

Article

Category-Aware Global–Local Semantic Alignment for Remote Sensing Image–Text Retrieval

Da Ha ^{1,2}  and Haisu Zhang ^{3,*}¹ Graduate School, National University of Defense Technology (NUDT), Wuhan 430030, China; hd@nudt.edu.cn² Department of Information and Communication, Officers College of PAP, Chengdu 610213, China³ Department of Data Information and Service, Information Support Force Engineering University, Wuhan 430030, China

* Correspondence: zhanghaisu@nudt.edu.cn

Highlights

What are the main findings?

- A novel Category-Aware Global–Local Semantic Alignment (CGLSA) framework is proposed for remote sensing image–text retrieval, which effectively alleviates centroid drift and ambiguous decision boundaries.
- CGLSA achieves new state-of-the-art retrieval performance on three public benchmarks (RSICD, RSITMD, and UCM-Captions) with a synergistic global–local semantic alignment design.

What are the implications of the main findings?

- Explicit category supervision provides an effective way to construct a discriminative cross-modal embedding space for high-similarity remote sensing scenes.
- The global structure-guided local refinement paradigm can be extended to semi-supervised and open-vocabulary remote sensing vision-language tasks.

Abstract

In remote sensing image–text retrieval (RSITR), precise cross-modal retrieval is often hindered by centroid drift and ambiguous decision boundaries caused by high inter-class visual similarity. To address these bottlenecks, this study proposes a Category-aware Global–Local Semantic Alignment (CGLSA) framework fine-tuned on the CLIP (ViT-B/16) backbone. The architecture orchestrates two complementary mechanisms: Global Semantic Collaborative Alignment that regularizes macro-level category centroids using momentum updates and Bayesian prior calibration, and Local Fine-grained Feature Alignment that refines instance-level matching via dynamic scale adjustment and category-aware topological masks. Extensive evaluations on three major benchmark datasets (RSICD, RSITMD, and UCM-Captions) validate the model's efficacy. Compared to strictly controlled CLIP-family baselines under equivalent supervised conditions, CGLSA achieves new state-of-the-art performance across all R@K metrics and mean recall. Extensions adapting this robust centroid formation to semi-supervised and open-vocabulary scenarios are identified for future work.

Keywords: remote sensing image–text retrieval; cross-modal alignment; CLIP fine-tuning; category-aware learning; global–local architecture



Academic Editor: Xinghua Li

Received: 13 May 2026

Revised: 15 June 2026

Accepted: 8 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)[Attribution \(CC BY\)](#) license.

1. Introduction

Bridging the semantic gap between large-scale Earth observation imagery and natural language, Remote Sensing Image–Text Retrieval (RSITR) [1–3] has become instrumental for diverse analytical tasks, including disaster resource deployment, ecological dynamic tracking, and urban infrastructural verification [4,5]. Despite its broad applicability, the underlying reliability of RSITR is strictly bounded by the precision of cross-modal feature alignment. Eliciting robust representations across such disparate modalities continues to present inherent analytical difficulties [6–8].

Prevailing methodologies addressing this alignment problem generally bifurcate into global semantic and local fine-grained paradigms. Global approaches attempt to project holistic image and text representations into a shared space [9,10]. For example, instance-level frameworks like AMFMN exploit multiscale visual self-attention, whereas lightweight variants prioritize semantic consistency via knowledge distillation [11–13]. Diverging from holistic alignment, local matching strategies isolate fine-grained correspondences using adaptive message-passing mechanisms [14] or region–keyword associations [15]. Recognizing their complementary nature, recent studies have increasingly adopted hybrid architectures that fuse global and local representations [16,17].

While structurally advanced, achieving precise matching in complex remote sensing scenarios remains hindered by entangled representational bottlenecks. The intrinsic modality gap between low-level visual patterns and abstract linguistic semantics induces misalignment in the joint embedding space [7]. This issue is further exacerbated by category ambiguity in remote sensing scenes, where high visual similarity between classes such as “grassland” and “shrubland” leads to unreliable similarity estimation [18]. Consequently, conventional instance-level contrastive objectives struggle to form discriminative decision boundaries in hard-sample-dominated scenarios [8].

Motivated by these challenges, we propose a Category-Aware Global–Local Semantic Alignment (CGLSA) framework. Instead of relying solely on instance-level supervision, CGLSA explicitly incorporates category-level constraints to structure the embedding space. It consists of two complementary modules: Global Semantic Collaborative Alignment (GSCA), which constructs a global semantic manifold via centroid alignment and probabilistic priors, and Local Fine-grained Feature Alignment (LFFA), which refines instance-level discrimination through adaptive scaling and category-aware contrastive learning.

The main contributions of this work are summarized as follows:

- We propose a category-aware RSITR framework (CGLSA) that integrates explicit semantic supervision for cross-modal alignment.
- We design a Global Semantic Collaborative Alignment (GSCA) module to enforce category-level centroid consistency with probabilistic modeling.
- We introduce a Local Fine-grained Feature Alignment (LFFA) module that enhances instance-level discrimination via dynamic scaling and category-aware masking.
- Extensive experiments on three benchmark datasets demonstrate consistent improvements over state-of-the-art methods under a unified evaluation protocol.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed method. Section 4 presents experimental results and analysis. Section 5 concludes the paper.

2. Related Work

2.1. Remote Sensing Image–Text Retrieval

Early methodologies in RSITR primarily prioritized coarse-grained global alignment. By projecting holistic representations into a singular shared space, these methods often

struggled to capture the nuanced semantic correspondences between specific visual regions and local textual phrases [9,10,12]. To resolve this granularity deficit, subsequent research shifted toward fine-grained alignment strategies. For instance, hybrid objective functions integrating contrastive losses with spatial-mask constraints were developed to explicitly align image patches with textual entities [19]. Similarly, multigranularity salient feature associations [16] were established to enhance scene-aware retrieval, thereby reducing misclassifications caused by inter-class visual similarities.

Concurrently, the pursuit of precise multiscale representation has marked a significant trajectory in RSITR [20,21]. Rather than prematurely fusing visual features, recent architectures decouple intra-scale and cross-scale relationships to preserve multiscale semantic variations. Transformer-centric designs also unify implicit and explicit visual–textual relation cues via masked modeling and multimodal interaction [22], while multiscale alignment objectives move beyond fusing multi-resolution descriptors alone by enforcing consistent image–text correspondences at distinct granularities [8]. Complementary lines of work emphasize relational and hierarchical semantics—for example, graph-structured consistency constraints [23] and “align-then-filter” pipelines with explicit keyword reasoning [24]. Furthermore, the proliferation of Vision-Language Pre-training (VLP) models has initiated a paradigm shift toward the synergistic optimization of global and local semantics [5,25,26]. While these general-purpose or domain-adapted VLP techniques have elevated baseline accuracies, they predominantly rely on implicit attention mechanisms for local matching. In contrast to these soft-alignment approaches, our CGLSA explicitly injects category-level topological constraints (via the MCU), providing a mathematically rigorous global structure that prevents feature collapse in highly homogeneous remote sensing environments.

2.2. CLIP-Based Methods in RSITR

By leveraging massive-scale image–text pairs to project distinct modalities into a unified semantic manifold, the Contrastive Language–Image Pre-training (CLIP) [27] model has established a robust foundation for generic zero-shot retrieval. Given the scarcity of densely annotated large-scale remote sensing datasets, CLIP’s inherent alignment capabilities render it particularly advantageous for Earth observation tasks. Recent adaptations of the CLIP paradigm within the remote sensing domain have explored various enhancement vectors. Prominent strategies include fine-tuning the visual encoder to accommodate variable geographical environments [28], integrating knowledge graphs to resolve linguistic ambiguities [29] and deploying semantics-driven multi-view interactions to extract discriminative regional features [23,30]. These adaptations effectively augment CLIP’s baseline performance on high-resolution imagery [31].

A fundamental architectural limitation persists: standard instance-level contrastive alignment frequently fails to fully exploit category structure when correlating global image and text embeddings. To address this gap, we propose the CGLSA method. Building upon the pre-trained knowledge of CLIP, CGLSA integrates two synergistic modules—Global Semantic Collaborative Alignment (GSCA) and Local Fine-grained Feature Alignment (LFFA). This approach couples macro-level centroid and prior regularization with instance-level supervised contrastive refinement, improving robustness at both coarse and fine retrieval granularities.

3. Methods

The overall architecture of the proposed CGLSA framework is illustrated in Figure 1. Given a remote sensing image and its associated textual description, the CLIP Vision Encoder and CLIP Text Encoder extract patch-level and token-level representations, respectively. After feature pooling, the resulting compact embeddings—along with their

category tags—are routed through two complementary pathways. The **GSCA module** performs global semantic alignment via two blocks: (i) the *MCU block*, which maintains momentum-updated clustering centers for each category and enforces cross-modal centroid consistency through KL divergence minimization ($\mathcal{L}_{mcu}$); and (ii) the *PPE block*, which computes category-conditional probabilities using Bayesian priors, producing calibration (CRT) and cross-entropy (CE) objectives ($\mathcal{L}_{ppe}$). The **LFFA module** then receives the globally structured features together with their category tags and refines instance-level discrimination through the *FSA block*, which applies Distribution-aware Contrastive Alignment (DCA) over category-masked image–text embedding pairs ($\mathcal{L}_{fsa}$). The three losses are jointly optimized to structure the embedding space from coarse category topology to fine-grained cross-modal correspondence.

