Supplementary Material: Robust Remote Sensing Image–Text Retrieval with Noisy Correspondence

Qiya Song[1], Yiqiang Xie[1], Yuan Sun[2]*, Renwei Dian[3], Xudong Kang[3]
[1]Hunan Normal University, Changsha, China 410081
[2]Sichuan University, Chengdu, China 610044
[3]Hunan University, Changsha, China 410082
sqyunb@hnu.edu.cn, yiqiang_xie@hunnu.edu.cn, sunyuan_work@163.com

This supplementary material provides additional details of the Robust Remote Sensing Image–Text Retrieval (RRSITR) framework to facilitate a deeper understanding of the proposed method. Section 1 presents a theoretical analysis of RRSITR. Section 2 details the training process with pseudocode for RRSITR. Section 3 introduces the datasets used in this paper. Section 4 provides a detailed introduction to the comparison methods. Section 5 presents the experiment results for the three datasets. Section 6 presents an analysis of the parameters. Section 7 presents ablation study results. Section 8 shows a visualization of the weights. Section 9 presents some identified noisy sample pairs from three benchmark datasets. Additionally, Section 10 presents the analysis of qualitative results.

Contents

1. Theoretical Analysis

For clarity, the following notations are used throughout this paper. Θ denotes the network parameters. ℓ_i denotes the sum of the global and local contrastive losses for the i-th sample pair. γ_1 and γ_2 are thresholds. $\mathcal{L}_{soft}$ refers to a robust triplet loss. λ_1 and λ_2 are the balancing factors.

In the main text, we derive the optimal weight solution as follows:

$$w_i^* = \begin{cases} \cos\left(\frac{\pi}{2} \cdot \frac{\ell_i}{\gamma}\right), & \text{if } \ell_i < \gamma, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We present some important properties of the objective function, which may help elucidate the rationale behind RRSITR.

Prior to analyzing RRSITR, for simplicity, ℓ_i is denoted as ℓ. We introduce two key variables: $F_\gamma(\ell)$ and $Q_\gamma(\Theta|\Theta^*)$. Specifically, $F_\gamma(\ell)$ is defined as the integral of $w^*(\ell,\gamma)$ with respect to ℓ, given by:

$$F_{\gamma_1}(\ell) = \int_0^\ell w^*(l,\gamma_1)dl = \begin{cases} \frac{2\gamma_1}{\pi}\sin\left(\frac{\pi}{2}\cdot\frac{\ell}{\gamma_1}\right), & \text{if } \ell < \gamma_1, \\ \frac{2\gamma_1}{\pi}, & \text{if } \ell \geq \gamma_1, \end{cases}$$

$$F_{\gamma_2}(\ell) = \int_{\gamma_1}^\ell w^*(l,\gamma_2)dl$$
$$= \begin{cases} \frac{2\gamma_2}{\pi}\left(\sin\left(\frac{\pi}{2}\cdot\frac{\ell}{\gamma_2}\right) - \sin\left(\frac{\pi}{2}\cdot\frac{\gamma_1}{\gamma_2}\right)\right), & \text{if } \gamma_1 \leq \ell < \gamma_2, \\ \frac{2\gamma_2}{\pi}\left(1 - \sin\left(\frac{\pi}{2}\cdot\frac{\gamma_1}{\gamma_2}\right)\right), & \text{if } \ell \geq \gamma_2. \end{cases} \tag{2}$$

Then we define $Q_\gamma(\Theta|\Theta^*)$ as the first-order expansion of $F_\gamma(\ell(\Theta))$ at $\ell(\Theta^*)$:

$$Q_{\gamma_1}(\Theta|\Theta^*) = F_{\gamma_1}(\ell(\Theta^*)) + w^*(\ell(\Theta^*),\gamma_1)\cdot(\ell(\Theta) - \ell(\Theta^*)), \tag{3}$$

$$Q_{\gamma_2}(\Theta|\Theta^*) = F_{\gamma_2}(\ell(\Theta^*)) + w^*(\ell(\Theta^*),\gamma_2)\cdot(\ell(\Theta) - \ell(\Theta^*)). \tag{4}$$

As indicated by $\sum_{i=1}^n F_{\gamma_1}(\ell_i(\Theta)) + \lambda_1 F_{\gamma_2}(\ell_i(\Theta)) + \lambda_2\mathcal{L}_{soft}(\Theta)$, we observe that the proposed objective function is analytically challenging, as it requires simultaneously solving for the self-paced weights w and the network parameters Θ. In the following proposition, we demonstrate that RRSITR actually optimizes a more simplified implicit objective function, where the self-paced weights w are completely eliminated.

