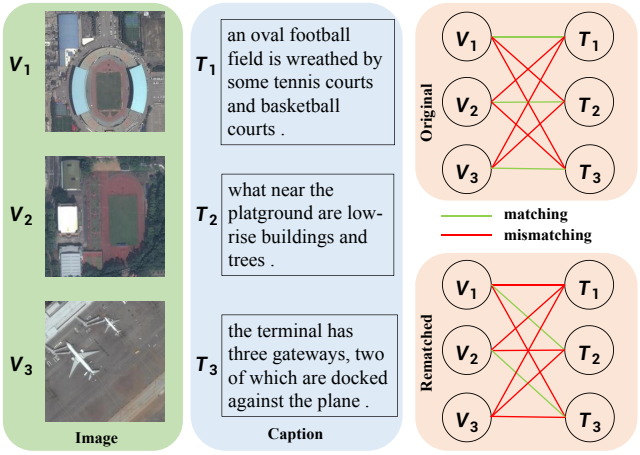

Figure 1: A simple example illustrates our key insight: Pseudo-Matched Pairs (PMPs), image–text samples with partial or incorrect alignment, are not merely noise but often contain latent semantic cues. Instead of discarding them, PMPGuard exploits these cues by rematching semantically relevant pairs (green links) and repelling irrelevant ones (red links), effectively transforming noisy supervision into useful alignment signals.

Most existing methods (Yu and Ji 2022; Wang et al. 2022; Chen et al. 2022) rely on well-annotated image–text pairs to learn effective joint representations. However, in the field of remote sensing, collecting high-quality and precisely aligned image–caption pairs is extremely costly and often infeasible. As a result, many datasets are constructed using web-crawled or automatically generated descriptions, which inevitably introduce Pseudo-Matched Pairs (PMPs)—pairs where the image and caption are only partially aligned or semantically inconsistent. To mitigate this issue, recent studies have attempted to down-weight the influence of PMPs or relax the loss margins in ranking-based objectives. Howeverese, these approaches tend to underexploit the mismatched samples, treating them merely as noise to suppress rather than as potentially informative signals to mine.

As shown in Fig. 1, a simple example illustrates our

*Corresponding Author

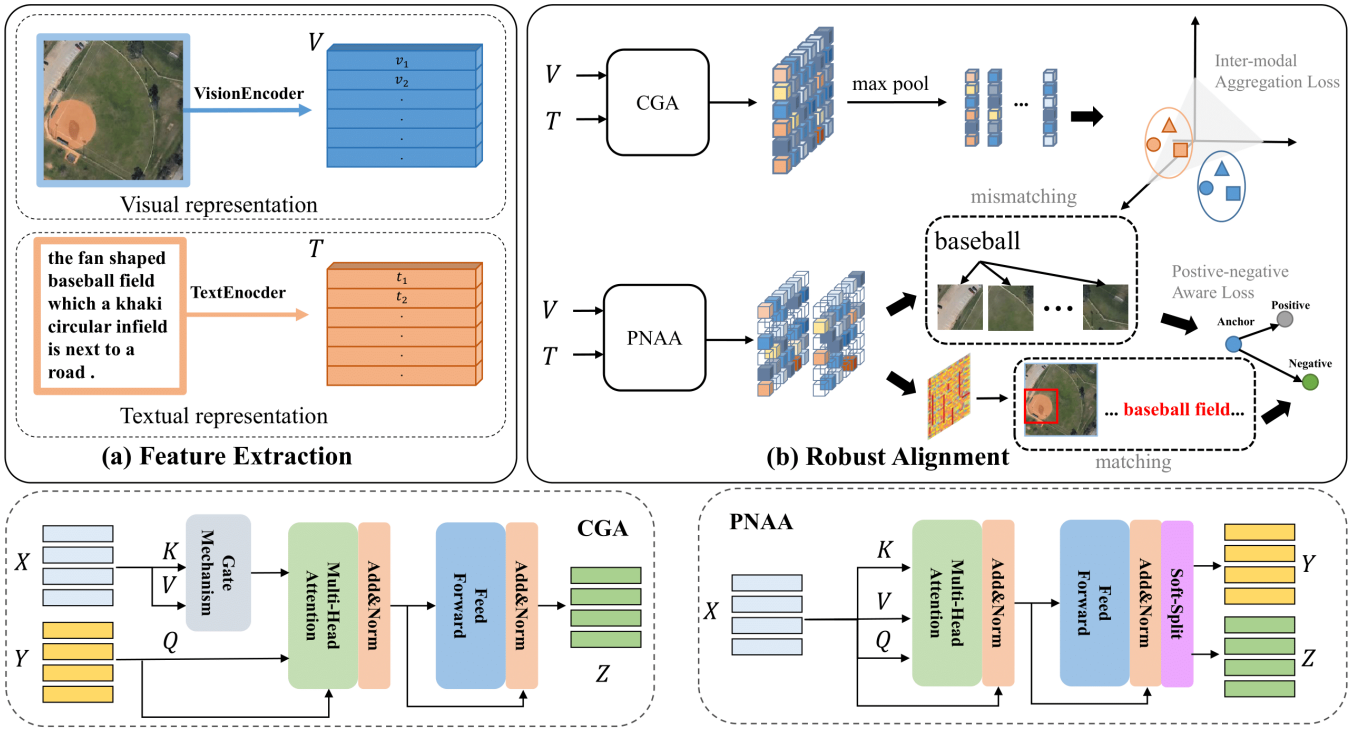

Figure 2: PMPGuard overview: vision and text encoders extract features, then Cross-Gated Attention (CGA) suppresses mismatched cues and Positive–Negative Awareness Attention (PNAA) contrasts reliable/unreliable pairs, jointly optimized to mitigate pseudo-matched pairs and strengthen cross-modal alignment.

