# Enhancing Semantic Clarity: Discriminative and Fine-grained Information Mining for Remote Sensing Image-Text Retrieval

**Yu Liu**[1,2], **Haipeng Chen**[1], **Yuheng Liang**[1], **Yuheng Yang**[1], **Xun Yang**[3,*] and **Yingda Lyu**[4,*]
[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[3]University of Science and Technology of China, Hefei, China
[4] Center for Public Education Research, Jilin University, China
{yul20, yhl24, yangyh20}@mails.jlu.edu.cn, xyang21@ustc.edu.cn, {chenhp, ydlv}@jlu.edu.cn

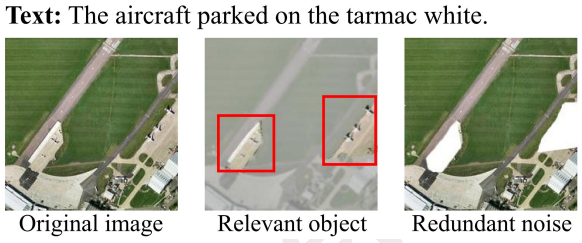
## Abstract

Remote sensing image-text retrieval is a fundamental task in remote sensing multimodal analysis, promoting the alignment of visual and language representations. The mainstream approaches commonly focus on capturing shared semantic representations between visual and textual modalities. However, the inherent characteristics of remote sensing image-text pairs lead to a semantic confusion problem, stemming from redundant visual representations and high inter-class similarity. To tackle this problem, we propose a novel **D**iscriminative and **F**ine-grained **I**nformation **M**ining (DFIM) model, which aims to enhance semantic clarity by reducing visual redundancy and increasing the semantic gap between different classes. Specifically, the Dynamic Visual Enhancement (DVE) module adaptively enhances the visual discriminative features under the guidance of multimodal fusion information. Meanwhile, the Fine-grained Semantic Matching (FSM) module cleverly models the matching relationship between image regions and text words as an optimal transport problem, thereby refining intra-instance matching. Extensive experiments on two benchmark datasets justify the superiority of DFIM in terms of retrieval accuracy and visual interpretability over the leading methods.

## 1 Introduction

Remote Sensing Image-Text Retrieval (RSITR) aims to retrieve texts or images with high semantic relevance to a given image or text from massive remote sensing databases collected by satellites or aerial drones. In recent years, there has been a growing interest in RSITR [Yuan *et al.*, 2022b; Liu *et al.*, 2024a] due to its important applications in disaster monitoring [Joyce *et al.*, 2009], natural resource exploration and remote sensing image captioning [Yang *et al.*, 2024a]. This task is of great significance yet highly challenging, as it requires precise visual-linguistic alignment.

* Corresponding author.




(a) A toy example to show redundant visual representations.

(b) High similarity across different scenes.

Figure 1: Causes of semantic confusion: (a) Redundant visual representations, and (b) high inter-class similarity.

Towards this end, many RSITR models have emerged [Pan *et al.*, 2023; Yang *et al.*, 2024b; Ma *et al.*, 2024]. Early works [Mao *et al.*, 2018; Lv *et al.*, 2021] achieve modal alignment by directly mapping different modalities to a semantic space or using cross-modal interaction. To optimize representation, [Yuan *et al.*, 2022a] proposes a multi-scale visual self-attention module to filter visual redundant features. [Mi *et al.*, 2022] improves text representation through a knowledge graph-based textual enhancement method. [Yuan *et al.*, 2022b] designs a module that dynamically fuses global and local visual features to better understand the relationships between different visual objects. The above studies have contributed prominently to RSITR. However, due to the performance limitations of feature extractors, the retrieval effect is often unsatisfactory. Recently, vision-

language pre-training (VLP) models have achieved significant success in the field of multimodal analysis. Inspired by this, several studies [Liu *et al.*, 2024a; Zhang *et al.*, 2024; Wang *et al.*, 2024] have successfully applied the Contrastive Language Image Pre-training (CLIP) model [Radford *et al.*, 2021] to the RSITR task. [Liu *et al.*, 2024a] first proposes a vision-language model for remote sensing. Both [Zhang *et al.*, 2024] and [Wang *et al.*, 2024] promote the performance improvement of remote sensing vision-language models (RSVLMs) by collecting large and high-quality remote sensing image-text datasets.