PROPOSITION 1 (Implicit Objective Function). For a fixed γ_1 and γ_2, the alternating optimization strategy (AOS) used to minimize $\sum_{i=1}^n F_{\gamma_1}(\ell_i(\Theta)) + \lambda_1 F_{\gamma_2}(\ell_i(\Theta)) + \lambda_2\mathcal{L}_{soft}(\Theta)$ is equivalent to the application of the Majorization–Minimization [5] algorithm to this function.

*Corresponding authors.

PROOF. The following holds:

$$\begin{aligned} F_{\gamma_1}(\ell_i(\Theta)) &\leq Q^{(i)}_{\gamma_1}(\Theta|\Theta^*), \\ F_{\gamma_2}(\ell_i(\Theta)) &\leq Q^{(i)}_{\gamma_2}(\Theta|\Theta^*). \end{aligned} \tag{5}$$

Additionally, the following inequality holds:

$$\begin{aligned} &\sum_{i=1}^{n} F_{\gamma_1}(\ell_i(\Theta)) + \lambda_1 F_{\gamma_2}(\ell_i(\Theta)) + \lambda_2 \mathcal{L}_{soft}(\Theta) \\ &\quad \leq \sum_{i=1}^{n} Q^{(i)}_{\gamma_1}(\Theta|\Theta^*) + \lambda_1 Q^{(i)}_{\gamma_2}(\Theta|\Theta^*) + \lambda_2 \mathcal{L}_{soft}(\Theta). \end{aligned} \tag{6}$$

Therefore, the Majorization-Minimization algorithm can be employed to minimize the objective function, with the key step being the minimization of the surrogate function $\sum_{i=1}^{n} F_{\gamma_1}(\ell_i(\Theta)) + \lambda_1 F_{\gamma_2}(\ell_i(\Theta)) + \lambda_2 \mathcal{L}_{soft}(\Theta)$. The steps for the k-th iteration are as follows:

Majorization Step: In this step, fix Θ as Θ^k. To obtain the surrogate function, it is necessary to compute $w_i(\ell_i(\Theta^k), \gamma_1)$ and $w_i(\ell_i(\Theta^k), \gamma_2)$ by solving the following problem:

$$\begin{aligned} w_i(\ell_i(\Theta^k), \gamma_1) &= \underset{w_i \in [0,1]}{\operatorname{argmin}} \, \mathcal{L}_i(\Theta^k, \gamma_1), \\ w_i(\ell_i(\Theta^k), \gamma_2) &= \underset{w_i \in [0,1]}{\operatorname{argmin}} \, \mathcal{L}_i(\Theta^k, \gamma_2), \end{aligned} \tag{7}$$

this computation follows procedure for updating w in AOS.

Minimization step: In this step, fixing w_i as w_i^*, update Θ by minimizing the surrogate function:

$$\Theta^{k+1} = \arg\min_{\Theta} \sum_{i=1}^{n} Q^{(i)}_{\gamma_1}(\Theta|\Theta^*) + \lambda_1 Q^{(i)}_{\gamma_2}(\Theta|\Theta^*) + \lambda_2 \mathcal{L}_{soft}(\Theta), \tag{8}$$

this process is consistent with the step of updating Θ in AOS.

It can be readily inferred that our AOS strategy is equivalent to the aforementioned optimization-minimization algorithm. Moreover, $\sum_{i=1}^{n} F_{\gamma_1}(\ell_i(\Theta)) + \lambda_1 F_{\gamma_2}(\ell_i(\Theta)) + \lambda_2 \mathcal{L}_{soft}(\Theta)$ can be regarded as the underlying objective function we practically optimize, while $F_{\gamma_1}(\ell_i(\Theta))$ and $F_{\gamma_2}(\ell_i(\Theta))$ represent the underlying loss function for the i-th sample.

According to Equation (2), the aforementioned proposition indicates that RRSITR minimizes an implicit loss function from which w is eliminated. Building upon this derivation, we now propose a subsequent proposition to demonstrate the robustness of RRSITR to hard samples.