core idea: the potential semantic similarity between unpaired samples enables the extraction of valuable knowledge from mismatched pairs. In remote sensing image-text retrieval, such PMPs are common, where image-text samples exhibit partial semantic relevance but are not perfectly aligned. This introduces two major challenges: (1) distinguishing truly mismatched elements from partially relevant ones, and (2) leveraging useful information from mismatched pairs without being misled by noisy signals. To address these issues, we propose **PMPGuard**, a novel framework that enhances cross-modal alignment by explicitly identifying and utilizing latent semantic cues within PMPs. Rather than discarding mismatched pairs as noise, PMPGuard rematches semantically relevant cross-modal samples (green links) and repels irrelevant ones (red links), leading to more robust retrieval performance. Our main contributions are summarized as follows:

- We propose a novel retrieval framework, **PMPGuard**, specifically designed to tackle semantic misalignment and noisy supervision caused by **Pseudo-Matched Pairs (PMPs)**, image–text pairs that are only partially or incorrectly aligned, which are prevalent in large-scale remote sensing datasets. Rather than simply suppressing PMPs, **PMPGuard** identifies and leverages them to strengthen cross-modal alignment.
- We introduce a **Cross-Gated Attention (CGA)** module that adaptively regulates cross-modal feature exchange. CGA selectively promotes semantically consistent content flow while filtering out modality-specific noise and PMP-induced mismatches during training.
- We design a **Positive–Negative Awareness Attention (PNAA)** module that explicitly distinguishes between aligned and misaligned regions in image–text pairs. By modeling both positive and negative cues via a dual-branch structure, PNAA improves robustness against noisy supervision from PMPs.
- Extensive experiments on three public remote sensing benchmarks, i.e., **RSICD**, **RSITMD**, and **RS5M**, demonstrate that **PMPGuard** consistently achieves **state-of-the-art (SOTA)** performance. Notably, its advantage is more evident under higher mismatch rates, confirming its robustness and generalization ability in the presence of PMPs.

Related Work

Image-Text Retrieval in Remote Sensing

Image-text retrieval in remote sensing (RS) seeks to align aerial imagery with natural language, enabling applications such as land cover analysis and disaster response. Recent years have witnessed a shift from traditional CNN-RNN frameworks to transformer-based models with more fine-grained alignment capabilities. For instance, PIR (Pan, Ma, and Bai 2023a) incorporates prior knowledge into the retrieval process, while SWAN (Pan, Ma, and Bai 2023b) introduces scene-aware aggregation to reduce semantic con-

fusion. DOVE (Ma, Pan, and Bai 2024) enhances directional alignment using visual-semantic embedding. These methods typically assume well-aligned training pairs, which limits their robustness in noisy real-world datasets. To address imperfect supervision, recent works like BiCro (Yang et al. 2023) and DEL (Feng et al. 2024) propose noise-aware training strategies. However, they mostly suppress mismatches instead of leveraging them. Our approach differs by actively mining mismatched pairs through gated attention and positive-negative awareness, enabling more robust alignment under noisy correspondence.

Mismatch-Robust Retrieval Models

Mismatch-robust retrieval focuses on learning reliable cross-modal alignments when training data contains noisy or pseudo-matched image-text pairs. This is especially relevant for remote sensing datasets, where large-scale annotation is often noisy or weakly supervised. DEL (Feng et al. 2024) introduces evidential learning to estimate uncertainty and reduce the influence of incorrect pairs. BiCro (Yang et al. 2023) enforces bi-directional similarity consistency to rectify noisy correspondences. L2RM (Han et al. 2024) explicitly rematches mismatched pairs during training using a joint re-alignment strategy. Although effective, most existing methods treat mismatched pairs as noise to be suppressed. In contrast, our approach not only identifies mismatches but also exploits their latent semantic relations through gated attention and discriminative mining, enabling more robust and informative alignment (Cheng et al. 2021).

Approach

Problem Definition

Without loss of generality, we take visual–text retrieval as an illustrative task to formalize the PMP (Pseudo-Matched Pair) problem in cross-modal retrieval. Let the training set be

$$\mathcal{D} = \{(V_i, T_i, m_i)\}_{i=1}^{N},$$

where (V_i, T_i) denotes a visual–text pair and $m_i \in \{0, 1\}$ indicates whether the pair is semantically matched.

The crux of cross-modal retrieval lies in measuring similarity across heterogeneous modalities. Existing approaches first project visual and textual inputs into a shared embedding space via modality-specific encoders f_v and f_t, respectively, and then compute the similarity of a pair (V_i, T_j) as

$$S_{ij} = g\big(f_v(V_i),\, f_t(T_j)\big),$$

where g is either a parametric or a non-parametric mapping.

Most prior methods tacitly assume that every pair labeled with $m_i = 1$ is indeed matched. In practice, however, real-world remote-sensing datasets—compiled through automated captioning pipelines, crowd-sourcing, or legacy metadata—frequently exhibit a non-negligible proportion of mismatched visual–text pairs that are erroneously annotated as positive. These pseudo-matched pairs (PMPs) propagate noisy supervision signals and compromise model robustness. Our objective is therefore to detect and suppress the influence of PMPs, thereby enabling robust cross-modal retrieval for remote-sensing data.

Cross-Gated Attention Mechanism

Addressing the noisy feature interaction problem caused by PMPs (as noted in Introduction), our Cross-Gated Attention Mechanism dynamically regulates information flow. Specifically, it solves two subproblems: (1) suppressing irrelevant features that may arise from mismatched pairs, while (2) preserving potentially useful cross-modal interactions.

Given a textual feature sequence $\mathcal{U} = \{u_i\}_{i=1}^{m}$ and a set of image region features $\mathcal{V} = \{v_j\}_{j=1}^{n}$, the goal is to compute modality-aware representations through mutual attention and gating operations.