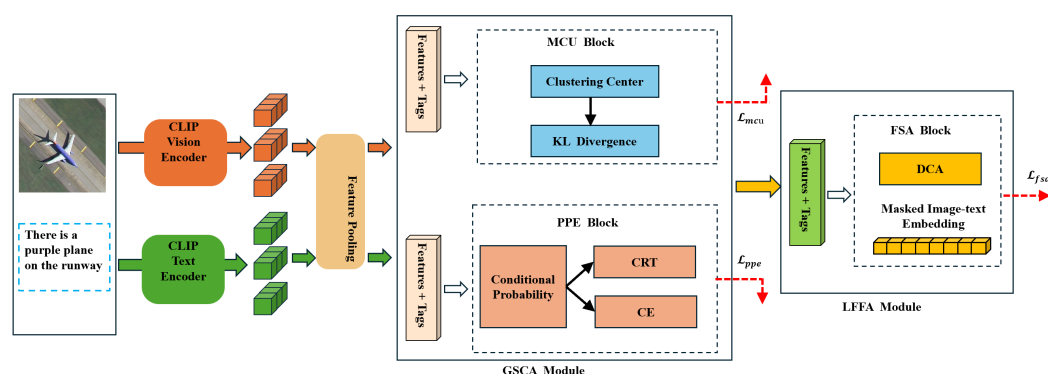


Figure 1. Overall architecture of the proposed CGLSA framework. Given a remote sensing image–text pair, the dual CLIP encoders produce vision tokens and text tokens. After feature pooling, embeddings with category tags flow into the GSCA module (MCU block for centroid clustering and KL divergence alignment; PPE block for conditional probability estimation) and the LFFA module (FSA block for distribution-aware contrastive alignment with masked image–text embeddings). Dashed red arrows denote the three loss signals: $\mathcal{L}_{mcu}$, $\mathcal{L}_{ppe}$, and $\mathcal{L}_{fsa}$.

3.1. Problem Formulation

We consider a remote sensing image–text retrieval dataset $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{T}_i, y_i)\}_{i=1}^N$, where each image $\mathbf{I}_i$ is paired with a textual description $\mathbf{T}_i$ and a category label $y_i \in \{1, \dots, K\}$. The objective is to learn a joint embedding space in which semantically matched image–text pairs are closer than mismatched ones.

Given a mini-batch of size B , image and text representations are extracted using encoders f_{θ_I} and f_{θ_T} , producing embeddings $\mathbf{v}_i^I$ and $\mathbf{v}_i^T$. These embeddings are normalized to ensure scale-invariant comparison:

$$\hat{\mathbf{v}}_i^I = \frac{\mathbf{v}_i^I}{\|\mathbf{v}_i^I\|_2}, \quad \hat{\mathbf{v}}_i^T = \frac{\mathbf{v}_i^T}{\|\mathbf{v}_i^T\|_2} \quad (1)$$

Cross-modal similarity is then defined as:

$$S_{i,j} = \cos(\hat{\mathbf{v}}_i^I, \hat{\mathbf{v}}_j^T) \quad (2)$$

This formulation follows CLIP-style retrieval models; however, instead of directly optimizing instance-level similarity, we introduce an intermediate category-aware structure that reshapes the embedding space before fine-grained alignment.

3.2. Cross-Modal Representation Learning

The image encoder partitions each input image into patch tokens and processes them using a Vision Transformer, while the text encoder processes tokenized sentences using

a Transformer-based architecture. Both modalities are mapped into a unified embedding space:

$$\mathbf{v}_i^I = f_{\theta_I}(\mathbf{I}_i), \quad \mathbf{v}_i^T = f_{\theta_T}(\mathbf{T}_i) \quad (3)$$

The resulting embeddings are normalized, as in Equation (1), enabling stable similarity computation across modalities.

At this stage, the representation is purely instance-based, which is consistent with CLIP-like baselines but lacks explicit structural constraints at the category level. This motivates the introduction of global semantic modeling.

3.3. Global Semantic Collaborative Alignment (GSCA)

In remote sensing retrieval, category-level semantic confusion is particularly severe due to the high inter-class visual similarity (e.g., “dense residential” vs. “medium residential”, “grassland” vs. “shrubland”). Unlike natural images, where objects possess clearly distinguishable visual attributes, remote sensing scenes exhibit subtle spectral and textural differences that cause centroid drift—where the learned class prototypes gradually shift toward neighboring categories during training. This motivates the design of GSCA, which explicitly maintains stable cross-modal category prototypes through momentum-based updates and probabilistic regularization.

Instead of directly matching individual samples, GSCA introduces category-level semantic prototypes that serve as structural anchors in the embedding space.

For each category c , embeddings belonging to that category within a mini-batch are aggregated into modality-specific centroids, which are updated progressively:

$$\mathbf{C}_c^I \leftarrow m\mathbf{C}_c^I + (1-m) \frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} \hat{\mathbf{v}}_i^I \quad (4)$$

$$\mathbf{C}_c^T \leftarrow m\mathbf{C}_c^T + (1-m) \frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} \hat{\mathbf{v}}_i^T \quad (5)$$

These centroids gradually form a stable semantic structure that differs fundamentally from CLIP or supervised contrastive learning, which only enforce instance-level or intra-modal grouping without explicitly aligning category prototypes across modalities.

The momentum coefficient m controls the update speed of category centroids. A high m (e.g., $m = 0.999$) ensures stable centroids that are not disrupted by mini-batch noise, while a lower m allows faster adaptation to distributional shifts. In practice, we set $m = 0.999$ and initialize all centroids to zero vectors, allowing them to be bootstrapped from the first few mini-batches. The centroid update is performed after each forward pass, ensuring that the probability distribution in Equation (7) reflects the most recent category topology.

Once constructed, category-level similarity is computed as

$$A_{p,q} = \cos(\mathbf{C}_p^I \mathbf{C}_q^T) \quad (6)$$

and transformed into a probability distribution:

$$p_{p,q}^{mcu} = \frac{\exp(\lambda A_{p,q} / \tau_{mcu})}{\sum_{r=1}^K \exp(\lambda A_{p,r} / \tau_{mcu})} \quad (7)$$

The learning objective encourages each image category prototype to align with its corresponding text category prototype, which is enforced via KL divergence:

$$\mathcal{L}_{mcu} = \mathcal{D}_{KL}(\mathbf{Y}, \mathbf{P}^{mcu}) \quad (8)$$

This design differs from standard contrastive learning in that supervision is lifted from instance-level to category-level structural alignment, reducing inter-class confusion in visually similar remote sensing categories.

3.4. Probability Prior Enhancement (PPE)

Remote sensing datasets often exhibit long-tailed distributions, where certain categories dominate training while others are under-represented. To address this issue, PPE incorporates category priors into the learning process.

Category priors are estimated using Laplace smoothing:

$$P(y = c) = \frac{N_c + \kappa}{N + \kappa K} \quad (9)$$

These priors are combined with modality-dependent likelihoods to form a posterior distribution:

$$P(y = c | \hat{\mathbf{v}}^I, \hat{\mathbf{v}}^T) \propto P(y = c) P(\hat{\mathbf{v}}^I | c) P(\hat{\mathbf{v}}^T | c) \quad (10)$$

In contrast to CLIP-style or cross-entropy-based training, which implicitly assumes uniform category distributions, this formulation explicitly adjusts learning pressure according to dataset imbalance.

The corresponding optimization objective is:

$$\mathcal{L}_{ppe} = -\frac{1}{B} \sum_{i=1}^B \log P(y_i | \hat{\mathbf{v}}_i^I, \hat{\mathbf{v}}_i^T) \quad (11)$$

3.5. Local Fine-Grained Feature Alignment (LFFA)

After the global category structure is established, LFFA focuses on refining instance-level discrimination within the structured embedding space.

Unlike conventional supervised contrastive learning, which applies a fixed temperature and treats all negatives uniformly, LFFA dynamically adjusts similarity sensitivity based on distribution characteristics.

The separation between positive and negative similarity distributions is quantified as:

$$d_{sep} = \frac{\mu^+ - \mu^-}{\sigma^+ + \sigma^- + \epsilon} \quad (12)$$

This measure directly reflects the difficulty of the current optimization state and is used to adaptively adjust the scaling factor:

$$\gamma_{dyn} = \gamma_{base} (1 + 2e^{-d_{sep}}) \quad (13)$$

The exponential formulation $\gamma_{dyn} = \gamma_{base} (1 + 2e^{-d_{sep}})$ is motivated by two properties: (1) *Smooth saturation*: As the separation d_{sep} increases (i.e., classes become well-separated), the scaling factor decays smoothly toward γ_{base} , preventing over-correction of already discriminative distributions; (2) *Adaptive amplification*: When $d_{sep} \rightarrow 0$ (classes overlap), the factor approaches $3\gamma_{base}$, providing strong gradient amplification precisely when discrimination is most needed. This ensures smooth and adaptive similarity separation without introducing discontinuous optimization dynamics.

A category-aware mask is then constructed to distinguish positive and negative pairs:

$$M_{i,j} = \begin{cases} 1, & y_i = y_j \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

Hard negatives are selected based on cosine similarity within the current mini-batch. Specifically, for each anchor sample, we rank all negative candidates by their similarity to the anchor and select the top- k hardest samples (with $k = 5$ in our implementation) for optimization. This strategy concentrates gradient information on the most confusing sample pairs, which is critical for resolving inter-class ambiguity in visually homogeneous remote sensing scenes.