PROPOSITION 2 (Robustness). For notational simplicity, denote $\ell_i(\Theta)$ by ℓ_i. Assume that $\min_k \ell_k \geq B_1$ and $\min_t \ell_t \geq B_2$ for some constants $B_1 < \infty$ and $\gamma_1 \leq B_2 < \infty$. Then, for any two distinct instances (i, j) in the training dataset D, the following holds:

For the function $F_{\gamma_1}(\ell)$, we have:

$$|F_{\gamma_1}(\ell_i) - F_{\gamma_1}(\ell_j)| \leq \kappa_1 \cdot |\ell_i - \ell_j|. \tag{9}$$

For $F_{\gamma_2}(\ell)$, when both $\ell_i \geq \gamma_1$ and $\ell_j \geq \gamma_1$, we have:

$$|F_{\gamma_2}(\ell_i) - F_{\gamma_2}(\ell_j)| \leq \kappa_2 \cdot |\ell_i - \ell_j|, \tag{10}$$

where the constants κ_1 and κ_2 are defined as:

$$\begin{aligned} \kappa_1 &= \begin{cases} \cos\left(\frac{\pi}{2} \cdot \frac{B_1}{\gamma_1}\right), & \text{if } B_1 < \gamma_1, \\ 0, & \text{if } B_1 \geq \gamma_1, \end{cases} \\ \kappa_2 &= \begin{cases} \cos\left(\frac{\pi}{2} \cdot \frac{B_2}{\gamma_2}\right), & \text{if } \gamma_1 \leq B_2 < \gamma_2, \\ 0, & \text{if } B_2 \geq \gamma_2. \end{cases} \end{aligned} \tag{11}$$

PROOF. Let $a = \min\{\ell_i, \ell_j\}$ and $b = \max\{\ell_i, \ell_j\}$. Since $F_{\gamma_1}(\ell)$ and $F_{\gamma_2}(\ell)$ are continuously differentiable on their respective domains, by the mean value theorem, there exists $\xi_1, \xi_2 \in [a, b]$ such that:

$$\begin{aligned} |F_{\gamma_1}(\ell_i) - F_{\gamma_1}(\ell_j)| &= F'_{\gamma_1}(\xi_1) \cdot |\ell_i - \ell_j|, \\ |F_{\gamma_2}(\ell_i) - F_{\gamma_2}(\ell_j)| &= F'_{\gamma_2}(\xi_2) \cdot |\ell_i - \ell_j|. \end{aligned} \tag{12}$$

Based on $F_{\gamma_1}(\ell)$ and $F_{\gamma_2}(\ell)$ as defined in Equation 2. For $F_{\gamma_1}(\ell)$, we have:

$$F'_{\gamma_1}(\ell) = \begin{cases} \cos\left(\frac{\pi}{2} \cdot \frac{\ell}{\gamma_1}\right), & \text{if } \ell < \gamma_1, \\ 0, & \text{if } \ell \geq \gamma_1. \end{cases} \tag{13}$$

For $F_{\gamma_2}(\ell)$, we have:

$$F'_{\gamma_2}(\ell) = \begin{cases} \cos\left(\frac{\pi}{2} \cdot \frac{\ell}{\gamma_2}\right), & \text{if } \gamma_1 \leq \ell < \gamma_2, \\ 0, & \text{if } \ell \geq \gamma_2. \end{cases} \tag{14}$$

Since $\xi_1 \geq B_1$ and $\xi_2 \geq B_2$, we have $F'_{\gamma_1}(\xi_1) \leq \kappa_1$ and $F'_{\gamma_2}(\xi_2) \leq \kappa_2$, where κ_1 and κ_2 are defined using B_1 and B_2 respectively.

Therefore, the inequalities hold for any two distinct instances (i, j) satisfying the respective conditions.

According to Proposition 2, the transformed loss functions $F_{\gamma_1}(\ell)$ and $F_{\gamma_2}(\ell)$ demonstrate stronger robustness toward hard samples with large loss values compared to the original loss function $\ell(\cdot)$. Specifically, when considering a hard sample i and an easy sample j, the loss discrepancy in RRSITR, measured by $|F_{\gamma_1}(\ell_i) - F_{\gamma_1}(\ell_j)|$ and $|F_{\gamma_2}(\ell_i) - F_{\gamma_2}(\ell_j)|$, becomes significantly smaller than the original loss discrepancy $|\ell_i - \ell_j|$, since $\kappa_1 < 1$ and $\kappa_2 < 1$. This result confirms that both $F_{\gamma_1}(\ell(\cdot))$ and $F_{\gamma_2}(\ell(\cdot))$ achieve enhanced robustness by substantially reducing sensitivity to large loss values.