Cross Attention. We first compute the attention score between each word u_i and image region v_j as:

$$a_{ij} = \frac{u_i^\top W_a v_j}{\|u_i\|\|v_j\|}, \tag{1}$$

where W_a is a learnable projection matrix. The attention weights are then normalized using softmax:

$$\alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{j'=1}^{n}\exp(a_{ij'})}, \beta_{ji} = \frac{\exp(a_{ij})}{\sum_{i'=1}^{m}\exp(a_{i'j})}. \tag{2}$$

Using these attention weights, we generate cross-modal attended features:

$$\tilde{v}_i = \sum_{j=1}^{n}\alpha_{ij}v_j, \tilde{u}_j = \sum_{i=1}^{m}\beta_{ji}u_i. \tag{3}$$

Gating Mechanism. To suppress irrelevant or noisy features and highlight important interactions, we introduce gating functions that control how much information is retained from each modality. The gated representations are computed as:

$$g_i^u = \sigma(W_g^u[\mathbf{u}_i; \tilde{\mathbf{v}}_i] + b_g^u), g_j^v = \sigma(W_g^v[\mathbf{v}_j; \tilde{\mathbf{u}}_j] + b_g^v), \tag{4}$$

where $[\cdot;\cdot]$ denotes concatenation, σ is the sigmoid activation, and W_g^u, W_g^v and b_g^u, b_g^v are learnable parameters. $\mathbf{u}_i$ and $\mathbf{v}_j$ are the original modality-specific embeddings, $\tilde{\mathbf{v}}_i$ and $\tilde{\mathbf{u}}_j$ are the cross-modal contexts, and $g_i^u, g_j^v \in [0,1]^d$ are the learned gating vectors.

The final cross-gated features are:

$$\hat{\mathbf{u}}_i = g_i^u \odot \mathbf{u}_i + (1 - g_i^u) \odot \tilde{\mathbf{v}}_i, \hat{\mathbf{v}}_j = g_j^v \odot \mathbf{v}_j + (1 - g_j^v) \odot \tilde{\mathbf{u}}_j, \tag{5}$$

where $\odot$ denotes element-wise multiplication. This mechanism allows the model to selectively combine intra- and inter-modal features, yielding enhanced representations for image-text matching.

Inter-modal Aggregation Loss To align the vision and language representations after the Cross-Gated Attention (CGA) refinement, we introduce the Inter-modal Aggregation Loss (IA). Let $\hat{\mathbf{u}}_i$ and $\hat{\mathbf{v}}_j$ denote the gated visual and textual features. The loss is formulated as an InfoNCE objective with a temperature parameter τ:

$$\mathcal{L}_{\text{IA}} = -\sum_{(i,j)\in\mathcal{P}} \log \frac{\exp\left(\frac{\hat{\mathbf{u}}_i^\top \hat{\mathbf{v}}_j}{\tau}\right)}{\sum_k \exp\left(\frac{\hat{\mathbf{u}}_i^\top \hat{\mathbf{v}}_k}{\tau}\right) + \sum_k \exp\left(\frac{\hat{\mathbf{u}}_k^\top \hat{\mathbf{v}}_j}{\tau}\right)}, \tag{6}$$

where $\mathcal{P}$ denotes the set of semantically matched pairs. By maximizing the agreement between paired samples while pushing non-pairs apart, this loss strengthens robust cross-modal alignment and alleviates the adverse impact of pseudo-matched pairs.

Positive-Negative Awareness Attention

To tackle the PMPs' dual nature (both misleading and potentially useful), PNAA explicitly models positive and negative signals through: (1) a negative branch that identifies and suppresses truly mismatched fragments (Eq.12-13), and (2) a positive branch that extracts useful semantic cues from partially matched regions (Eq.14-15). This dual strategy directly addresses the challenge of utilizing imperfect training pairs raised in our problem statement.

Catching Pseudo-Matched Pairs To distinguish matched and mismatched image–text pairs, GeoRSCLIP (Zhang et al. 2024) models their similarity distributions as Gaussian functions:

$$f_k^+(s) = \frac{1}{\sigma_k^+\sqrt{2\pi}} \exp\left(-\frac{(s-\mu_k^+)^2}{2(\sigma_k^+)^2}\right), \tag{7}$$

$$f_k^-(s) = \frac{1}{\sigma_k^-\sqrt{2\pi}} \exp\left(-\frac{(s-\mu_k^-)^2}{2(\sigma_k^-)^2}\right), \tag{8}$$

where μ_k^+, σ_k^+ and μ_k^-, σ_k^- denote the means and standard deviations of similarity scores for matched and mismatched pairs, respectively.

To automatically separate the two distributions, GeoRSCLIP learns a decision boundary t_k by minimizing the weighted overlap between the two distributions:

$$t_k = \arg\min_{t\geq 0} \left[\alpha \int_t^{+\infty} f_k^-(s)\, ds + \int_{-\infty}^{t} f_k^+(s)\, ds\right], \tag{9}$$

where α is a penalty parameter balancing false positives and false negatives. The learned boundary t_k is then used to separate positive and negative fragments during attention calculation, enabling more robust alignment under noisy supervision.

Negative Branch. For each word u_i, we compute its maximum similarity with all image regions:

$$s_i = \max_j \left(\cos(u_i, v_j) - t_k\right), \tag{10}$$

and apply a negative mask to suppress matched words:

$$s_i^{\text{neg}} = s_i \cdot \text{Mask}_{\text{neg}}(s_i), \tag{11}$$

where $\text{Mask}_{\text{neg}}(s_i) = 1$ if $s_i < 0$, and 0 otherwise.

Positive Branch. We compute inter-modal attention weights:

$$w_{ij}^{\text{inter}} = \text{softmax}\left(\text{Mask}_{\text{pos}}(s_{ij} - t_k)\right), \tag{12}$$

where $\text{Mask}_{\text{neg}}(s_i) = 1$ if $s_i > 0$, and 0 otherwise. And aggregate matched image features to get the positive score:

$$s_i^{\text{pos}} = \cos(u_i, \hat{v}_i) + \sum_j w_{ij}^{\text{relev}} s_{ij}, \tag{13}$$

where $\hat{v}_i = \sum_j w_{ij}^{\text{inter}} v_j$, and w_{ij}^{relev} is a relevance-based attention weight.

Positive-negative Aware Loss The final similarity score is computed as:

$$S(\mathcal{V},\mathcal{U}) = \frac{1}{m}\sum_{i=1}^{m}\left(s_i^{\text{pos}} + s_i^{\text{neg}}\right). \tag{14}$$

We adopt the bidirectional triplet ranking loss for training:

$$\begin{aligned}\mathcal{L}_{\mathcal{PA}} = &\sum_{(\mathcal{U},\mathcal{V})} \left[\gamma - S(\mathcal{U},\mathcal{V}) + S(\mathcal{U},\mathcal{V}')\right]_+ \\ &+ \left[\gamma - S(\mathcal{U},\mathcal{V}) + S(\mathcal{U}',\mathcal{V})\right]_+,\end{aligned} \tag{15}$$

where γ is a margin hyperparameter, and $[\cdot]_+$ denotes the hinge loss.