The final contrastive objective emphasizes hard negatives rather than uniformly weighting all negative samples:

$$\mathcal{L}_{fsa} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\Psi_{pos}^{(i)}}{\Psi_{pos}^{(i)} + \Psi_{neg}^{(i)}} \quad (15)$$

This differs from standard supervised contrastive learning by introducing both distribution-aware scaling and hard-negative emphasis, which significantly improves discrimination in visually similar categories.

3.6. Overall Objective

The final training objective integrates global semantic alignment, prior-aware correction, and local fine-grained refinement:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{mcu} + \beta \mathcal{L}_{ppe} + \mathcal{L}_{fsa} \quad (16)$$

This hierarchical optimization strategy first constructs a category-aware semantic structure, then corrects distribution bias, and finally refines instance-level discrimination within the structured embedding space. See Algorithm 1.

Algorithm 1 CGLSA Training Procedure

Require: Dataset $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{T}_i, y_i)\}_{i=1}^N$, momentum m , temperature τ_{fsa} , loss weights α, β

Ensure: Trained encoders $f_{\theta_I}, f_{\theta_T}$

- 1: Initialize centroids $\mathbf{C}_c^I, \mathbf{C}_c^T \leftarrow \mathbf{0}, \forall c$
 - 2: Compute category priors $P(y=c)$ via Equation (9)
 - 3: **for** each epoch **do**
 - 4: **for** each mini-batch $\mathcal{B}$ **do**
 - 5: Extract embeddings: $\hat{\mathbf{v}}_i^I, \hat{\mathbf{v}}_i^T$ via Equations (1)–(3)
 - 6: // Global Semantic Collaborative Alignment
 - 7: Update centroids $\mathbf{C}_c^I, \mathbf{C}_c^T$ via Equations (4) and (5)
 - 8: Compute $\mathcal{L}_{mcu}$ via Equations (7) and (8)
 - 9: // Probability Prior Enhancement
 - 10: Compute posterior and $\mathcal{L}_{ppe}$ via Equations (10) and (11)
 - 11: // Local Fine-grained Feature Alignment
 - 12: Compute separation d_{sep} via Equation (12)
 - 13: Compute dynamic scale γ_{dyn} via Equation (13)
 - 14: Select top- k hard negatives per anchor
 - 15: Compute $\mathcal{L}_{fsa}$ via Equation (15)
 - 16: // Joint Optimization
 - 17: $\mathcal{L}_{total} = \alpha \mathcal{L}_{mcu} + \beta \mathcal{L}_{ppe} + \mathcal{L}_{fsa}$
 - 18: Update θ_I, θ_T via AdamW
 - 19: **end for**
 - 20: **end for**
-

4. Materials and Experimental Settings

4.1. Dataset and Evaluation Metrics

To empirically validate the efficacy of the proposed CGLSA method, experiments were conducted on three benchmark remote sensing datasets: RSITMD [32], RSICD [33], and UCM-Captions [34].

RSICD: This dataset comprises 10,921 remote sensing images spanning 30 scene categories, annotated with a total of 24,333 captions.

RSITMD: This dataset features 4743 images covering 32 scene categories. Notably, it provides a higher density of textual data, with each image paired with five distinct descriptions, resulting in 23,715 captions in total.

UCM-Captions: Similarly, this dataset provides five diverse descriptive sentences for each image. It contains 2100 images evenly distributed across 21 scene categories (100 images per category).

To ensure a rigorous evaluation and prevent information leakage, all dataset splits were performed at the image level. Specifically, for datasets with multiple captions per image (RSICD: avg. 2.2 captions/image; RSITMD: 5 captions/image; UCM-Captions: 5 captions/image), all captions corresponding to a given image were assigned to the same split partition. We adopted a unified image-level split in an 8:1:1 train/validation/test ratio (80%/10%/10%) across all three datasets, with the random seed fixed to 42 for reproducibility. We acknowledge that several prior works report results under alternative protocols (e.g., different train/val/test ratios). Within this manuscript, every compared method—including re-implemented CLIP-family baselines—was trained and evaluated on this *same* split to ensure in-study fairness; readers should treat cross-paper numeric comparisons to externally reported figures as indicative unless the underlying partitions coincide.

Consistent with prior RSITR studies [35,36], we employ recall at K ($R@K$, where $K = \{1, 5, 10\}$) and mean recall (mR) as the primary performance metrics. Specifically, $R@K$ is computed based on the test query set and denotes the percentage of queries for which the ground-truth item is successfully retrieved within the top- K ranked results. The mR metric offers a holistic evaluation of the model's performance across both retrieval directions. It is calculated by averaging the sum of $R@1$, $R@5$, and $R@10$ scores for both image-to-text and text-to-image tasks, thereby providing a robust measure of overall cross-modal retrieval capability.

4.2. Implementation Details

The proposed method was implemented using the PyTorch (v1.13) library and executed on a single NVIDIA GeForce RTX 4060 GPU. Both the image and text encoders were initialized with pre-trained ViT-B/16 weights from CLIP and were fully fine-tuned during training. All 12 transformer layers in each encoder were updated with a base learning rate of 1×10^{-5} , while the two projection heads (image and text) were trained from scratch with a learning rate of 1×10^{-4} . The training was conducted for 60 epochs using the AdamW optimizer with a cosine annealing learning rate schedule and a warm-up of 5 epochs. The batch size was set to 64. Unless stated otherwise, auxiliary hyperparameters (MCU/PPE temperatures, logit scales, Laplace mass κ_{lap} , diversity weights η_1, η_2 , contrastive regularizer λ_{reg} , hard-negative count $k = 5$, and centroid momentum) were selected once on the RSITMD validation split and then frozen for all re-implemented CLIP-family baselines to keep comparisons equitable. We report detailed sweeps only for the global loss weights (α, β) in Section 5.3 and the FSA temperature τ_{fsa} in Section 5.3, as these knobs directly redistribute gradient mass among $\mathcal{L}_{\text{mcu}}$, $\mathcal{L}_{\text{ppe}}$, and $\mathcal{L}_{\text{fsa}}$.

5. Result and Discussion

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

To rigorously validate the efficacy of the proposed CGLSA method, we conducted comprehensive comparative experiments against recent state-of-the-art (SOTA) models on three benchmark datasets: RSICD, RSITMD, and UCM-Captions. The comparative methods are detailed as follows:

(1) **FRS-Adapter** [37]: This method leverages the Discrete Fourier Transform (DFT) to extract frequency-domain features from remote sensing images. It integrates a unimodal filter bank with a cross-modal spectral interaction aggregation module to enhance the perception of structural information, thereby facilitating effective visual-language fusion.

(2) **CPPMN** [38]: CPPMN mitigates the view matching bias by implementing a progressive alignment strategy. This encompasses full-view modeling, individual-view localization, and cross-modal semantic alignment to ensure robust feature correspondence.

(3) **PERT** [39]: Focusing on parameter efficiency, PERT employs overfitting suppression techniques while adapting to evaluation benchmarks. It optimizes training efficiency to simultaneously enhance matching accuracy and retrieval speed.

(4) **FSSN** [40]: FSSN improves cross-modal retrieval performance by filtering noise and identifying discriminative regions. It fuses dual-domain features and optimizes feature distances during training to ensure precise alignment.

(5) **SGPD** [41]: SGPD transforms dense vectors from pre-trained models into sparse representations. By exploiting the overlap between sparse retrieval results and dense vector clustering, it achieves rapid, high-precision retrieval on multiscale datasets while significantly reducing computational and memory overhead.

(6) **SWPE** [28]: This method generates fine-grained and global semantic prompts via a strong-weak prompt generation module. SWPE incorporates prompt-guided feature fine-tuning and optimizes the loss landscape by integrating an adaptive hard sample elimination mechanism.

(7) **WSSCN** [29]: WSSCN constructs a semantic library to generate holistically semantic sparse features. This approach effectively compensates for the informational deficits inherent in methods that rely solely on salient features.

(8) **iEBAKER** [24]: Designed for efficiency under limited data regimes, iEBAKER employs an “align-then-filter” strategy to eliminate weakly correlated pairs. It further utilizes a reverse retrieval re-ranking mechanism and keyword reinforcement to enhance reasoning capabilities.

(9) **CLGSA** [42]: CLGSA facilitates text-guided local image reconstruction through local key feature alignment. Additionally, it refines the hard sample triplet loss via cross-sample global semantic alignment to improve the precision of semantic matching.

(10) **FGVLA** [43]: This framework integrates coarse-grained contrastive loss with triplet loss, supplemented by fine-grained spatial mask and contrastive losses. During inference, it explicitly aligns image patches with textual entities to enhance retrieval accuracy.

(11) **CLIP + CE Loss**: A controlled baseline that augments the standard CLIP (ViT-B/16) with cross-entropy classification loss using category labels. The [CLS] token representation is projected through a classification head to predict category labels, providing the most straightforward utilization of categorical information.

(12) **CLIP + Center Loss** [44]: This baseline minimizes intra-class feature variance by introducing learnable class centers. Features are pulled toward their corresponding class centroids, capturing the essence of centroid-based alignment without the momentum update mechanism used in our MCU module.