2. Training Procedure

To clearly illustrate the training process of the RRSITR framework. Algorithm 1 provides the corresponding pseudocode, which details the key operations and data flow.

3. Dataset Details

Three benchmark datasets are used to evaluate the proposed RRSITR, including:

- RSITMD [11]: The RSITMD comprises 4,743 remote sensing images belonging to 32 categories. Each remote sensing image has a size of 256 × 256 pixels and is accompanied by five descriptive texts. These texts provide

fine-grained information for the remote sensing scenes. We follow the data partitioning scheme described in [12], which consists of 3,861 training images, 430 validation images, and 452 test images.

- RSICD [4]: The RSICD comprises 10,921 remote sensing images collected from Google Earth that belong to 30 different categories. The sizes of these remote sensing images are 224 × 224 pixels, and their resolutions vary. Each remote sensing image is associated with five texts, resulting in 54,605 description texts. Compared with RSITMD, the similarity of text descriptions in RSICD is higher, which increases the difficulty of cross-modal learning. We follow the data partition in [12] which consists of 8,845 training images, 983 validation images, and 1,093 test images.
- NWPU [1]: The NWPU comprises 31,500 images, each with a pixel resolution of 256 × 256, spanning 45 scene categories. Every image is annotated with five descriptive captions. We follow the data partition in [12] which consists of 25,515 training images, 2,835 validation images, and 3,150 test images.

Algorithm 1 Robust Remote Sensing Image–Text Retrieval

Require: Dataset $\mathcal{D} = \{(I_i, T_i, y_i)\}_{i=1}^{N}$; Hyperparameters $\gamma_1, \gamma_2, \lambda_1, \lambda_2, \sigma, \alpha$
Ensure: Parameters Θ

Initialize Θ with CLIP visual encoder φ and text encoder ϕ;
while not converged **do**
  for each batch $\mathcal{B} \subset \mathcal{D}$ **do**
    Compute $f_v^g, f_v^l = \varphi(I_i)$ and $f_t^g, f_t^l = \phi(T_i)$;
    Compute $S^g = \cos\left(f_v^g, f_t^g\right)$;
    Compute $S^l = \| \cos\left(f_v^l, f_t^l\right) \|_{2,2}$;
    Compute $\mathcal{L}_{info}^{gl} = \mathcal{L}_{info}^{g}(S^g) + \mathcal{L}_{info}^{l}(S^l)$;
    for each sample in $\mathcal{B}$ **do**
      $\ell_i = \mathcal{L}_{info}^{gl}$;
      if $\ell_i < \gamma_1$ **then**
        $w_i = \cos(\frac{\pi}{2} \cdot \frac{\ell_i}{\gamma_1})$;
        Compute the loss $\mathcal{L}_{S1}$;
      else if $\gamma_1 \leq \ell_i < \gamma_2$ **then**
        $w_i = \cos(\frac{\pi}{2} \cdot \frac{\ell_i}{\gamma_2})$;
        Compute the loss $\mathcal{L}_{S2}$;
      else
        $w_i = 0$;
        Compute the loss $\mathcal{L}_{soft}$;
      end if
    end for
    Compute $\mathcal{L}_{overall}$;
    Update Θ using gradient descent;
  end for
end while
return Θ

4. Comparison Methods

To validate the effectiveness of our model, we select ten state-of-the-art remote sensing image-text retrieval models as baseline comparisons, as follows.

- SIRS [13]: The SIRS model utilizes a multitask joint framework for efficient plug-and-play and end-to-end model training.
- MSA [10]: The MSA model separately aligns image-text pairs at multiple scales to improve the ability to learn joint representations that enable effective retrieval.
- AMFMN [11]: The AMFMN model is an asymmetric multimodal feature matching network that allows for adaptation to multiscale feature inputs and dynamic filtering of redundant features.
- HVSA [12]: The HVSA model is a curriculum learning-based hypersphere visual semantic alignment network for advanced retrieval performance.
- SWAN [8]: The SWAN model enhances the perception of remote sensing image scenes to reduce confusion in the semantic space and improve retrieval performance.
- PIR [7]: The PIR model employs a progressive attention encoder to model long-range dependencies, thereby enhancing the characterization of key features and improving retrieval performance.
- KAMCL [3]: The KAMCL model constructs a knowledge augmented learning framework to provide valuable concepts and learn discriminative representations.
- S-CLIP [6]: The S-CLIP model introduces a contrastive learning and language modality-based pseudo-labeling strategy to improve the training of pre-trained models.
- SEMICLIP [2]: The SEMICLIP model leverages a small amount of image-text paired data with a large number of images without textual descriptions to improve the image-text alignment capabilities of pre-trained models.
- CUP [9]: The CUP model introduces prompt tuning into pre-trained models to alleviate the computational burden of optimization.