The Training Objective

The final training objective is the joint minimization of the Inter-modal Aggregation Loss and the Positive–Negative Awareness Loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{IA}} + \lambda\, \mathcal{L}_{\text{PA}}, \tag{16}$$

where $\lambda > 0$ balances the two terms. Minimizing $\mathcal{L}_{\text{total}}$ simultaneously aligns cross-modal representations via the InfoNCE-based $\mathcal{L}_{\text{IA}}$ and suppresses the influence of noisy or mismatched pairs via the margin-based $\mathcal{L}_{\text{PA}}$, yielding robust retrieval under imperfect supervision.

Experiments

Datasets

RSICD RSICD (Lu et al. 2018) is a dedicated remote-sensing image-captioning benchmark that comprises 10,921 geo-diverse images—drawn from Google Earth and Street View—each paired with five concise English captions. Spanning urban, rural, agricultural, mountainous, and aquatic scenes, the dataset has been uniformly resized to 224×224 pixels to facilitate consistent processing.

RSITMD RSITMD (Yuan et al. 2022) is a fine-grained benchmark for cross-modal remote-sensing image retrieval. It comprises 4,743 images paired with 23,715 diverse, low-redundancy captions and 1–5 precise keywords per image, yielding richer object-level descriptions. Designed to minimize intra-class similarity and amplify inter-class diversity, RSITMD delivers superior retrieval accuracy compared with existing datasets.

MRate	Method	RSICD							RSITMD						
		Sentence Retrieval			Image Retrieval			mR	Sentence Retrieval			Image Retrieval			mR
		R1	R5	R10	R1	R5	R10		R1	R5	R10	R1	R5	R10	
0	L2RM	5.01	13.19	21.84	4.32	15.21	26.30	14.15	7.37	23.07	35.73	7.42	29.76	49.77	25.45
	HarMA-Vit	1.49	3.83	7.53	1.15	3.47	6.80	3.97	2.89	4.82	9.67	1.61	4.61	9.63	5.57
	PIR	8.61	23.57	35.40	5.49	21.84	38.04	21.99	11.14	33.14	45.21	8.60	33.02	56.43	30.99
	SWAN	9.82	24.61	36.25	6.70	22.51	38.67	23.48	11.91	34.29	46.16	9.50	33.66	57.14	32.63
	DOVE	10.54	25.31	38.27	7.93	23.54	40.13	23.79	13.51	35.25	48.12	10.21	34.52	58.45	33.29
	PMPGuard	**12.56**	**28.20**	**39.37**	**9.76**	**25.55**	**42.49**	**26.21**	**14.94**	**37.39**	**49.11**	**12.58**	**37.06**	**59.01**	**34.95**
0.2	L2RM	4.03	12.35	21.04	3.59	14.49	25.51	13.50	6.64	22.35	34.96	6.86	28.94	48.94	24.78
	HarMA-Vit	0.82	3.29	6.50	0.49	2.71	5.65	3.24	2.21	3.98	8.63	0.88	3.76	8.89	4.73
	PIR	7.78	22.69	34.40	4.85	20.70	36.94	21.23	10.4	32.08	44.25	7.65	32.08	55.31	30.29
	SWAN	8.95	23.72	35.07	5.85	21.66	37.91	22.66	11.25	33.25	45.06	8.85	32.85	56.04	31.87
	DOVE	9.88	24.25	37.34	7.04	22.82	39.43	23.11	12.57	34.55	47.01	9.61	33.66	57.46	32.48
	PMPGuard	**11.90**	**27.26**	**38.20**	**9.01**	**24.42**	**41.67**	**25.42**	**14.06**	**36.64**	**48.46**	**11.87**	**36.09**	**57.88**	**34.07**
0.4	L2RM	2.84	9.06	17.11	2.73	11.40	20.55	10.61	5.09	17.92	30.09	4.47	20.13	35.49	18.86
	HarMA-Vit	1.01	3.20	6.86	0.59	1.99	4.96	3.10	1.77	5.31	7.96	0.93	3.98	7.30	4.54
	PIR	6.22	17.38	29.37	4.41	18.50	32.81	18.12	9.51	26.33	40.27	7.08	29.56	51.59	27.39
	SWAN	7.44	22.31	33.60	4.41	20.29	36.50	21.19	9.82	31.77	43.69	7.33	31.36	54.47	30.26
	DOVE	8.35	22.95	35.96	5.48	21.31	37.86	21.77	11.06	33.17	45.60	8.21	32.18	56.01	30.91
	PMPGuard	**10.38**	**25.77**	**36.90**	**7.44**	**22.92**	**40.18**	**24.03**	**12.61**	**35.29**	**47.25**	**10.41**	**34.63**	**56.62**	**32.58**
0.6	L2RM	0.88	2.21	3.32	0.35	2.35	4.56	2.28	3.54	15.04	25.44	4.20	18.10	29.47	15.97
	HarMA-Vit	0.64	2.84	5.40	0.60	2.51	4.56	2.76	1.11	5.75	7.74	0.40	3.50	6.33	4.14
	PIR	4.67	13.72	25.07	2.95	14.82	26.55	14.63	9.96	27.88	38.72	6.46	26.28	48.36	26.28
	SWAN	6.59	21.61	32.44	3.63	19.51	35.58	20.34	8.94	31.21	43.22	6.59	30.80	53.65	29.31
	DOVE	7.49	22.30	34.80	4.64	20.48	36.86	20.97	10.28	32.66	44.93	7.53	31.51	55.11	30.34
	PMPGuard	**9.51**	**24.94**	**35.70**	**6.66**	**22.05**	**39.23**	**23.03**	**11.71**	**34.74**	**46.61**	**9.58**	**33.87**	**55.83**	**31.54**
0.8	L2RM	0.22	1.55	3.32	0.58	2.35	4.56	2.49	3.54	11.28	17.92	1.86	7.88	14.56	9.51
	HarMA-Vit	0.27	1.92	3.66	0.40	1.70	3.17	1.85	0.88	3.32	4.87	0.35	2.04	4.96	2.74
	PIR	3.66	11.16	21.04	2.34	11.4	21.67	11.88	4.65	17.04	27.43	3.63	15.44	28.98	16.19
	SWAN	4.61	19.64	31.30	1.69	17.55	33.71	18.53	7.95	29.97	41.88	4.55	28.91	51.86	27.41
	DOVE	5.57	20.24	33.57	2.75	18.56	35.02	19.01	8.38	30.91	43.94	5.53	29.65	53.97	28.68
	PMPGuard	**7.48**	**22.80**	**34.40**	**4.74**	**20.15**	**37.30**	**21.12**	**10.01**	**32.90**	**44.97**	**7.48**	**31.72**	**54.77**	**29.64**