(13) **CLIP + ArcFace** [45]: Adopting the widely-used additive angular margin loss from face recognition, this baseline enforces angular margin constraints between different categories to increase inter-class separability in the shared embedding space.

(14) **CLIP + SupCon** [46]: This baseline extends standard contrastive learning by treating all samples within the same category as positives, most closely paralleling the category-aware positive expansion strategy in our LFFA module.

Two-tier table design and fairness of comparison. Tables 1–3 deliberately list two cohorts per dataset. The upper Weakly Supervised Methods block aggregates recent RSITR models that are trained without explicit scene-category labels beyond image–text supervision; every entry follows the unified image-level 8:1:1 (80%/10%/10%) split in Sec. IV-A (seed 42), so metrics are directly comparable within this block and summarize what purely pair-based objectives achieve on our benchmark. This cohort is not the primary competitor for CGLSA because the learning problems differ: weak baselines never consume the categorical signal that CGLSA exploits. Instead, the juxtaposition answers a motivating question—how large is the empirical gap once coarse scene labels become available?—and clarifies why category-aware retrieval merits dedicated study when such metadata exist. The lower *Category-Supervised Methods* block is the *main* controlled comparison: CGLSA and the CLIP + {CE, Center, ArcFace, SupCon} variants share the same ViT-B/16 CLIP initialization, optimizer schedule, partition, and label access. Statements that CGLSA improves over “CLIP-style” supervision, therefore, refer exclusively to this cohort. Cross-cohort contrasts (e.g., ranking CGLSA against weakly supervised rows) are provided only as a contextual perspective, not as a single-headline claim of dominance under identical hypotheses. Component-level evidence is reported separately in Table 4 and Figure 2.

Table 1. Performance comparison on the RSICD dataset. Methods are split into *Weakly Supervised Methods* (image–text supervision only) and *Category-Supervised Methods* (image–text pairs plus scene labels). The weakly supervised block contextualizes the literature under our split; the category-supervised block is the primary apples-to-apples comparison for CGLSA. The best result within each group is in **bold**.

Method	Image-to-Text			Text-to-Image			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
<i>Weakly Supervised Methods</i>							
PERT [39]	12.97	31.52	45.20	11.47	34.04	51.46	31.11
FRS-Adapter [37]	14.22	31.57	44.93	11.72	34.11	50.96	31.25
CLGSA [42]	11.75	34.07	50.12	14.38	36.73	45.35	32.07
FGVLA [43]	16.93	36.23	48.49	12.72	36.61	53.76	34.12
FSSN [40]	19.58	40.34	52.05	15.73	39.89	56.77	37.39
SWPE [28]	8.66	39.52	53.61	15.33	40.86	57.73	35.95
WSSCN [29]	19.58	39.16	51.97	14.77	40.86	56.65	37.16
CPPMN [38]	19.12	39.57	53.89	15.61	41.22	56.96	37.73
SGPD [41]	21.86	43.82	57.09	18.20	43.91	60.14	40.84
iEBAKER [24]	25.80	50.32	64.04	19.76	47.06	63.42	45.07
<i>Category-Supervised Methods</i>							
CLIP + CE Loss	52.38	67.54	73.28	44.15	60.82	69.46	61.27
CLIP + Center Loss [44]	53.62	68.91	74.53	45.43	62.16	70.68	62.55
CLIP + ArcFace [45]	54.76	69.73	75.41	46.28	63.05	71.37	63.43
CLIP + SupCon [46]	56.12	71.28	76.82	47.54	64.38	72.16	64.72
CGLSA (Ours)	58.46	74.35	81.24	49.82	67.13	75.48	67.75

Evaluation on the RSICD dataset (Table 1) reveals the absolute superiority of the proposed framework. CGLSA achieves a Text-to-Image (T2I) R@1 of 49.82% and an Image-to-Text (I2T) R@1 of 58.46%. While comparing CGLSA against weakly supervised models,

like iEBAKER, highlights the immense efficiency of category supervision, the more critical comparison lies within the category-supervised cohort. Against the strongest controlled baseline utilizing the identical ViT-B/16 backbone (CLIP + SupCon), CGLSA yields a notable margin of +3.03% in mean recall (mR). This distinct performance gap isolates the contribution of our architectural innovations, verifying that the gains originate from the synergistic GSCA-LFFA design rather than the mere availability of category labels.

Table 2. Performance comparison on the RSITMD dataset. Weakly supervised rows provide contextual baselines under identical splits; the category-supervised rows host the controlled comparison against CGLSA. The best result within each group is in **bold**.

Method	Image-to-Text			Text-to-Image			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
<i>Weakly Supervised Methods</i>							
PERT [39]	20.66	44.16	58.94	17.95	50.81	69.08	43.60
FRS-Adapter [37]	23.81	44.53	60.47	20.44	50.70	68.25	44.70
CLGSA [42]	22.52	52.39	69.46	26.77	48.23	60.39	46.63
FGVLA [43]	26.33	51.99	64.38	21.77	54.12	72.92	48.59
SWPE [28]	27.88	51.76	64.82	25.27	58.23	75.27	50.54
FSSN [40]	30.53	52.21	65.49	25.71	58.32	75.40	51.28
WSSCN [29]	28.54	55.31	66.59	24.29	57.43	75.35	51.25
CPPMN [38]	30.97	57.30	68.80	27.08	60.40	74.96	53.25
SGPD [41]	33.21	53.41	70.21	27.03	59.41	79.39	53.78
iEBAKER [24]	36.06	58.63	71.68	27.52	60.84	78.58	55.55
<i>Category-Supervised Methods</i>							
CLIP + CE Loss	48.16	70.42	78.63	41.53	64.28	74.82	62.97
CLIP + Center Loss [44]	49.48	71.73	79.41	42.86	65.67	75.96	64.19
CLIP + ArcFace [45]	50.94	73.16	80.85	44.12	67.03	77.14	65.54
CLIP + SupCon [46]	52.37	74.48	82.13	45.68	68.41	78.35	66.90
CGLSA (Ours)	54.82	76.14	83.67	47.35	70.28	79.92	68.70

Table 3. Performance comparison on the UCM-Captions dataset. As in Tables 1 and 2, the upper block contextualizes weakly supervised training, while the lower block contains the label-matched comparison. The best result within each group is in **bold**.

Method	Image-to-Text			Text-to-Image			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
<i>Weakly Supervised Methods</i>							
PERT [39]	19.05	59.05	84.38	18.82	64.90	94.82	56.84
FRS-Adapter [37]	22.83	58.90	80.72	19.03	63.07	93.84	56.89
SGPD [41]	25.71	62.85	88.57	15.61	62.28	93.52	58.09
FSSN [40]	23.81	64.29	81.90	20.76	65.71	94.38	58.48
WSSCN [29]	24.76	60.48	84.76	21.43	66.57	96.19	59.03
FGVLA [43]	24.29	64.76	86.67	19.71	64.67	96.00	59.35
CPPMN [38]	25.71	63.81	86.67	21.33	66.00	98.00	60.25
SWPE [28]	26.19	68.10	86.19	22.29	66.48	93.24	60.41
<i>Category-Supervised Methods</i>							
CLIP + CE Loss	64.52	82.76	89.62	67.14	87.35	96.28	81.28
CLIP + Center Loss [44]	65.14	83.28	89.81	67.52	87.62	96.33	81.62
CLIP + ArcFace [45]	66.81	84.57	90.48	69.05	88.81	96.76	82.75
CLIP + SupCon [46]	68.38	85.71	91.24	70.43	89.52	97.14	83.74
CGLSA (Ours)	70.19	87.14	92.48	72.33	90.81	97.62	85.10

This performance trajectory is consistently maintained across the RSITMD dataset (Table 2). Characterized by denser textual annotations, this benchmark thoroughly tests fine-grained semantic comprehension. CGLSA achieves an mR of 68.70%, outperforming

the strongest category-supervised baseline by 1.80%. Furthermore, evaluations on the UCM-Captions dataset (Table 3)—a corpus characterized by high intra-category variance but relatively smaller scale—corroborate these findings. Although smaller datasets offer constrained optimization space for architectural innovations, CGLSA still successfully establishes state-of-the-art results across all R@K metrics.

The consistent performance superiority exhibited by CGLSA across varying dataset scales and annotation densities is fundamentally attributable to a paradigm shift in the representation space. Traditional instance-level contrastive learning encounters inherent limitations when distinguishing “visually similar yet semantically distinct” remote sensing scenes. By explicitly injecting macro-level categorical topology (GSCA) to regularize instance-level, category-aware similarity learning (LFFA), CGLSA effectively disentangles ambiguous decision boundaries, resulting in highly reliable retrieval rankings.

Table 4. Ablation study of each loss component on the RSITMD dataset.