5. Comparison with the State-of-the-Art

In all experimental tables, the highest scores are highlighted in **bold**, the second-highest scores are marked with underlines, and '(±)' denotes the standard deviation.

As shown in Tab.1, under the 0% noise condition, the proposed RRSITR achieves the best performance across all evaluation metrics on all three datasets. These results demonstrate the effectiveness and robustness of RRSITR, indicating that RRSITR not only possesses strong anti-interference capabilities but also maximizes retrieval performance under ideal conditions.

Table 1. Image-text retrieval performance on RSITMD, RSICD, and NWPU under 0% noise ratio.

Dataset	Method	Ref.	Image-to-Text Retrieval			Text-to-Image Retrieval			mR
			R@1	R@5	R@10	R@1	R@5	R@10	
RSITMD	AMFMN	TGRS'22	12.92±1.96	30.49±2.34	43.19±1.74	10.86±1.02	36.31±0.67	56.19±1.21	31.66±1.28
	HVSA	TGRS'23	13.76±2.24	32.17±1.89	45.00±1.33	11.40±0.83	38.86±0.93	57.18±1.16	33.06±1.10
	KAMCL	TGRS'23	13.85±1.66	31.32±2.67	44.65±2.90	12.13±0.78	38.94±1.33	56.58±1.29	32.91±1.50
	SWAN	ICMR'23	14.73±2.28	33.85±0.94	47.17±1.59	11.35±0.49	39.54±0.73	59.60±1.19	34.37±0.80
	PIR	ACMMM'23	17.30±1.16	39.07±1.29	52.26±1.83	13.28±0.54	42.20±0.90	62.58±1.03	37.78±0.81
	S-CLIP	NeurIPS'23	9.16±0.98	32.57±1.20	50.22±0.73	9.47±0.62	31.90±0.81	49.29±1.27	30.44±0.73
	SIRS	TGRS'24	14.07±0.87	32.92±1.28	46.42±1.59	11.66±0.60	40.55±0.54	59.11±0.88	34.12±0.46
	MSA	TGRS'24	15.35±0.75	34.69±1.45	46.28±1.13	13.64±0.55	41.84±1.05	59.39±1.12	35.20±0.59
	SEMICLIP	ICLR'25	11.95±0.34	37.65±2.63	55.84±1.84	11.68±1.38	39.25±1.89	57.66±1.94	35.67±1.19
	CUP	TNNLS'25	20.44±1.20	40.62±1.02	53.98±0.95	16.78±0.47	46.64±0.90	63.80±0.79	40.38±0.62
	RRSITR	Ours	**25.44±0.54**	**46.82±1.86**	**58.98±2.04**	**20.88±0.49**	**53.59±1.74**	**70.90±0.81**	**46.10±1.08**
RSICD	AMFMN	TGRS'22	6.68±0.79	19.32±0.63	30.07±0.86	5.62±0.57	20.44±0.50	34.12±0.69	19.38±0.36
	HVSA	TGRS'23	8.05±1.07	20.90±1.26	31.77±1.10	5.69±0.33	21.54±0.75	34.77±1.20	20.45±0.77
	KAMCL	TGRS'23	9.31±0.34	23.42±0.67	35.85±1.03	6.92±0.45	24.14±0.59	38.22±1.38	22.98±0.67
	SWAN	ICMR'23	7.65±0.91	21.25±1.09	32.66±1.00	6.29±0.23	23.29±0.66	38.46±0.49	21.60±0.52
	PIR	ACMMM'23	9.59±0.65	23.39±0.88	36.27±1.09	7.29±0.45	24.91±0.53	40.53±0.85	23.66±0.47
	S-CLIP	NeurIPS'23	4.71±0.84	19.30±1.92	32.44±1.84	5.31±0.52	19.69±1.27	32.57±2.59	19.00±1.38
	SIRS	TGRS'24	6.90±0.33	20.24±0.43	31.43±0.76	5.43±0.40	21.07±0.33	35.30±0.65	20.06±0.10
	MSA	TGRS'24	6.84±0.60	20.68±0.72	32.04±0.96	6.13±0.42	23.59±0.32	39.35±0.55	21.44±0.33
	SEMICLIP	ICLR'25	7.19±0.65	24.43±0.96	39.62±1.26	7.05±0.47	26.02±1.25	40.07±1.83	24.06±0.90
	CUP	TNNLS'25	11.27±0.55	32.04±1.16	48.01±0.57	9.45±0.48	30.94±1.04	47.45±1.30	29.86±0.48
	RRSITR	Ours	**15.86±0.82**	**35.50±2.11**	**48.54±1.88**	**12.05±0.59**	**34.68±1.66**	**50.41±2.61**	**32.84±1.52**
NWPU	AMFMN	TGRS'22	12.49±0.43	41.14±0.63	59.59±0.60	9.50±0.07	32.15±0.18	48.49±0.20	33.89±0.23
	HVSA	TGRS'23	12.79±0.41	41.13±0.50	59.65±0.80	9.42±0.20	32.26±0.26	48.78±0.13	34.01±0.22
	KAMCL	TGRS'23	20.17±1.02	54.04±0.89	70.83±1.09	11.65±0.44	35.71±0.52	51.70±0.61	40.68±0.69
	SWAN	ICMR'23	12.03±0.24	38.63±0.50	56.97±0.65	8.70±0.13	29.82±0.31	46.06±0.29	32.04±0.28
	PIR	ACMMM'23	22.39±0.45	57.84±0.38	74.61±0.61	12.91±0.10	38.71±0.26	54.59±0.21	43.51±0.23
	S-CLIP	NeurIPS'23	4.30±0.27	16.45±0.78	27.69±0.64	3.65±0.26	14.62±0.13	25.08±0.70	15.30±0.36
	SIRS	TGRS'24	11.00±0.41	36.50±0.25	54.65±0.78	8.46±0.11	29.59±0.18	45.90±0.14	31.02±0.18
	MSA	TGRS'24	7.28±0.11	25.94±0.68	41.66±0.51	7.44±0.17	26.91±0.61	42.70±0.83	25.32±0.17
	SEMICLIP	ICLR'25	8.22±0.54	23.43±0.23	37.26±0.69	6.38±0.21	23.01±0.63	37.12±0.54	22.57±0.22
	CUP	TNNLS'25	9.94±1.21	32.01±3.46	47.35±4.09	5.94±0.53	21.16±1.70	34.22±2.04	25.10±1.66
	RRSITR	Ours	**25.18±0.78**	**59.66±0.87**	**76.06±0.64**	**14.78±0.23**	**40.78±0.34**	**56.30±0.30**	**45.46±0.35**