Although promising, RSVLMs focus primarily on broad application scenarios and overlook the unique semantic confusion challenges that RSITR encounters. Specifically, current methods face semantic confusion for two key reasons: 1) redundant visual representations, and 2) high inter-class similarity. Redundant visual representations mean that in remote sensing images, the proportion of the foreground is often small, making its semantic representation easily disturbed by retrieval-irrelevant areas. As depicted in Fig.1 (a), retrieval-relevant areas (*i.e.*, aircraft and tarmac) are often affected by the surrounding irrelevant objects (*e.g.*, buildings and grass). High inter-class similarity refers to the obvious similarity of image-text pairs with different scenes, which undermines the accuracy of multimodal semantic representation. As shown in Fig.1 (b), although the remote sensing images belong to different scenes (*i.e.*, school, playground, commercial area, and dense residential), their visual contents are similar, and the corresponding descriptions (*e.g.*, buildings) are apparently similar as well. These insights prompt a pivotal research question that motivates this study: "*How can we reduce visual semantic redundancy and increase inter-class representation distance to enhance semantic clarity?*"

To answer this, we present a novel Discriminative and Fine-grained Information Mining (DFIM) model, which consists of two crucial modules: 1) Dynamic Visual Enhancement (DVE), and 2) Fine-grained Semantic Matching (FSM). Concretely, the DVE module first utilizes a multimodal interaction guidance strategy to integrate semantic information from different modalities, offering precise direction for enhancing discriminative features. It then adopts a dynamic fusion strategy to flexibly refine these features, thereby effectively suppressing visual redundancy. The FSM module cleverly models the fine-grained alignment between image regions and text words as an optimal transport problem [Liu *et al.*, 2020]. By striving to minimize the transport cost between the distributions of regions and words, it can accurately distinguish subtle differences between similar pairs. Finally, we integrate the advantages of DVE and FSM to alleviate semantic redundancy and improve overall performance. The comprehensive experimental results show that our method has a significant advantage over the current state-of-the-art methods, while also possessing good visual interpretability. The main contributions of this study are summarized as follows:

- We propose a new remote sensing image-text retrieval model, DFIM, which can effectively solve the semantic confusion problem and significantly strengthen the connection between vision and language.
- To reduce visual redundancy, DVE adaptively enhances the visual discriminative features by combining multimodal interaction guidance with dynamic fusion strategies.
- To alleviate inter-class semantic confusion, FSM employs optimal transport learning strategy to identify fine-grained correspondences between image regions and text words.
- We justify the superiority of DFIM on two popular benchmark datasets (*i.e.*, RSICD and RSITMD) with extensive experiments, where our design outperforms the state-of-the-art models. Moreover, our method exhibits excellent visual interpretability.

## 2 Related Work

**Remote Sensing Image-Text Retrieval** (RSITR) aims to ensure semantic consistency between remote sensing images and text descriptions. Recently, RSITR has become a research hotspot. According to modal interaction methods [Pan *et al.*, 2023] before entering the latent semantic space, RSITR methods can be roughly classified into two categories: intra-modal interaction methods, and inter-modal interaction methods. The former [Yuan *et al.*, 2022b; Zhang *et al.*, 2022] performs information interactions only for the same modality, while the latter [Lv *et al.*, 2021; Yuan *et al.*, 2022a] performs information interactions for different modalities. Although existing methods advance through enhanced unimodal features or cross-modal interactions, their effectiveness remains constrained by limited feature extractor capabilities. Recent studies [Liu *et al.*, 2024a; Pan *et al.*, 2025; Yang *et al.*, 2024c; Liu *et al.*, 2024b]have shown that the contrastive language image pre-training (CLIP) model has been successfully applied to various tasks including RSITR. Instead, we focus on improving model retrieval accuracy while enhancing visual interpretability.