Table 1: Image-text retrieval performance under different MRate on RSICD and RSITMD.

RS5M RS5M (Zhang et al. 2024) is a five-million pair remote sensing vision language corpus whose captions were automatically scraped from open platforms. The sheer scale comes at the cost of pervasive pseudo-matched pairs (PMPs): brief, template-like, or outright irrelevant text routinely accompanies images of agriculture, infrastructure, or land-use, yielding noisy, weak, and geographically skewed alignments. A quality-scored subset reveals that up to 30% of pairs exhibit significant mismatch, making RS5M an ideal testbed for studying retrieval robustness under realistic, imperfect supervision.

Implementation Details

All experiments in this study were conducted using two NVIDIA RTX A6000 GPUs. For consistency across different datasets with varying image sizes, we uniformly resized them to 224 × 224 pixels and input them into the network. Data augmentation techniques, including rotation and flip, were applied to enhance model robustness. The word vector's representation dimension was 768, while the embedding space for images and text was 512. The parameter τ was initialized to 0.07 and learned during contrast loss calculation, and a 0.5 margin was enforced for the triplet loss calculation. The network was trained for 20 epochs using the AdamW optimizer, which combined both contrast loss and triplet loss, with a batch size of 128 per GPU. To ensure the robustness of the experimental results, each set of experiments was conducted five times, and the average result was recorded.

Baselines

We compare PMPGuard with five state-of-the-art cross-modal retrieval methods, including five general methods: PIR (Pan, Ma, and Bai 2023a), SWAN (Pan, Ma, and Bai 2023b), HarMA (Huang 2024), DOVE (Ma, Pan, and Bai 2024), Furthermore, L2RM (Han et al. 2024). L2RM and HarMa using GeoRSCLIP (Zhang et al. 2024) as the backbone are treated as large-parameter models.

Main Results

Construction of Pseudo-matched Training Data To evaluate robustness under imperfect supervision, we sim-

MRate	Method	RSICD							RSITMD						
		Sentence Retrieval			Image Retrieval			mR	Sentence Retrieval			Image Retrieval			mR
		R1	R5	R10	R1	R5	R10		R1	R5	R10	R1	R5	R10	
0	L2RM-Geo	18.84	42.30	53.65	14.86	38.60	55.55	37.42	27.98	52.80	62.83	22.95	55.20	72.34	48.71
	HarMA-Geo	19.51	43.72	55.68	15.41	40.06	57.73	38.90	29.04	54.56	65.03	23.82	56.95	74.61	50.39
	PMPGuard-Geo	**20.21**	**45.05**	**57.28**	**15.93**	**41.30**	**59.39**	**40.23**	**30.12**	**56.21**	**67.00**	**24.55**	**58.75**	**76.84**	**52.01**
0.2	L2RM-Geo	17.75	40.65	52.55	13.68	37.50	53.82	36.10	27.00	51.05	61.38	21.85	53.40	70.40	47.25
	HarMA-Geo	18.39	42.36	54.53	14.29	38.96	56.25	37.46	28.10	53.10	63.72	22.70	55.62	73.58	49.47
	PMPGuard-Geo	**19.02**	**43.85**	**56.30**	**14.85**	**40.32**	**58.11**	**38.75**	**29.15**	**54.86**	**66.10**	**23.50**	**57.48**	**76.00**	**51.10**
0.4	L2RM-Geo	17.44	38.76	51.80	12.92	36.85	52.28	34.95	26.75	50.10	61.92	21.78	53.25	70.50	47.30
	HarMA-Geo	18.12	40.35	53.98	13.50	38.19	54.47	36.43	27.88	52.21	64.38	22.74	55.53	73.72	49.41
	PMPGuard-Geo	**18.75**	**41.96**	**56.18**	**14.02**	**39.50**	**56.86**	**37.70**	**28.95**	**54.10**	**66.90**	**23.45**	**57.30**	**76.05**	**51.12**
0.6	L2RM-Geo	16.45	38.45	51.47	11.75	33.90	50.90	33.72	25.40	48.70	61.80	19.85	51.20	67.25	45.65
	HarMA-Geo	17.11	40.07	53.71	12.28	35.24	53.14	35.26	26.55	50.88	64.16	20.75	53.50	70.31	47.69
	PMPGuard-Geo	**17.68**	**41.45**	**55.60**	**12.78**	**36.55**	**55.35**	**36.50**	**27.58**	**52.80**	**66.50**	**21.50**	**55.25**	**73.00**	**49.35**
0.8	L2RM-Geo	17.54	36.33	48.33	11.56	34.27	50.50	33.06	25.01	47.08	60.33	19.50	49.65	66.38	44.66
	HarMA-Geo	18.30	38.24	50.87	12.17	36.07	53.16	34.80	26.33	49.56	63.50	20.53	52.26	69.87	47.01
	PMPGuard-Geo	**19.07**	**40.15**	**53.41**	**12.78**	**37.87**	**55.82**	**36.54**	**27.65**	**52.04**	**66.68**	**21.56**	**54.87**	**73.36**	**49.36**

Table 2: The image-text retrieval performance of the large model GeoRSCLIP as the backbone network under different mismatch rates (MRate) on RSICD and RSITMD.