6. Parameter Analysis

This study investigates the impact of α and σ on retrieval performance. As illustrated in Fig.1, both excessively large and small parameter values result in performance degradation. Specifically, we could observe that the model maintains satisfactory performance in both noisy and noise-free environments when α falls within the range of 0.7 to 1.0 and σ lies in the range of 0.5 to 0.8. This demonstrates that such a parameter configuration effectively balances the contributions of global and local similarity measures, as well as the triplet margin σ, thereby ensuring stable and superior performance under varying data quality conditions.

To effectively categorize samples into clean, ambiguous, and noisy subsets in RRSITR, we first theoretically determine the optimal ranges for parameters γ_1 and γ_2. Based on the inherent cross-entropy properties of the InfoNCE loss, we derive a threshold range $\gamma \in [\ln N, 4 \ln N]$, where N is the batch size.

Specifically, when the model is in a random guessing state with a prediction probability of $1/N$, the expected value of the loss ℓ, which comprises four cross-entropy terms, converges to approximately $4 \ln N$. Samples exhibiting loss values exceeding this theoretical upper bound are statistically highly likely to be noise. Conversely, for high-confidence clean samples, the loss is expected to be significantly lower than this random baseline. Theoretically, reducing the loss to $\ln N$ indicates strong semantic alignment and high confidence in positive pairs. In our implementation with $N = 100$, the parameters $\gamma_1 = 5$ and $\gamma_2 = 18$ align with the theoretical values $\ln N$ and $4 \ln N$.