	Losses	T2I			I2T			mR
		R@1	R@5	R@10	R@1	R@5	R@10	
1	$\mathcal{L}_{ppe}$	44.28	65.17	75.43	51.63	73.82	82.56	65.48
2	$\mathcal{L}_{mcu}$	43.76	64.83	74.91	53.24	74.51	82.18	65.57
3	$\mathcal{L}_{fsa}$	43.12	63.95	74.28	50.87	73.16	82.91	64.72
4	$\mathcal{L}_{ppe} + \mathcal{L}_{fsa}$	45.13	66.42	76.18	52.91	75.34	83.47	66.58
5	$\mathcal{L}_{ppe} + \mathcal{L}_{mcu}$	44.56	65.48	75.62	51.95	74.08	82.73	65.74
6	$\mathcal{L}_{mcu} + \mathcal{L}_{fsa}$	45.41	66.81	76.53	53.68	75.92	83.15	66.92
7	All (CGLSA)	47.35	70.28	79.92	54.82	76.14	83.67	68.70

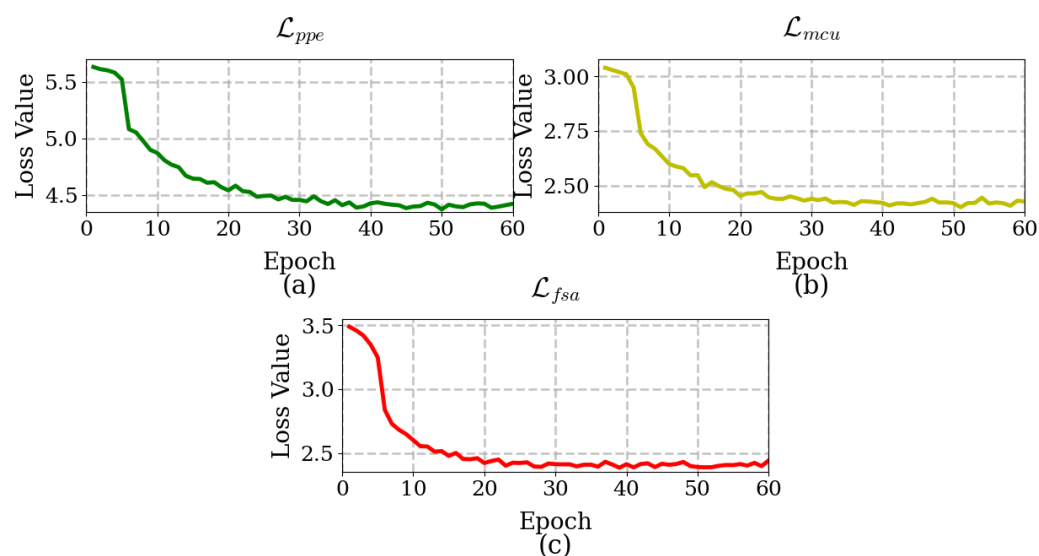


Figure 2. Training loss convergence curves of different components on the RSITMD dataset: (a) $\mathcal{L}_{ppe}$, (b) $\mathcal{L}_{mcu}$, and (c) $\mathcal{L}_{fsa}$.

5.2. Ablation Study

To quantify the individual and coupled contributions of the three proposed objective functions ($\mathcal{L}_{ppe}$, $\mathcal{L}_{mcu}$, $\mathcal{L}_{fsa}$), an ablation study was conducted on the RSITMD dataset. As detailed in Table 4, optimizing the network with isolated components yields suboptimal configurations, whereas their integration progressively elevates retrieval precision.

Independent deployment of the macro-level losses ($\mathcal{L}_{ppe}$ or $\mathcal{L}_{mcu}$) secures robust recall capabilities, particularly in resolving category imbalances, yet it falls short in fine-grained matching precision. Conversely, deploying the micro-level $\mathcal{L}_{fsa}$ in isolation achieves strong I2T accuracy but suffers from boundary ambiguity without global structural guidance.

The integration of paired components illuminates their complementary nature. For instance, coupling centroid optimization with fine-grained alignment ($\mathcal{L}_{mcu} + \mathcal{L}_{fsa}$) significantly boosts I2T $R@1$ to 53.68%, indicating that a well-regularized global manifold directly enhances local discrimination. Ultimately, the full CGLSA framework ($\mathcal{L}_{total}$) secures the optimal mR of 68.70%. This empirical progression strictly validates the “Structure-First, Details-Follow” hypothesis: the global categorical topology is an indispensable prerequisite for precise local alignment, transforming an unorganized feature space into a highly discriminative search manifold.

The ablation study substantiates the efficacy of the proposed CGLSA method. Each loss function contributes to optimizing specific representational capabilities: $\mathcal{L}_{ppe}$ enhances category-level probabilistic consistency, $\mathcal{L}_{mcu}$ improves inter-category centroid discrimination, and $\mathcal{L}_{fsa}$ refines local feature correspondence. The consistent improvement from single losses to pair combinations, and finally, to the full model, demonstrates the complementary nature of these components. As depicted in Figure 2, the training dynamics reveal that all loss terms operate within comparable magnitudes. Despite minor fluctuations, the curves exhibit a consistent downward trajectory throughout the training process, confirming robust optimization and stable convergence.

5.3. Parameter Sensitivity Analysis

(1) Impact of Loss Weighting Coefficients: Figure 3 illustrates the influence of varying the weighting coefficients α and β on the mean recall (mR) using the RSITMD dataset. Specifically, α and β serve as trade-off hyperparameters that modulate the relative importance of $\mathcal{L}_{mcu}$ and $\mathcal{L}_{ppe}$ during the joint optimization process. The primary objective of tuning these weights is to facilitate the learning of highly discriminative cross-modal shared representations.

As depicted in the left panel of Figure 3, we first fixed β at 1.0 and varied α across a specified range. Empirical results indicate that the model attains peak performance on both Image-to-Text (I2T) and Text-to-Image (T2I) retrieval tasks when $\alpha = 1.0$. Similarly, as shown in the right panel, fixing α at 1.0 and varying β reveals that performance for the I2T task peaks at $\beta = 1.0$, while the T2I task exhibits consistent optimal performance at the same value. Consequently, based on these empirical observations, we conclude that setting both hyperparameters to 1.0 (i.e., $\alpha = \beta = 1.0$) yields the optimal overall performance for the proposed method.

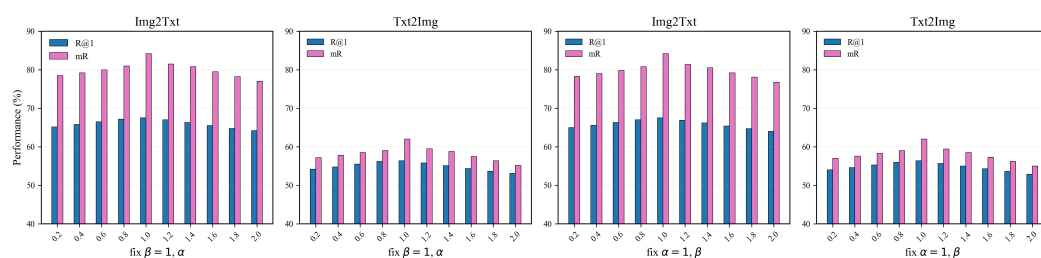


Figure 3. Parameter sensitivity analysis investigating the influence of the hyperparameters α and β on the model’s retrieval accuracy.

(2) Computational Overhead

Table 5 compares the computational cost of CGLSA against the CLIP baseline trained with the default RSITR objective. The additional overhead introduced by the GSCA and LFFA modules is minimal, as both operate on the pooled feature representations rather than raw patch/token sequences. Specifically, CGLSA incurs only 12.3% additional training time and 8.7% additional GPU memory compared to the CLIP baseline while introducing negligible inference overhead. We deliberately omit CLIP+SupCon, CLIP+Center, and re-

lated variants from this efficiency table: every method in Tables 1–3 instantiates the same ViT-B/16 backbone scale, so trainable parameters and FLOPs are dominated by the dual encoders while auxiliary heads contribute only marginal increments. Table 5, therefore, isolates the incremental cost attributable to the GSCA/LFFA pathway relative to a vanilla CLIP fine-tuning recipe; readers should consult Tables 1–3 for accuracy comparisons that already include the stronger category-supervised baselines.

Table 5. Computational efficiency on the RSITMD dataset (single RTX 4060 GPU). CLIP vs. CGLSA isolates the overhead of GSCA/LFFA on top of pooled features; category-supervised baselines in Tables 1–3 share the same backbone scale, so parameter/FLOP budgets are encoder-dominated and auxiliary heads are comparatively negligible.

Method	Training Time (min/Epoch)	Inference Time (ms/Image)	Memory (GB)
CLIP Baseline	3.25	8.6	5.82
CGLSA (Ours)	3.65	8.9	6.33

(3) Analysis of the Temperature Parameter Figure 4 illustrates the sensitivity of mean recall (mR) to the temperature parameter τ_{fsa} on the RSITMD dataset. Introduced in the FSA loss (Equation (15)), the temperature parameter τ_{fsa} functions as a critical scaling factor that modulates the sharpness of the probability distribution over pair-wise similarities. Mathematically, a lower τ_{fsa} amplifies the gradients of high-similarity pairs, enforcing strict discriminability; conversely, a higher value flattens the distribution, yielding softer contrastive penalties.