**Token Pruning** has become a hot research topic [Wei *et al.*, 2023; Cao *et al.*, 2024], aiming to dynamically reduce less important tokens based on input-dependent importance. Researchers have proposed various methods [Bolya *et al.*, 2022; Wei *et al.*, 2023] to remove redundant tokens in ViT [Alexey, 2020] to improve the model's computational efficiency. In addition, some studies [Chen *et al.*, 2025; Cao *et al.*, 2024] have reduced the computational cost of vision-language models through pruning techniques. Although these methods effectively reduce computational overhead, they may result in a certain degree of accuracy loss. On the contrary, we reduce the redundancy of visual information via token pruning, thereby extracting discriminative information and further enhancing the task-relevant representation.

**Optimal Transport Strategy** (OT) is initially developed to quantify the distance between two probability distributions. It is frequently used to establish correspondences between learnable features or to measure the distances between distributions. OT has been applied to various fields such as domain adaptation [Courty *et al.*, 2016] and person re-identification [Wang *et al.*, 2022]. [Cuturi, 2013] proposes a lightning-fast approach that efficiently tackles large-scale problems by using the Sinkhorn algorithm. Unbalanced Optimal Transport [Chapel *et al.*, 2021] reformulates the corresponding opti-

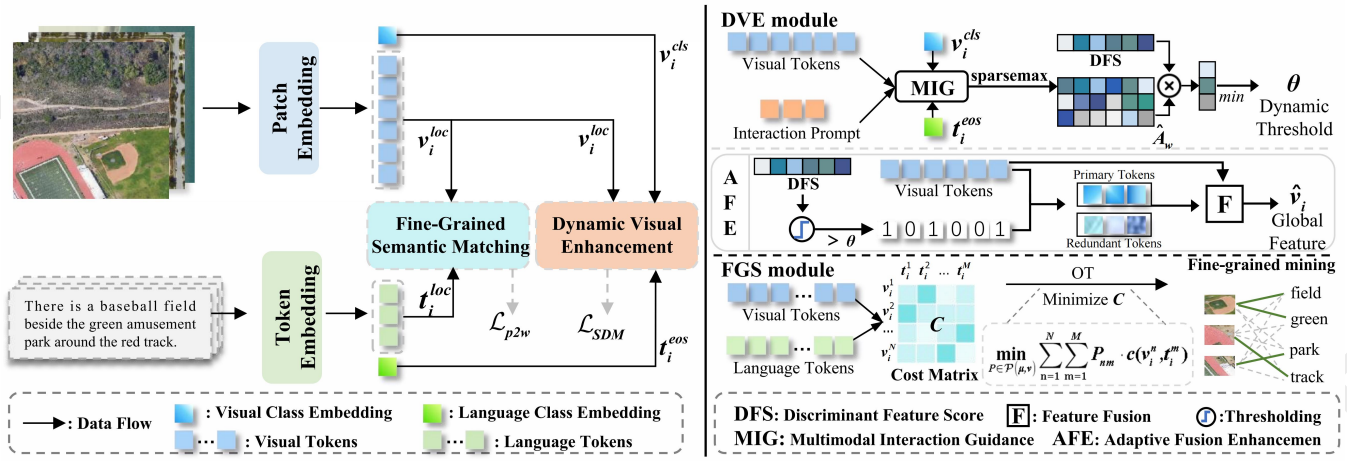



Figure 2: Overview of our proposed DFIM framework, which incorporates two principal components: the DVE module and the FSM module. The DVE and FSM modules effectively mitigate the semantic confusion problem by enhancing global discriminative visual information and mining fine-grained matching, respectively.

mization problem as a non-negative penalized linear regression problem. In contrast, we apply OT to minimize the transport cost between image patches and word embeddings, effectively capturing fine-grained semantic alignment.