To investigate the influence of parameters γ_1 and γ_2 on the robustness of the RRSITR, we conduct a sensitivity analysis on their values. As shown in Tab.2, the results indicate that setting γ_1 and γ_2 either too high or too low leads to a decline in model performance. Specifically, the model achieves the best overall performance when γ_1 is set to 5 and γ_2 to 18. Further analysis reveals that moderate values of γ_1 and γ_2 help the model establish a more reasonable discrimination mechanism among clean, ambiguous, and noisy samples, thereby enhancing overall robustness. In contrast, overly small values of γ_1 and γ_2 cause the model to overlook substantial useful information, while excessively large values introduce additional noise and impair generalization capability.

7. Ablation Studies

To investigate the role of each component in the RRSITR model, we conduct ablation studies on the RSITMD dataset with different noise ratios. We compare the proposed RRSITR with four variants. Specifically, #1, #2, #3, and #4 represent the removal of the local contrastive learning module, SPL module, RTL module, and all components, respec-

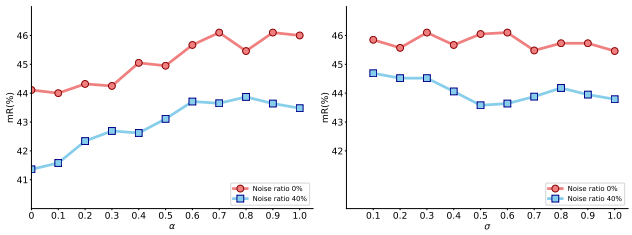

Figure 1. Parameters analysis (i.e., α and σ) on RSITMD.

Table 2. Different γ_1 and γ_2 on RSITMD with 0% noise

γ_1	γ_2	Image-to-Text Retrieval			Text-to-Image Retrieval			mR
		R@1	R@5	R@10	R@1	R@5	R@10	
1	6	1.24	4.03	6.55	0.45	2.49	4.22	3.16
2	9	6.90	14.51	19.78	5.54	15.89	23.20	14.30
3	12	23.18	44.83	58.10	20.81	53.28	70.28	45.08
4	15	24.29	45.31	58.50	**21.45**	**54.18**	**71.51**	45.87
5	18	**25.44**	**46.82**	**58.98**	20.88	53.59	70.90	**46.10**
6	21	24.47	45.31	57.48	20.91	53.62	71.03	45.47

tively. As shown in Tab.3 and Tab.4, we draw the following conclusions: 1) The complete RRSITR framework achieves the best overall performance, which validates the effectiveness of each component. 2) The synergistic interactions among the modules significantly enhance model robustness, enabling the framework to effectively handle noisy correspondence challenges.

Table 3. Ablation studies on RSITMD with 0% noise ratio.

Method	Image-to-Text Retrieval			Text-to-Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
#1	23.14	44.56	57.34	**21.07**	53.35	**71.38**	45.14
#2	23.58	44.96	58.54	20.87	53.40	71.09	45.41
#3	24.60	45.00	58.63	21.06	**53.87**	70.55	45.62
#4	23.19	44.69	**59.07**	20.05	52.59	70.35	44.99
RRSITR	**25.44**	**46.82**	58.98	20.88	53.59	70.90	**46.10**

Table 4. Ablation studies on RSITMD with 20% noise ratio.

Method	Image-to-Text Retrieval			Text-to-Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
#1	23.36	43.23	57.04	**20.92**	52.99	70.05	44.60
#2	22.70	41.90	56.46	19.68	50.74	68.99	43.41
#3	**24.78**	44.56	58.10	20.57	52.26	**70.21**	45.08
#4	23.01	42.61	56.33	18.91	49.94	68.17	43.16
RRSITR	24.60	**46.68**	**58.85**	20.19	**53.04**	70.12	**45.58**

8. Weight Visualization

The weight assignment strategy during training is visualized on the original RSITMD dataset in Fig.2. This strategy assigns higher weights to high-quality clean pairs to effectively guide the learning process, while adaptively allocating weights to ambiguous pairs based on their quality to fully exploit the useful information within the training data. In contrast, noisy samples are assigned zero weight to prevent them from misleading the training. Through this dynamic weighting mechanism, the model utilizes the training data more robustly, leading to enhanced robustness and retrieval performance.