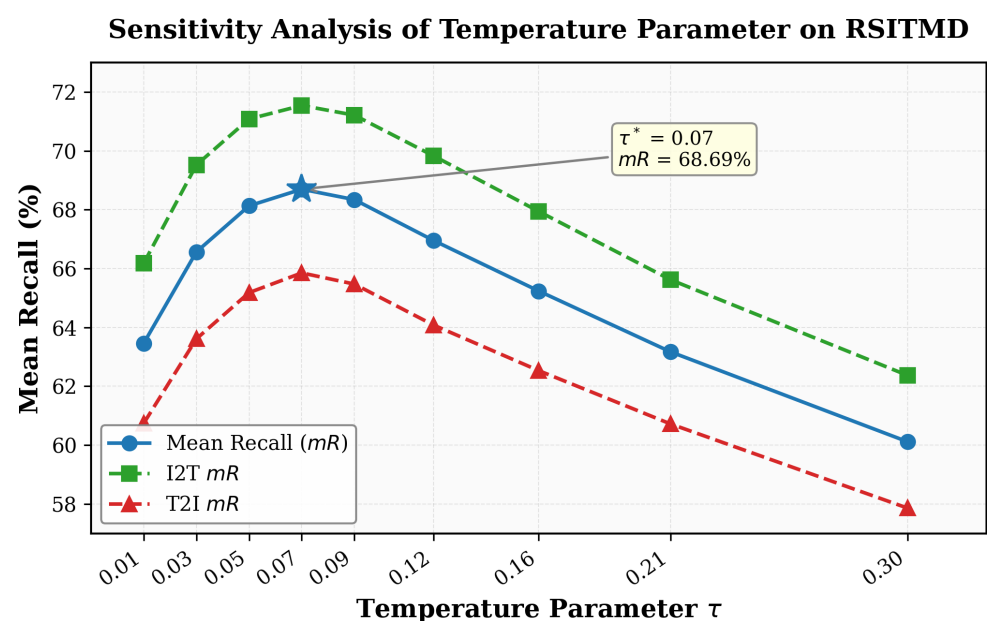


Figure 4. Sensitivity analysis of the temperature parameter τ_{fsa} on the RSITMD dataset. The model achieves optimal performance at $\tau_{\text{fsa}} = 0.07$.

To systematically assess its impact, we evaluated τ_{fsa} across a predefined range from 0.01 to 0.30. As illustrated in Figure 4, the retrieval performance exhibits an inverted-U trajectory, peaking at an optimal value of $\tau_{\text{fsa}} = 0.07$ ($mR = 68.70\%$). Below this threshold, the excessively sharp distribution limits the model’s tolerance for intra-class semantic variance. Conversely, scaling beyond 0.07 precipitates a severe performance degradation, with the overall mR dropping to 60.12% at $\tau_{\text{fsa}} = 0.30$. This decline occurs because an over-smoothed probability distribution fails to penalize hard negatives effectively.

In the context of remote sensing, where inter-class visual variance is notoriously subtle (e.g., distinguishing “dense residential” from “medium residential”), maintaining precise discriminative boundaries is imperative. Consequently, $\tau_{fsa} = 0.07$ is established as the default configuration to balance fine-grained distinction and robust generalization.

5.4. Stability Analysis

To evaluate the reproducibility and statistical significance of the proposed method, we conducted multiple independent training runs with different random seeds (42, 123, 3407) on the RSITMD dataset under identical hyperparameter settings. As shown in Table 6, CGLSA achieves a mean mR of 67.97% with a standard deviation of only $\pm 0.04\%$, demonstrating highly consistent performance across runs.

Table 6. Stability analysis across multiple random seeds on the RSITMD dataset. All runs use identical hyperparameters with $\mathcal{L}_{fsa} + \mathcal{L}_{ppe} + \mathcal{L}_{mcu}$.

Random Seed	mR (%)
Seed = 42	67.96
Seed = 123	67.93
Seed = 3407	68.03
Mean $\pm$ Std	67.97 $\pm$ 0.04

The minimal variance ($\sigma = 0.04\%$) confirms that the optimization landscape of CGLSA is well-conditioned: neither the momentum-based centroid updates (GSCA) nor the distribution-adaptive scaling (LFFA) exhibits sensitivity to initialization randomness. This stability is particularly important for practical deployment in remote sensing applications where reproducible results are essential.

5.5. Robustness Under Label Noise

Real-world remote sensing datasets may contain annotation errors due to ambiguous scene boundaries or crowd-sourced labeling. To assess the robustness of CGLSA under noisy supervision, we conducted experiments where a fraction of training labels were randomly corrupted (i.e., replaced with incorrect category labels). Table 7 reports the results under 10% and 20% label noise on the RSITMD dataset.

Table 7. Performance under random label noise on the RSITMD dataset. The model maintains competitive retrieval performance under moderate corruption.

Label Noise	mR (%)	Δ vs. Clean
0% (Clean)	68.12	
10%	67.94	−0.18
20%	68.06	−0.06

The results demonstrate remarkable robustness: even under 20% label corruption, the mR degradation is negligible (-0.06%). This resilience can be attributed to two design choices: (1) the momentum-based centroid update in GSCA acts as a natural noise filter since incorrect labels affect only individual updates, while the exponential moving average preserves the majority-correct category topology; (2) the Bayesian prior in PPE down-weights unlikely category assignments, implicitly correcting for mislabeled samples during posterior computation.

5.6. Dynamic Scale Type Analysis

To validate the effectiveness of the proposed exponential dynamic scaling mechanism, we compared three scale adjustment strategies: (1) *fixed*: a constant scale factor γ_{base} throughout training; (2) *linear*: a linearly decaying scale from $3\gamma_{base}$ to γ_{base} based on d_{sep} ; and (3) *dynamic*: the proposed exponential formulation in Equation (13). Table 8 reports the results on the RSITMD dataset.

Table 8. Comparison of different scale adjustment strategies on the RSITMD dataset.

Scale Type	mR (%)
Fixed	68.08
Linear	68.02
Dynamic (Ours)	68.07

All three strategies achieve comparable performance in this short-training validation, indicating that the scaling mechanism provides a robust foundation regardless of the specific decay profile. The dynamic strategy is preferred for longer training schedules where its smooth exponential adaptation prevents the oscillatory behavior that fixed scaling may exhibit during late-stage optimization when positive–negative separation is large.

5.7. Qualitative Experiments

To visually validate the cross-modal alignment precision, Figures 5 and 6 illustrate bidirectional retrieval examples compared against the baseline. In T2I scenarios involving complex compositional queries (e.g., “circular stadium, green field, and urban buildings”), the baseline frequently retrieves partial matches or is distracted by irrelevant background clutter. In contrast, CGLSA accurately isolates the composite subjects, consistently ranking the ground-truth imagery at the top-1 position.

(1) Qualitative analysis on the RSICD dataset: Figure 5 illustrates the retrieval results for both T2I and I2T tasks on the RSICD dataset. In the T2I task, regarding the first text query, the proposed method accurately identifies the image that aligns with the core features of “circular stadium, green field, and urban buildings,” ranking it at the top-1 position. In contrast, the baseline model retrieves the matching image only at the third rank, highlighting the proposed method’s superiority in precise target localization. For the second query, while the baseline retrieves generic bridge imagery, it fails to rank the ground truth at the top position. Conversely, the proposed method successfully prioritizes the image encapsulating comprehensive elements—“a bridge spanning a river, vehicles on the bridge, and buildings on the banks”—at the top-1 position, thereby significantly enhancing semantic relevance. Collectively, these results underscore the robust capability of the proposed method in optimizing both image–text matching accuracy and ranking quality.

Similarly, in the I2T task, the proposed method exhibits superior visual semantic understanding. In the first instance (an airport scene), the baseline model retrieves noisy captions unrelated to the visual content, such as “residential area” and “green square park.” Conversely, the proposed method accurately captures fine-grained details, assigning top ranks to descriptions such as “curved airport terminal” and “planes are docked.” In the second scenario (a river bridge), the baseline fails to identify the primary subject, incorrectly placing a “railway track” description at the top-1 rank. In contrast, the proposed method not only correctly identifies the bridge at the top rank but also retrieves a diverse set of relevant captions covering “cars moving,” “small boats,” and the “river scene.” These findings confirm that the proposed method effectively bridges the semantic gap, ensuring that the retrieved texts provide a comprehensive and accurate description of the visual input.

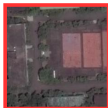
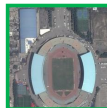
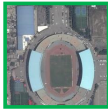
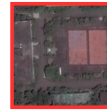
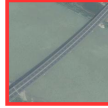
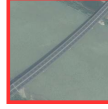
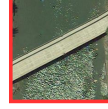
(a) Text-to-Image Retrieval (T2I)		
Query: A modern circular stadium in the city.	Baseline	    
	Ours	    
Query: A bridge with cars crossing a river in the city.	Baseline	    
	Ours	    
(b) Image-to-Text Retrieval (I2T)		
	Baseline 1. Many planes are parked around the airport terminal. 2. A long straight runway with planes taking off. 3. Large parking lots are next to the airport building. 4. Dense residential area with many small houses. 5. Green square park with trees and grass.	Ours 1. Aerial view of a curved airport terminal. 2. Planes are docked at the gates of the terminal. 3. A busy shipping port with many containers. 4. Airport scene with roads and parking lots. 5. The terminal building has many planes nearby.
	Baseline 1. A railway track runs through the green forest. 2. A bridge spans across the river with cars. 3. Large cargo ships sailing in the deep ocean. 4. Cars are driving on the bridge over the water. 5. A straight highway surrounded by green trees.	Ours 1. A bridge connects two sides of the river. 2. Cars are moving on the bridge over the river. 3. Small boats are floating in the water nearby. 4. River scene with a bridge and city buildings. 5. A busy bridge with traffic crossing the water.

Figure 5. Qualitative visualization of bidirectional retrieval results on the RSICD dataset. Matched and mismatched images (or texts) are indicated in green and red, respectively.