# 3 Methodology

## 3.1 Problem Overview

Given a remote sensing image-text dataset $D$, RSITR task is to learn the cross-modal similarity for retrieval. Formally, $D$ consists of $K$ image-text pairs, denoted as $D = \{v_i, t_i\}_{i=1}^{K}$. Drawing inspiration from the partial success of transferring knowledge from CLIP [Radford *et al.*, 2021] to RSITR [Liu *et al.*, 2024a; Zhang *et al.*, 2024], we directly initialize our DFIM with the complete CLIP image and text encoders, thereby bolstering its inherent cross-modal alignment capabilities. In detail, given an image-text pair $(v_i, t_i)$, the visual encoder performs patch embedding to the image $v_i$ to generate the visual representation $V_i = \{v_i^{cls}, v_i^1, ..., v_i^N\} \in \mathbb{R}^{(N+1)\times d}$, which contains global visual feature $v_i^{cls} \in \mathbb{R}^{1\times d}$ and visual patch features $v_i^{loc} = \{v_i^1, ..., v_i^N\} \in \mathbb{R}^{N\times d}$. The language encoder processes words in text $t_i$ using token embedding, converting them into a textual representation $T_i = \{t_i^{sos}, t_i^1, ...t_i^{eos}\} \in \mathbb{R}^{(M+2)\times d}$, which consists of global textual feature $t_i^{eos} \in \mathbb{R}^{1\times d}$ and token features $t_i^{loc} = \{t_i^1, ..., t_i^M\} \in \mathbb{R}^{M\times d}$. Finally, we compute cosine similarity as a similarity score between the output vectors of the two modalities. Considering the above introduction of the RSITR task, it is intuitive that the alignment of the two modalities determines the model's performance. However, as illustrated in Fig.1, learning an effective feature extraction network is challenging due to the semantic confusion problem. Therefore, our primary research questions are how to accurately identify visual discriminative features and how to distinguish similar pairs by learning fine-grained matching. The pipeline of our **D**iscriminative and **F**ine-grained **I**nformation Mining (DFIM) method is shown in Fig.2.

## 3.2 Our Proposed Method: DFIM

Our DFIM method consists of two key components: *Dynamic Visual Enhancement* and *Fine-grained Semantic Matching*, which will be elaborated in Sec. 3.2 and 3.2.

**Dynamic Visual Enhancement**

As discussed in Sec.1, redundant visual representations exacerbate the semantic confusion problem. To address this challenge, we propose a Dynamic Visual Enhancement (DVE) module, which dynamically enhances discriminative features to promote multimodal alignment from a global perspective. There are two important strategies: multimodal interaction guidance, and adaptive fusion enhancement.

*Multimodal Interaction Guidance (MIG).* To ensure that the retained tokens are essential for both modalities, we first obtain guidance information via the MIG strategy. The core of this strategy is to establish the correlation between visual and textual modalities by using learnable interaction prompts $p = \{p_1, p_2, ...p_l\} \in \mathbb{R}^{l\times d}$, where $l$ is the length of tokens. Specifically, we use an attention layer to obtain visual token attention weights $A_w \in \mathbb{R}^{l\times N}$ guided by multimodal interaction information. The detailed calculation process is as:

$$A_w = softmax(\frac{p(t_i^{eos})^T v_i^{cls} (v_i^{loc})^T}{\sqrt{d_k}}), \quad (1)$$

where $d_k$ represents a scaling factor. Afterward, the visual token attention maps $A_w$ are fed into the adaptive fusion enhancement strategy to guide the fusion process of the visual features, ensuring that the enhanced features are meaningful in both modalities, which is exemplified in Fig.2.

*Adaptive Fusion Enhancement (AFE).* The dynamic token pruning methods [Chen *et al.*, 2025; Cao *et al.*, 2024] have been demonstrated to be more effective than static token pruning because they can adaptively adjust the model's compression rate in accordance with the complexity of the input instances. Inspired by this, we propose the AFE strategy to dynamically strengthen discriminative visual features. As depicted in Fig.2, we first compute the discriminant feature

score for each token. Then, a learnable threshold is used to dynamically filter the tokens based on the complexity of the instances. Finally, we introduce feature fusion operations to further enhance the visual representation. Detailed description as follows:

(1) Discriminant Feature Score (DFS). To effectively avoid discarding key tokens, our approach not only considers token importance based on class attention map [Liu *et al.*, 2022; Cao *et al.*, 2024], but also extends to the importance of tokens across different modalities. The DFS is obtained by averaging two types of scores:

$$\mathrm{DFS} = (S_{\mathrm{cls}} + S_{\mathrm{token}})/2. \tag{2}$$

Here, $S_{\mathrm{cls}}$ is the class attention score as implemented by [Liu *et al.*, 2022]. $S_{\mathrm{token}}$ denotes the token attention score. We use the visual token attention weights $A_w \in \mathbb{R}^{l\times N}$ from the MIG strategy to calculate the $S_{\mathrm{token}}$ as follows:

$$S^n_{\mathrm{token}} = \frac{\max(A^n_w)}{\sum_{n=1}^{N}\max(A^n_w)}, \tag{3}$$

where $N$ is the length of tokens and $\max(A^n_w)$ is the maximum value for the $n$-th token in the attention weights $A_w$.