Method	RS5M						
	Sentence Retrieval			Image Retrieval			mR
	R1	R5	R10	R1	R5	R10	
L2RM-Geo	13.58	34.42	45.21	9.84	31.26	46.90	30.42
HarMA-Geo	14.84	37.13	48.39	10.96	33.74	50.12	32.78
PMPGuard-Geo	**16.12**	**39.80**	**51.60**	**12.10**	**36.05**	**53.43**	**35.21**

Table 3: Image-text retrieval performance on RS5M.

ulate realistic noise via pseudo-matched training sets with controlled mismatches. For a clean dataset $\mathcal{D} = (V_i, T_i)_{i=1}^{N}$, we generate $\mathcal{D}_x$ by replacing T_i in $x\%$ of pairs with a random T_j ($j \neq i$). To ensure a true semantic mismatch, we use the pretrained **GeoRSCLIP** model to compute similarity scores and retain only replacements with scores below a threshold τ. This filtering prevents trivial or semantically similar substitutions, increasing the realism of the noise. We construct multiple variants with different mismatch rates ($x = 0\%, 20\%, \ldots, 80\%$) to benchmark retrieval models under varying noise levels.

Results on Synthesized PMPs Table 1 presents the experimental results on RSICD and RSITMD, where our model adopts Swin-Transformer (Liu et al. 2021) as the image backbone and BERT (Devlin et al. 2019) as the text backbone. From the results, we observe that PMPGuard achieves the best performance across all metrics compared to other state-of-the-art methods, demonstrating its strong capability for robust cross-modal retrieval. Moreover, when the mismatch ratio is higher, such as 0.6 and 0.8, the improvements brought about by PMPGuard become even more pronounced, validating the effectiveness of mismatched pair mining to enhance retrieval robustness.

Evaluation Results Based on the Large Pretrained Backbone Model Table 2 presents the performance of different methods under varying mismatch rates (MRate) on the RSICD and RSITMD datasets. In general, our proposed method, **PMPGuard-Geo**, consistently outperforms baselines (L2RM-Geo and HarMA-Geo) in all settings and evaluation metrics. In RSICD, PMPGuard-Geo achieves the highest mR at each mismatch level, with a notable improvement from 37.42 (L2RM-Geo) and 38.90 (HarMA-Geo) to 40.23 at MRate = 0. Even as the mismatch rate increases to 0.8, PMPGuard-Geo maintains strong performance (mR = 36.54), demonstrating superior robustness to pseudo-matched pairs. On RSITMD, PMPGuard-Geo exhibits a similar advantage, consistently achieving the best R@1, R@5, R@10, and mR values. For example, at MRate = 0.8, PMPGuard-Geo obtains an mR of 49.36, significantly surpassing L2RM-Geo (44.66) and HarMA-Geo (47.01). These results validate that PMPGuard-Geo not only captures fine-grained cross-modal semantics more effectively but is also more resilient to noisy supervision, making it well-suited for real remote sensing image-text retrieval tasks.

Results on Real PMPs Our key findings highlight the effectiveness of PMPGuard-Geo in addressing pseudo-matched pairs (PMPs) in remote sensing image-text retrieval. As shown in Table 3, PMPGuard-Geo achieves best-in-class retrieval performance, with substantial gains in R1 scores for both sentence and image retrieval, demonstrating strong precision and the ability to suppress noise via Cross-Gated Attention (CGA). Consistent improvements in R@10 further validate Positive-Negative Awareness Attention (PNAA), which effectively distinguishes latent semantic cues from noise. In particular, PMPGuard-Geo achieves a 4.79% improvement in mean recall (mR), confirming its robustness in the retrieval directions and thresholds. Furthermore, the performance gap in the RS5M data set,

MRate	Method	RSICD							RSITMD						
		Sentence Retrieval			Image Retrieval			mR	Sentence Retrieval			Image Retrieval			mR
		R1	R5	R10	R1	R5	R10		R1	R5	R10	R1	R5	R10	
0	w/o CGA	18.32	41.65	52.70	14.22	37.45	54.80	36.52	27.15	51.90	61.30	22.10	54.30	71.40	47.53
	w/o PNAA	19.35	43.20	55.12	15.12	39.65	56.90	38.56	28.45	53.95	64.20	23.35	56.40	73.95	49.38
	full	**20.21**	**45.05**	**57.28**	**15.93**	**41.30**	**59.39**	**40.23**	**30.12**	**56.21**	**67.00**	**24.55**	**58.75**	**76.84**	**52.01**
0.2	w/o CGA	17.25	39.45	50.83	13.10	36.40	52.70	34.96	26.10	49.90	60.20	20.70	52.30	69.20	46.57
	w/o PNAA	18.15	41.30	53.20	13.85	38.05	54.85	36.90	27.30	52.40	62.90	22.05	54.90	72.60	48.69
	full	**19.02**	**43.85**	**56.30**	**14.85**	**40.32**	**58.11**	**38.75**	**29.15**	**54.86**	**66.10**	**23.50**	**57.48**	**76.00**	**51.10**
0.4	w/o CGA	16.85	37.80	49.65	12.50	35.50	51.40	33.45	25.60	48.80	59.80	20.35	50.80	67.90	45.54
	w/o PNAA	17.76	39.90	52.30	13.20	37.10	53.20	35.91	26.65	51.30	63.30	21.45	53.30	71.45	48.24
	full	**18.75**	**41.96**	**56.18**	**14.02**	**39.50**	**56.86**	**37.70**	**28.95**	**54.10**	**66.90**	**23.45**	**57.30**	**76.05**	**51.12**
0.6	w/o CGA	15.90	36.40	48.75	11.30	32.80	49.85	32.50	24.15	47.10	59.60	18.80	49.85	65.20	44.12
	w/o PNAA	16.68	38.20	51.10	12.05	34.40	51.70	34.69	25.65	49.85	62.30	20.00	52.40	68.95	46.52
	full	**17.68**	**41.45**	**55.60**	**12.78**	**36.55**	**55.35**	**36.50**	**27.58**	**52.80**	**66.50**	**21.50**	**55.25**	**73.00**	**49.35**
0.8	w/o CGA	15.40	34.50	46.00	10.75	31.40	48.10	31.36	23.45	45.85	58.00	18.10	48.05	63.30	42.79
	w/o PNAA	16.88	36.80	49.45	11.80	33.85	50.65	33.57	24.95	48.60	61.30	19.30	51.20	66.75	45.02
	full	**19.07**	**40.15**	**53.41**	**12.78**	**37.87**	**55.82**	**36.54**	**27.65**	**52.04**	**66.68**	**21.56**	**54.87**	**73.36**	**49.36**