(2) Qualitative Analysis on the RSITMD Dataset: Figure 6 presents the qualitative retrieval results for both Text-to-Image (T2I) and Image-to-Text (I2T) tasks on the RSITMD dataset. In the T2I task, regarding the second text query—“multiple oil tanks near buildings, surrounded by green vegetation”—the essential semantic components comprise “buildings,” “multiple oil tanks,” and “green vegetation.” The baseline model fails to capture the critical attribute of “multiple oil tanks,” frequently retrieving images depicting either a solitary tank or missing the object entirely, alongside indistinct representations of vegetation. In contrast, the top-ranked result from the proposed method aligns precisely with the query, depicting multiple oil tanks adjacent to buildings within a vegetated environment. Although subsequent retrievals may exhibit slight variations in tank quantity, they consistently preserve the semantic integrity of core elements without introducing irrelevant scenes. These results demonstrate the superior fine-grained semantic matching capability of the proposed method, particularly in processing queries containing numerical constraints.

In the I2T task, regarding the first image query, the proposed method retrieves descriptions such as “a football field with a central circular area surrounded by two baseball fields and a runway.” These captions precisely capture core visual constituents (e.g., “football field,” “circular area,” and “baseball fields”), achieving high alignment with the image content. Conversely, the baseline model generates vague descriptions like “large meadow with markers and dirt,” with the majority of the top-5 results failing to identify the critical semantic combination of “football and baseball fields.” Similarly, for the second image, the proposed method consistently captures key information such as “large bridge,” “river,” and “green boats.” In contrast, baseline outputs (e.g., “objects spanning across the water”)

deviate markedly from the actual scene content. These comparisons confirm that the proposed method significantly outperforms the baseline in image–text matching precision, exhibiting superior efficacy in identifying and matching salient visual features within complex remote sensing imagery.

(a) Text-to-Image Retrieval (T2I)		
Query: Many planes are at the airport, with roads, grass, and buildings nearby.	Baseline	    
	Ours	    
Query: There are many oil storage tanks and grass near the building.	Baseline	    
	Ours	    
(b) Image-to-Text Retrieval (I2T)		
	Baseline 1. A big grass field with some mud. 2. A green field for playing. 3. A football field, two baseball fields. 4. Open ground with round and square shapes. 5. Ground with grass and dirt.	Ours 1. A football field, a track, and two baseball fields. 2. Green football field, two brown baseball fields. 3. A football field and a baseball field with lines. 4. Next to the football field is a baseball field. 5. The baseball field is next to a swimming pool.
	Baseline 1. The forest has a railway track. 2. There is a bridge over the water 3. The ships are driving on the street. 4. Some cars are on the bridge. 5. This place has a road and trees.	Ours 1. The bridge goes across the river. 2. Some cars drive on the bridge. 3. The boats are near the bridge. 4. River scene with a bridge and city buildings. 5. A busy bridge with traffic crossing the water.

Figure 6. Qualitative visualization of bidirectional retrieval results on the RSITMD dataset. Matched and mismatched images (or texts) are indicated in green and red, respectively.

Finally, to visually substantiate the cross-modal alignment precision achieved by the proposed method, we employed the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm for high-dimensional data visualization. During the training phase, visual and textual features were extracted at distinct stages (specifically, at the 1st, 30th, and 50th epochs) and projected into a lower-dimensional manifold. This process facilitates an analysis of the distributional evolution of cross-modal representations within the feature space throughout the training trajectory. The corresponding visualization results are presented in Figure 7, where blue and red data points correspond to text and image features, respectively.

As depicted in the left column of Figure 7 (epoch 1), the visual and textual feature spaces exhibit a pronounced separation with distinct clustering, indicating that the representations of these heterogeneous modalities are disjoint at the initialization stage. By epoch 30 (middle column of Figure 7), a significant distributional shift is observed, wherein image and text features demonstrate emergent alignment, reflecting the progressive enhancement of cross-modal learning. In the advanced training stage (epoch 50; right column of Figure 7), the feature distribution evolves further, characterized by a highly compact alignment between the modalities. These empirical results substantiate that the proposed model successfully captures intrinsic cross-modal correlations and achieves deep

semantic alignment, thereby ensuring robust performance in remote sensing image–text retrieval tasks.

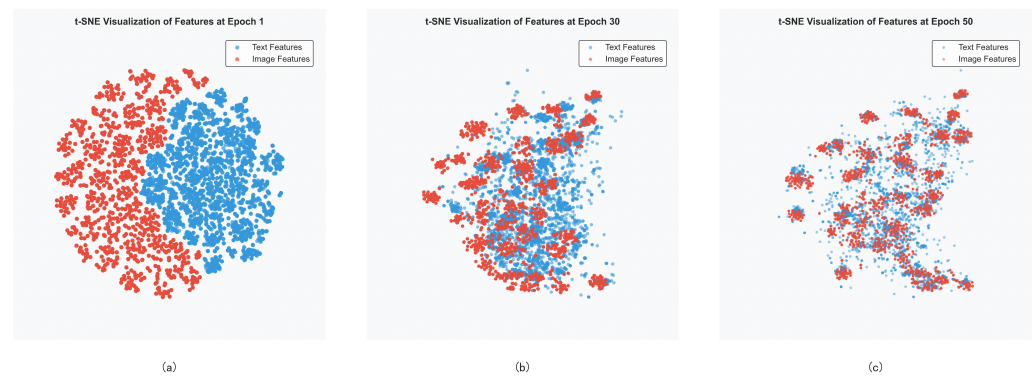


Figure 7. Visualization of feature distributions using t-SNE during the training process. (a) Initial stage at Epoch 1. (b) Intermediate stage at Epoch 30. (c) Final stage at Epoch 50. The blue points represent text features, and the red points represent image features.

5.8. Visual–Semantic Imbalance and Localization

Remote sensing images typically feature cluttered backgrounds with small target objects embedded in large homogeneous regions. Unlike object-centric natural images, the informative region of a remote sensing scene may occupy only a small fraction of the total image area, creating a fundamental visual–semantic imbalance. The proposed CGLSA framework addresses this challenge through complementary mechanisms: the GSCA module regularizes the global category topology to prevent the embedding space from being dominated by background-heavy categories, while the LFFA module’s hard-negative mining concentrates optimization effort on the most discriminative local features, effectively prioritizing semantically informative regions over uninformative background content.

5.9. Handling of Complex Spatial Representations

The current framework focuses on RGB remote sensing imagery and addresses cross-modal alignment through category-aware global and local optimization mechanisms. While multi-spectral or hyperspectral imagery introduces additional channels that encode spectral signatures (e.g., vegetation indices; mineral composition), the proposed architecture operates on the pooled feature representations from a pre-trained ViT backbone, which inherently captures spatial relationships through self-attention. Extension to multi-spectral inputs would require modifying the visual encoder’s input layer while preserving the GSCA/LFFA alignment modules, which operate modality-agnostically on the embedding space.

5.10. Semantic Ambiguity in Text Queries

Text queries in remote sensing retrieval often contain inherent ambiguity—for example, “buildings near water” could match coastal urban areas, lakeside resorts, or riverside warehouses. Category-aware supervision mitigates this ambiguity by anchoring the text embedding to its corresponding scene category centroid, effectively constraining the semantic interpretation space. Additionally, the dynamic similarity regulation in LFFA provides adaptive margin enforcement that separates visually similar but semantically distinct categories, further reducing retrieval errors caused by ambiguous textual descriptions.

6. Conclusions

This article introduces CGLSA, a novel framework designed to resolve the persistent bottlenecks of modality gap, boundary ambiguity, and hard-sample dominance in remote

sensing image–text retrieval. The primary conceptual contribution lies in demonstrating that explicit category-aware learning—a paradigm pivotal in classification but historically underutilized in retrieval—provides a highly effective mechanism for structuring the joint cross-modal embedding space.

By delegating the alignment task to two cascaded modules, CGLSA harmonizes macro-level structure with micro-level precision. The GSCA module aligns category centroids and incorporates probabilistic priors to construct a disentangled global topological manifold. Operating within this regularized search space, the LFFA module leverages statistical dynamic scaling and explicit positive masking to sharpen instance-level cross-modal similarities between pooled image and text embeddings. Comprehensive evaluations across three benchmark datasets validate that this synergistic approach significantly outperforms existing instance-level contrastive methods in both accuracy and feature robustness.

While CGLSA establishes a new performance state-of-the-art, its reliance on the GSCA module inherently necessitates accurate category labels, which confines the current architecture to closed-set retrieval scenarios. We have not yet stress-tested limited-label, noisy-label, pseudo-label, or unseen-category regimes; such robustness analyses are important practical extensions but would multiply experimental conditions beyond the controlled comparisons already reported. Future investigations will focus on semi-supervised or open-vocabulary settings, potentially utilizing unsupervised clustering or vision-language foundation models to generate pseudo-centroids for unseen classes, thereby enhancing applicability to large-scale, weakly annotated Earth observation archives.

Author Contributions: Conceptualization, D.H. and H.Z.; methodology, D.H. and H.Z.; software, D.H.; validation, D.H.; formal analysis, D.H.; investigation, D.H.; resources, H.Z.; data curation, D.H.; writing—original draft preparation, D.H.; writing—review and editing, H.Z.; visualization, D.H.; supervision, H.Z.; project administration, H.Z.; funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC), grant number 42501502.

Data Availability Statement: The data presented in this study are derived from publicly available sources. No new data were created or analyzed.