(2) Token Filtering. To implement instance-wise adaptive token filtering, we first compute learnable thresholds $\theta$ using the token attention weight $A_w$ learned from the MIG strategy. The computation of $\theta$ is defined as follows:

$$\theta = \min(\hat{A}_w \otimes \mathrm{DFS}), \tag{4}$$

where $\hat{A}_w$ represents the sparse token attention maps obtained by applying the sparsemax function [Martins and Astudillo, 2016] on $A_w$. Based on the above DFS and threshold $\theta$, we perform discriminative tokens filtering. Specifically, we compare the DFS of each token with $\theta$ to obtain the filtering mask $M_m$, which is expressed as follows:

$$M_m(v^n_i) = \begin{cases} 1, & if\ \mathrm{DFS}(v^n_i) > \theta, \\ 0, & otherwise. \end{cases} \tag{5}$$

Here, $v^n_i$ is the $n$-th token in the visual patch features $v^{loc}_i$. Retain primary tokens $v^{imp}_i$ with scores above the threshold $\theta$ and exclude the rest according to the filter mask $M_m$.

(3) Feature Fusion. To prevent the loss of information caused by directly discarding tokens, we first aggregate the primary tokens $v^{imp}_i$ to obtain a new global feature $\hat{v}^{cls}_i$, which contains more discriminative information than the original global feature $v^{cls}_i$. In detail, we employ an embedding transformation [Qin *et al.*, 2024] to the updated sequence of important tokens. Subsequently, we fuse $\hat{v}^{cls}_i$ with $v^{cls}_i$ in a specific ratio. The process is defined as follows:

$$\hat{v}^{cls}_i = \mathrm{MaxPool}(mlp(v^{imp}_i + f(v^{imp}_i))), \tag{6}$$

$$\hat{v}_i = \lambda\hat{v}^{cls}_i + v^{cls}_i, \tag{7}$$

where $mlp(\cdot)$ refers to a multi-layer perceptron, $f(\cdot)$ represents a linear layer. $\lambda$ is the balance coefficient. So far, we obtain a non-redundant global visual representation $\hat{v}_i$ that contains both contextual information and enhanced discriminative properties.

**Fine-grained Semantic Matching**

While $v^{cls}_i$ and $t^{eos}_i$ provide a comprehensive understanding of global alignment relationships, due to high inter-class similarity, it may lack some fine-grained matching information to distinguish subtle differences between similar objects or scenarios. Although existing methods [Yuan *et al.*, 2022a; Zhou *et al.*, 2024] have recognized the significance of extracting fine-grained information, most of them focus exclusively on visual fine-grained details, neglecting multimodal fine-grained matching, which is of vital importance to RSITR. To mitigate this issue, we propose a Fine-grained Semantic Matching (FSM) module to enhance multimodal fine-grained awareness. Specifically, for the given $N$ suppliers (image patches) and $M$ demanders (words). The supplier supplies image patches to the demander, described as a vector $\mathbf{x}$, and the demander receives image patches from the supplier, described as a vector $\mathbf{y}$. We formulate the fine-grained matching task as an optimal transport problem [Liu *et al.*, 2020], which is to find an optimal transportation plan $P \in \mathbb{R}^{N\times M}$ to minimize the transport cost $C$. It can be expressed as:

$$P^* = \arg\max_{P\in\mathcal{P}} \langle P, C\rangle_F + \lambda_1 H(P), \tag{8}$$

where $\langle P, C\rangle_F$ is the Frobenius inner product between the transportation (matching) plan $P$ and cost matrix $C$. The matrix $c(v^n_i, t^m_i)$ is an element of $C$, which denotes the transport costs between $v^n_i$ and $t^m_i$. It can be expressed as $c(v^n_i, t^m_i) = 1 - \cos(v^n_i, t^m_i)$. In addition, we add entropy regularization on $P$ as $H(P) = \sum_{nm} P_{nm}\log P_{nm}$ to ensure that $P$ is not over-concentrated on a few elements. Ultimately, the solution of Eq. 8 can limit a transportation polytope:

$$\mathcal{P} = \{P \in \mathbb{R}^{N\times M} | P\mathbb{1}_M = \mathbf{x}, P^T\mathbb{1}_N = \mathbf{y}\}, \tag{9}$$

where $P_{nm}$ represents the transport plan between the $n$-th image patch and the $m$-th word, and $P$ contains all non-negative $N\times M$ elements, with row and column sums equal to $\mathbf{x}$ and $\mathbf{y}$, respectively. We employ the inexact proximal point method for optimal transport (IPOT) [Xie *et al.*, 2020] to approximate the transport plan, which is formulated as follows:

$$P^* = Diag(\mu)\exp(C/\lambda_2)Diag(v), \tag{10}$$

where $\mu$ and $v$ are row and column normalized vectors, respectively, and can be computed via the iterative Sinkhorn-Knopp algorithm [Cuturi, 2013]. To maximize inter-class distinctions and learn fine-grained matching, each image patch should be closely aligned with the corresponding text word. Thus, we define the patch-to-word alignment loss $\mathcal{L}_{p2w}$ as:

$$\mathcal{L}_{p2w} = \min_{P\in\mathcal{P}(\mu,v)} \sum_{n=1}^{N}\sum_{m=1}^{M} P_{nm}\cdot c(v^n_i, t^m_i). \tag{11}$$

Overall, the FSM module reduces the inter-class similarity by explicitly establishing fine-grained alignment between patches and words, thereby effectively alleviating the semantic confusion problem.

**Learning Strategies**

Our training objective consists of two parts: 1) We leverage the bi-directional similarity distribution matching loss [Jiang

and Ye, 2023] $\mathcal{L}_{SDM}$ to supervise the learning of the global representations across different modalities:

$$\mathcal{L}_{SDM} = \mathcal{L}_{sdm}(\hat{v}_i, t_j^{eos}) + \mathcal{L}_{sdm}(t_i^{eos}, \hat{v}_j),$$
$$\mathcal{L}_{sdm} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{i,j}\log(\frac{p_{i,j}}{q_{i,j}+\epsilon}), \tag{12}$$

where $p_{i,j}$ is the probability of matching pairs, and $q_{i,j} = y_{i,j}/\sum_{k=1}^{K} y_{i,k}$ is the true matching probability. 2) We also adopt the proposed $\mathcal{L}_{p2w}$ to supervise the fine-grained matching as described in Sec. 3.2. The total loss is given by:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{SDM} + \beta\mathcal{L}_{p2w}, \tag{13}$$

where $\alpha$ and $\beta$ denote the balance coefficient.

# 4 Experiments

This section conducts experiments to answer the following questions. RQ1: How effective is DFIM in improving the model performance across different settings? RQ2: How do the Dynamic Visual Enhancement (DVE) module and Fine-grained Semantic Matching (FSM) module contribute to the performance? RQ3: What are the learning patterns and insights of DFIM?

## 4.1 Settings

**Datasets.** We evaluate our method on the RSICD and RSITMD datasets. **RSICD** [Lu *et al.*, 2017] consists of a total of 10,921 images, each image is associated with five text descriptions. Following [Yuan *et al.*, 2022a], we divide the dataset into 7,862 training images, 1,966 validation images, and 1,093 test images. **RSITMD** [Yuan *et al.*, 2022a] includes 4,743 images, with each image annotated by five sentences offering a finer-grained description than RSICD. Utilizing the same division strategy as [Yuan *et al.*, 2022b], we split the dataset into 3,435 training images, 856 validation images, and 452 test images. Additionally, to investigate the model's performance on both significant and insignificant sample pairs, we divide the RSICD and RSITMD test sets according to [Ma *et al.*, 2024]. The distinction between significant and insignificant images is determined by whether they contain prominent objects.

**Evaluation Protocol.** Following standard practices in image-text retrieval, we evaluate the performance of DFIM by *R@K* (*K* = 1, 5, 10) and *mR*. *R@K* denotes the proportion of accurately matched pairs within the top *K* retrieval outcomes, while *mR* represents the mean of these *R@K* values.