Table 4: Ablation experiments of the PMPGuard model under different mismatch rates (MRate) on RSICD and RSITMD.

compared to synthetic noise baselines, demonstrates PMPGuard's unique ability to handle real-world PMP challenges, including natural alignment inconsistencies, geographic variation, and large-scale noisy supervision. These results confirm that explicitly modeling positive-negative relationships and adaptively gating cross-modal signals transform PMPs from learning obstacles into valuable supervision.

Ablation Experiment

As shown in Table 4, we performed ablation studies with varying MRate in the RSICD and RSITMD datasets to evaluate the effectiveness of each component in the proposed PMPGuard-Geo framework. We examine the performance of variants without the CGA or PNAA modules (denoted as "w/o CGA" and "w/o PNAA") and compare them with the full model. Removing either CGA or PNAA results in a consistent performance drop across all metrics and mismatch levels. In particular, excluding PNAA causes a sharper decline at higher mismatch rates: for example, mean recall (mR) on RSITMD drops from 49.36 to 45.02 at MRate = 0.8, demonstrating PNAA's effectiveness in handling noisy supervision. Omitting **CGA** broadly reduces recall, particularly in sentence retrieval, indicating its role in enhancing cross-modal alignment. Overall, **CGA** and **PNAA** play complementary roles: CGA strengthens semantic correspondence, while PNAA improves robustness to noisy or weakly aligned pairs, jointly enabling PMPGuard-Geo to maintain

Visualization Experiment

Fig. 3 visualizes PMPGuard's ability to correct mismatched remote sensing (RS) image-text pairs. Three initial incorrect caption-image match cases are presented, where PMPGuard successfully rematches each image with more appropriate descriptions—green highlights correct matches, and red denotes original mismatches. This demonstrates PMPGuard's

Figure 3: PMPGuard rematches mismatched RS image-text pairs; green and red words denote correct and incorrect matches, respectively.

effectiveness in enhancing RS image-text semantic alignment and cross-modal retrieval accuracy, a key capability for real-world applications requiring precise image understanding.

Conclusion

We present **PMPGuard**, a robust remote-sensing (RS) image–text retrieval framework addressing pseudo-matched pairs (PMPs) via two core innovations: **Cross-Gated Attention (CGA)** (dynamic cross-modal feature modulation) and **Positive–Negative Awareness Attention (PNAA)** (discriminative alignment learning). Extensive experiments on RSICD, RSITMD, and RS5M demonstrate its superior performance across 0–0.8 mismatch rates, achieving SOTA results with robustness to noisy correspondences, providing a principled solution for real-world imperfect alignment retrieval scenarios and laying a foundation for robust cross-modal RS learning.

Acknowledgments

This work is partially supported by the Zhejiang Provincial Natural Science Foundation of China under Grants No. LY24F020020 and No. LRG25F020002 and the Natural Science Foundation of China under Grants No. U20A20196 and No. 62302453.

References

Chen, J.; Huang, H.; Peng, J.; Zhu, J.; Chen, L.; Tao, C.; and Li, H. 2022. Contextual Information-Preserved Architecture Learning for Remote-Sensing Scene Classification. *IEEE Trans. Geosci. Remote. Sens.*, 60: 1–14.

Chen, Y.; Du, C.; Zi, Y.; Xiong, S.; and Lu, X. 2024a. Scale-aware adaptive refinement and cross interaction for remote sensing audio-visual cross-modal retrieval. *IEEE Transactions on Geoscience and Remote Sensing*.

Chen, Y.; Huang, J.; Li, X.; Xiong, S.; and Lu, X. 2023. Multiscale salient alignment learning for remote-sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13.

Chen, Y.; Huang, J.; Xiong, S.; and Lu, X. 2024b. Integrating multisubspace joint learning with multilevel guidance for cross-modal retrieval of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.

Cheng, Q.; Zhou, Y.; Fu, P.; Xu, Y.; and Zhang, L. 2021. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 4284–4297.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Elman, J. L. 1990. Finding Structure in Time. *Cogn. Sci.*, 14(2): 179–211.

Faes, M. G.; Broggi, M.; Spanos, P. D.; and Beer, M. 2022. Elucidating appealing features of differentiable autocorrelation functions: A study on the modified exponential kernel. *Probabilistic Engineering Mechanics*, 69: 103269.

Feng, Z.; Zeng, Z.; Guo, C.; Li, Z.; and Hu, L. 2024. Learning From Noisy Correspondence With Tri-Partition for Cross-Modal Matching. *IEEE Trans. Multim.*, 26: 3884–3896.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, 1735–1742. IEEE.

Han, H.; Zheng, Q.; Dai, G.; Luo, M.; and Wang, J. 2024. Learning to Rematch Mismatched Pairs for Robust Cross-Modal Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 26669–26678. IEEE.

He, Y.; Xu, X.; Chen, H.; Li, J.; and Pu, F. 2024. Visual Global-Salient Guided Network for Remote Sensing Image-Text Retrieval. *IEEE Transactions on Geoscience and Remote Sensing*.

Hu, P.; Huang, Z.; Peng, D.; Wang, X.; and Peng, X. 2023. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9595–9610.

Huan, R.; Zhong, G.; Chen, P.; and Liang, R. 2024a. TriSAT: Trimodal Representation Learning for Multimodal Sentiment Analysis. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32: 4105–4120.

Huan, R.; Zhong, G.; Chen, P.; and Liang, R. 2024b. UniMF: A Unified Multimodal Framework for Multimodal Sentiment Analysis in Missing Modalities and Unaligned Multimodal Sequences. *IEEE Trans. Multim.*, 26: 5753–5768.