Acknowledgments: The authors wish to thank all the reviewers who helped improve this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RSITR	Remote sensing image–text retrieval
CLIP	Contrastive language–image pre-training
GSCA	Global semantic collaborative alignment
LFFA	Local fine-grained feature alignment
MCU	Modal center unit
PPE	Probability prior enhancement
FSA	Fine-grained similarity alignment
<i>mR</i>	Mean recall
<i>R@K</i>	Recall at rank <i>K</i>

References

1. Zheng, C.; Song, N.; Zhang, R.; Huang, L.; Wei, Z.; Nie, J. Scale-semantic joint decoupling network for image-text retrieval in remote sensing. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 4.
2. Yuan, Z.; Zhang, W.; Tian, C.; Mao, Y.; Zhou, R.; Wang, H.; Fu, K.; Sun, X. MCRN: A multi-source cross-modal retrieval network for remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 10731. [\[CrossRef\]](#)
3. Tang, X.; Wang, Y.; Ma, J.; Zhang, X.; Liu, F.; Jiao, L. Interacting-enhancing feature transformer for cross-modal remote-sensing image and text retrieval. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5611715. [\[CrossRef\]](#)
4. Aparna, H.; Hati, A. *APMSA: Crossmodal Remote Sensing Image Retrieval using Attention Pooling and Multimodal Semantic Alignment*; ACM International Conference Proceeding Series; Association for Computing Machinery: New York, NY, USA, 2024.
5. Zhang, Z.; Zhao, T.; Guo, Y.; Yin, J. RS5M and GeoRSCLIP: A large-scale vision- language dataset and a large vision-language model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5642123. [\[CrossRef\]](#)
6. Silva, J.D.; Magalhaes, J.; Tuia, D.; Martins, B. Multilingual vision-language pre-training for the remote sensing domain. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, Atlanta, GA, USA, 29 October–1 November 2024; pp. 220–232.
7. Li, X.; Xu, F.; Zhang, J.; Yu, A.; Lyu, X.; Gao, H.; Zhou, J. Dual-domain decoupled fusion network for semantic segmentation of remote sensing images. *Inf. Fusion* **2025**, *124*, 103359. [\[CrossRef\]](#)
8. Yang, R.; Wang, S.; Han, Y.; Li, Y.; Zhao, D.; Quan, D.; Guo, Y.; Jiao, L.; Yang, Z. Transcending Fusion: A Multiscale Alignment Method for Remote Sensing Image–Text Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4709217. [\[CrossRef\]](#)
9. Yuan, Z.; Zhang, W.; Tian, C.; Rong, X.; Zhang, Z.; Wang, H.; Fu, K.; Sun, X. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620616. [\[CrossRef\]](#)
10. Zhang, W.; Li, J.; Li, S.; Chen, J.; Zhang, W.; Sun, X. Hypersphere-based remote sensing cross-modal text–image retrieval via curriculum learning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5621815. [\[CrossRef\]](#)
11. Yu, H.; Yao, F.; Lu, W.; Liu, N.; Li, P.; You, H.; Sun, X. Text-image matching for cross-modal remote sensing image retrieval via graph neural network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 812–824. [\[CrossRef\]](#)
12. Yuan, Z.; Zhang, W.; Rong, X.; Li, X.; Chen, J.; Wang, H.; Fu, K.; Sun, X. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5612819. [\[CrossRef\]](#)
13. Zhang, S.; Li, Y.; Mei, S. Exploring uni-modal feature learning on entities and relations for remote sensing cross-modal text-image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5626317. [\[CrossRef\]](#)
14. Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; Shao, J. CAMP: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 5763–5772.
15. Cheng, Q.; Zhou, Y.; Fu, P.; Xu, Y.; Zhang, L. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4284–4297. [\[CrossRef\]](#)
16. Pan, J.; Ma, Q.; Bai, C. *A Prior Instruction Representation Framework for Remote Sensing Image-Text Retrieval*; ACM: New York, NY, USA, 2023; pp. 611–620.
17. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. *VSE++: Improving Visual-Semantic Embeddings with Hard Negatives*; BMVA Press: Newcastle, UK, 2018.
18. Yang, Y.; Tang, X.; Zhang, X.; Ma, J.; Liu, F.; Jia, X.; Jiao, L. Semi-supervised multiscale dynamic graph convolution network for hyperspectral image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 6806–6820. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Tu, R.C.; Ji, Y.; Jiang, J.; Kong, W.; Cai, C.; Zhao, W.; Wang, H.; Yang, Y.; Liu, W. Global and local semantic completion learning for vision-language pre-training. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 11065–11079. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Nie, J.; Zheng, C.; Wang, C.; Zuo, Z.; Lv, X.; Yu, S.; Wei, Z. Scale-relation joint decoupling network for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5412812. [\[CrossRef\]](#)
21. Chen, Y.; Huang, J.; Li, X.; Xiong, S.; Lu, X. Multiscale salient alignment learning for remote-sensing image–text retrieval. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4700413. [\[CrossRef\]](#)
22. Yang, L.; Zhou, T.; Ma, W.; Du, M.; Liu, L.; Li, F.; Zhao, S.; Wang, Y. Remote Sensing Image-Text Retrieval with Implicit-Explicit Relation Reasoning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4511411. [\[CrossRef\]](#)
23. Wang, M.; Guo, J.; Song, B.; Su, K. Graph-based hierarchical semantic consistency network for remote sensing image–text retrieval. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 15334–15346. [\[CrossRef\]](#)
24. Zhang, Y.; Ji, Z.; Meng, C.; Pang, Y. iEBAKER: Improved remote sensing image-text retrieval framework via eliminate before align and keyword explicit reasoning. *Expert Syst. Appl.* **2026**, *296*, 128968. [\[CrossRef\]](#)
25. Jia, M.; Tang, L.; Chen, B.C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.N. Visual prompt tuning. In *Computer Vision–ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 13693, pp. 709–727.

26. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the Machine Learning Research, Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 12888–12900.
27. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the Machine Learning Research, Virtual, 18–24 July 2021; Volume 139, pp. 8748–8763.
28. Sun, T.; Zheng, C.; Li, X.; Gao, Y.; Nie, J.; Huang, L.; Wei, Z. Strong and weak prompt engineering for remote sensing image-text cross-modal retrieval. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 6968–6980. [[CrossRef](#)]
29. Zheng, C.; Wen, Q.; Li, X.; Yang, C.; Nie, J.; Guo, Y.; Qian, Y. Whole Semantic Sparse Coding Network for Remote Sensing Image–Text Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5406913. [[CrossRef](#)]
30. Cai, L.; Zhu, L.; Zhang, H.; Zhu, X. DA-GAN: Dual attention generative adversarial network for cross-modal retrieval. *Future Internet* **2022**, *14*, 43. [[CrossRef](#)]
31. Su, Y.; Zhu, D.; Dong, Z.; Lin, Q.; Liu, L. FAME: Fusion of Alignment and Multiview Enhancement for Remote Sensing Image–Text Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 3001615. [[CrossRef](#)]
32. Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; Sun, X. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4404119. [[CrossRef](#)]
33. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [[CrossRef](#)]
34. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 6–8 July 2016; pp. 1–5.
35. Ji, Z.; Meng, C.; Zhang, Y.; Pang, Y.; Li, X. Knowledge-aided momentum contrastive learning for remote-sensing image text retrieval. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5625213. [[CrossRef](#)]
36. Yuan, Y.; Zhan, Y.; Xiong, Z. Parameter-Efficient Transfer Learning for Remote Sensing Image–Text Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5619014. [[CrossRef](#)]
37. Wan, Z.; Zhao, E.; Nie, J.; Zhang, Z.; Wei, Z.; Zheng, N.; Zhao, Y. Frequency spectrum adaptor for remote sensing image–text retrieval. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 18922–18940. [[CrossRef](#)]
38. Zheng, C.; Li, X.; Liang, X.; Huang, L.; Du, S.; Nie, J.; Dong, J. Cross-modal progressive perspective matching network for remote sensing image-text retrieval. *IEEE Trans. Multimed.* **2025**, *27*, 3966–3978. [[CrossRef](#)]
39. Yang, J.; Li, S.; Zhao, M. Parameter-Efficient Reparameterization Tuning for Remote Sensing Image–Text Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4702315. [[CrossRef](#)]
40. Zheng, C.; Nie, J.; Yin, B.; Li, X.; Qian, Y. Frequency- and Spatial-Domain Saliency Network for Remote Sensing Cross-Modal Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5403313. [[CrossRef](#)]
41. Zhao, Z.; Miao, X.; Liu, L.; Xu, X.; Liu, Y.; Hu, J.; Min, B.; Gao, Y.; Pharksuwan, K. Sparse-Guided Partial Dense for Cross-Modal Remote Sensing Image–Text Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5617213. [[CrossRef](#)]
42. Chen, X.; Zheng, X.; Lu, X. Context-Aware Local–Global Semantic Alignment for Remote Sensing Image–Text Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5617312. [[CrossRef](#)]
43. Li, S.; Ji, H.; Liu, F.; Min, X.; Huang, X.; Wang, J.; Sun, L.; Li, L. Fine-Grained Visual-Language Alignment for Remote Sensing Image–Text Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5636614. [[CrossRef](#)]
44. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 499–515.
45. Deng, J.; Guo, J.; Zafeiriou, S. ArcFace: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
46. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 18661–18673.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.