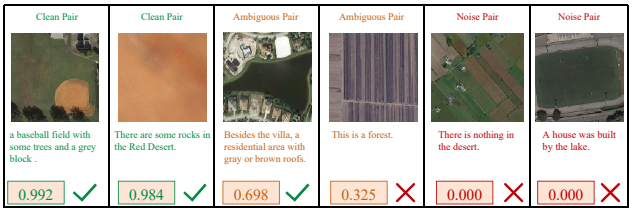

Figure 2. Weight distribution across clean, ambiguous, and noisy data pairs in RSITMD training.

9. Visualization of Noisy Sample Pairs

To verify that the original datasets indeed contain mismatched image-text pairs, we employ RRSITR to analyze three widely used benchmark datasets, RSITMD, RSICD, and NWPU. This analysis identifies several samples with matching noise. As shown in Fig.3, RRSITR effectively detects these mismatched sample pairs, thereby confirming the presence of noisy correspondences in the original datasets. Therefore, conducting robust learning on noisy datasets holds significant importance for further improving remote sensing image-text retrieval performance.

Figure 3. Some noisy sample pairs correctly identified by RRSITR.

10. Qualitative Results

As shown in Fig.4, the proposed RRSITR accurately retrieves corresponding target text given a query image on the RSITMD dataset, with most of the top-5 returned results being matched texts. For texts that correspond to only a single matching image, RRSITR consistently ranks the correct image first among all retrieval results. These results

confirm that robust learning on datasets with noisy correspondence effectively improves remote sensing image-text retrieval performance.

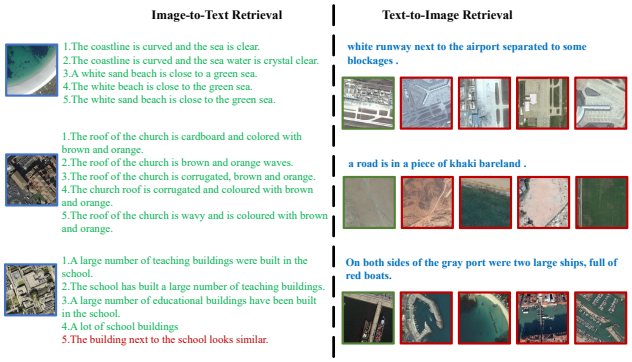

Figure 4. The visualization of the top 5 retrieval results on RSITMD, including image-to-text retrieval (left) and text-to-image retrieval (right). Blue indicates the query content, green represents matching results, and red indicates non-matching results.

References

[1] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 3

[2] Kai Gan, Bo Ye, Min-Ling Zhang, and Tong Wei. Semi-supervised CLIP adaptation by enforcing semantic and trapezoidal consistency. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[3] Zhong Ji, Changxu Meng, Yan Zhang, Yanwei Pang, and Xuelong Li. Knowledge-aided momentum contrastive learning for remote-sensing image text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 3

[4] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017. 3

[5] Deyu Meng, Qian Zhao, and Lu Jiang. What objective does self-paced learning indeed optimize? *arXiv preprint arXiv:1511.06049*, 2015. 1

[6] Sangwoo Mo, Minkyu Kim, Kyungmin Lee, and Jinwoo Shin. S-clip: Semi-supervised vision-language learning using few specialist captions. In *Advances in Neural Information Processing Systems*, pages 61187–61212. Curran Associates, Inc., 2023. 3

[7] Jiancheng Pan, Qing Ma, and Cong Bai. A prior instruction representation framework for remote sensing image-text retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 611–620, 2023. 3

[8] Jiancheng Pan, Qing Ma, and Cong Bai. Reducing semantic confusion: Scene-aware aggregation network for remote sensing cross-modal retrieval. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 398–406, 2023. 3

[9] Yijing Wang, Xu Tang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Cross-modal remote sensing image-text retrieval via context and uncertainty-aware prompt. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11384–11398, 2025. 3

[10] Rui Yang, Shuang Wang, Yingping Han, Yuanheng Li, Dong Zhao, Dou Quan, Yanhe Guo, Licheng Jiao, and Zhi Yang. Transcending fusion: A multiscale alignment method for remote sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. 3

[11] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 2, 3

[12] Weihang Zhang, Jihao Li, Shuoke Li, Jialiang Chen, Wenkai Zhang, Xin Gao, and Xian Sun. Hypersphere-based remote sensing cross-modal text–image retrieval via curriculum learning. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 3

[13] Zicong Zhu, Jian Kang, Wenhui Diao, Yingchao Feng, Junxi Li, and Jingen Ni. Sirs: Multitask joint learning for remote sensing foreground-entity image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. 3