Huan, R.; Zhong, G.; Chen, P.; and Liang, R. 2025. MulDeF: A Model-Agnostic Debiasing Framework for Robust Multimodal Sentiment Analysis. *IEEE Trans. Multim.*, 27: 2304–2319.

Huang, T. 2024. Efficient Remote Sensing with Harmonized Transfer Learning and Modality Alignment. *CoRR*, abs/2404.18253.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 1746–1751. ACL.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Wang, N.; Zhang, L.; Du, B.; and Tao, D. 2020a. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7760–7768.

Li, Y.; Ma, J.; and Zhang, Y. 2021. Image retrieval from remote sensing big data: A survey. *Inf. Fusion*, 67: 94–115.

Li, Y.; Song, L.; Chen, Y.; Li, Z.; Zhang, X.; Wang, X.; and Sun, J. 2020b. Learning dynamic routing for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8553–8562.

Lin, J.; Rao, Y.; Lu, J.; and Zhou, J. 2017. Runtime neural pruning. *Advances in neural information processing systems*, 30.

Lin, Z.; Wang, Z.; Qian, T.; Mu, P.; Chan, S.; and Bai, C. 2025. NeighborRetr: Balancing Hub Centrality in Cross-Modal Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 9263–9273. Computer Vision Foundation / IEEE.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision

Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 9992–10002. IEEE.

Lu, X.; Wang, B.; Zheng, X.; and Li, X. 2018. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote. Sens.*, 56(4): 2183–2195.

Ma, Q.; Pan, J.; and Bai, C. 2024. Direction-oriented visual-semantic embedding model for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*.

Ouyang, P.; Chen, J.; Ma, Q.; Wang, Z.; and Bai, C. 2024. Distinguishing Visually Similar Images: Triplet Contrastive Learning Framework for Image-text Retrieval. In *IEEE International Conference on Multimedia and Expo, ICME 2024, Niagara Falls, ON, Canada, July 15-19, 2024*, 1–6. IEEE.

Ouyang, P.; Ma, Q.; and Bai, C. 2025. Sparse Information Perception Network for Remote Sensing Cross-Modal Retrieval. *IEEE Trans. Geosci. Remote. Sens.*, 63: 1–15.

Pan, J.; Ma, Q.; and Bai, C. 2023a. A Prior Instruction Representation Framework for Remote Sensing Image-text Retrieval. In El-Saddik, A.; Mei, T.; Cucchiara, R.; Bertini, M.; Vallejo, D. P. T.; Atrey, P. K.; and Hossain, M. S., eds., *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 611–620. ACM.

Pan, J.; Ma, Q.; and Bai, C. 2023b. Reducing Semantic Confusion: Scene-aware Aggregation Network for Remote Sensing Cross-modal Retrieval. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 398–406.

Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. Deep Evidential Learning with Noisy Correspondence for Cross-modal Retrieval. In *ACM Multimedia*.

Qu, B.; Li, X.; Tao, D.; and Lu, X. 2016. Deep semantic understanding of high resolution remote sensing image. In *International Conference on Computer, Information and Telecommunication Systems, CITS 2016, Kunming, China, July 6-8, 2016*, 1–5. IEEE.

Qu, L.; Liu, M.; Wu, J.; Gao, Z.; and Nie, L. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1104–1113.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 815–823. IEEE Computer Society.

Tang, X.; Wang, Y.; Ma, J.; Zhang, X.; Liu, F.; and Jiao, L. 2023. Interacting-enhancing feature transformer for cross-modal remote-sensing image and text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.

Wang, F.; Pan, J.; Xu, S.; and Tang, J. 2022. Learning Discriminative Cross-Modality Features for RGB-D Saliency Detection. *IEEE Trans. Image Process.*, 31: 1285–1297.

Wright, J.; Ma, Y.; Mairal, J.; Sapiro, G.; Huang, T. S.; and Yan, S. 2010. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6): 1031–1044.

Wu, Y.; Li, L.; Jiao, L.; Liu, F.; Liu, X.; and Yang, S. 2024. TrTr-CMR: Cross-Modal Reasoning Dual Transformer for Remote Sensing Image Captioning. *IEEE Transactions on Geoscience and Remote Sensing*.

Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023. BiCro: Noisy Correspondence Rectification for Multi-modality Data via Bi-directional Cross-modal Similarity Consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 19883–19892. IEEE.

Yu, D.; and Ji, S. 2022. A New Spatial-Oriented Object Detection Framework for Remote Sensing Images. *IEEE Trans. Geosci. Remote. Sens.*, 60: 1–16.

Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; and Sun, X. 2022. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote. Sens.*, 60: 1–19.

Zhang, D.; Nan, F.; Wei, X.; Li, S.; Zhu, H.; McKeown, K.; Nallapati, R.; Arnold, A.; and Xiang, B. 2021. Supporting clustering with contrastive learning. *arXiv preprint arXiv:2103.12953*.

Zhang, L.; Xu, D.; Arnab, A.; and Torr, P. H. 2020. Dynamic graph message passing networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3726–3735.

Zhang, L.; and Zhang, L. 2022. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2): 270–294.

Zhang, Z.; Zhao, T.; Guo, Y.; and Yin, J. 2024. 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*.

Zhong, G.; Huan, R.; Wu, M.; Liang, R.; and Chen, P. 2025. Towards Robust Multimodal Emotion Recognition under Missing Modalities and Distribution Shifts. *arXiv preprint arXiv:2506.10452*.

Zhou, J.; Jampani, V.; Pi, Z.; Liu, Q.; and Yang, M.-H. 2021. Decoupled dynamic filter networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6647–6656.

Zhou, Y.; Suo, J.; Wang, Y.; Su, J.; Xiao, W.; Hong, Z.; Ranjan, R.; Wang, L.; and Wen, Z. 2025. MMCANet: A Multi-Modal and Cross-Attention Network for Cloud Removal and Exploration of Progressive Remote Sensing Images Restoration Algorithm. *IEEE Transactions on Geoscience and Remote Sensing*.