**Implementation Details.** All experiments are conducted in a station equipped with RTX4090 24GB GPU. We use CLIP (ViT-B-32) [Radford *et al.*, 2021] as the backbone, and employ ViT-B-32-RET-2 [Zhang *et al.*, 2024] for parameter initialization. All input images are resized to $224 \times 224$ and then randomly cropped and rotated to enhance the training samples. The maximum length of the textual token sequence is set to 77. We set the epoch to 20 and the batchsize to 128 on both datasets. We use Adam [Kingma, 2014] as the model optimizer. A cosine learning rate decay is applied, starting with an initial learning rate of $1 \times 10^{-4}$ for RSICD and $1 \times 10^{-5}$ for RSITMD. The hyper-parameters $\lambda$, $\alpha$ and $\beta$ are set to 0.5, 1.0 and 0.7 for RSICD, respectively. The hyper-parameters $\lambda$, $\alpha$ and $\beta$ are set to 0.6, 0.8 and 1.1 for RSITMD, respectively.

## 4.2 Quantitative Comparison (RQ1)

In this section, we comprehensively evaluate our DFIM on two widely-used RSITR datasets (*i.e.*, RSICD [Lu *et al.*, 2017] and RSITMD [Yuan *et al.*, 2022a]). According to different training paradigms, the methods for comparison with DFIM can be roughly divided into two categories: 1) Traditional cross-modal retrieval methods, including AMFMN [Yuan *et al.*, 2022a], GaLR [Yuan *et al.*, 2022b], SWAN [Pan *et al.*, 2023], HVSA[Zhang *et al.*, 2023] and DOVE [Ma *et al.*, 2024]. 2) CLIP-based methods, such as AdaptFormer [Chen *et al.*, 2022], PE-RSITR [Yuan *et al.*, 2023], GeoRSCLIP [Zhang *et al.*, 2024] and HarMA [Huang, 2024].

**Results on RSICD Dataset.** According to the comparison results on RSICD reported in Tab.1, we can find that our DFIM shows a significant increase compared to the SOTA approach HarMA. For example, *R@1* improves 14.1% (23.42 vs. 20.52) and 11.7% (17.70 vs. 15.84) in sentence and image retrieval, respectively. In general, there is a 10.7% (43.14 vs. 38.95) improvement on the *mR* metric. Thus, our method outperforms traditional and CLIP-based retrieval methods on the RSICD dataset, reflecting its superior performance in resolving semantic confusion problem.

**Results on RSITMD Dataset.** Tab.1 presents the results on RSITMD. It is evident that DFIM has a high performance advantage, as it achieves considerable improvements across all metrics relative to state-of-the-art methods. Specifically, compared to HarMA, the *R@1* of image-query-text is improved by 6.8% (34.97 vs. 32.74), and *mR* is improved by 5.2% (55.01 vs. 52.27). The above results demonstrate the effectiveness of DFIM in reducing the semantic confusion and the superiority of the scene perception capability.

**Results on significant and insignificant test sets.** To further explore the capability of DFIM in matching significance and insignificance objects, we conduct controlled experiments: 1) the **complete** test set; 2) the **significant** test set; and 3) the **insignificant** test set. The difference between significant and insignificant images depends on whether they contain salient objects or not. Fig.3 displays results on the complete, significant and insignificant test sets of the RSICD and RSITMD datasets. DFIM has achieved substantial improvements over traditional methods (*i.e.*, SWAN and DOVE). Besides, we implement a baseline approach CLIP*, which uses ViT-B-32-RET-2 [Zhang *et al.*, 2024] for parameter initialization. Our method achieves the best performance on the *mR* metric for all three experimental setups. That is, our method significantly improves the retrieval effect by reducing visual redundancy and mining fine-grained alignment information, regardless of whether the object is salient or not.

## 4.3 In-depth Studies of DFIM (RQ2)

**Contributions of the DFIM's components.** To fully understand the DFIM, we explore the effectiveness of the DVE and FSM modules on the RSITMD dataset. The corresponding performances are reported in Tab.2. **Effectiveness of DVE.** Observing w/o FSM, we can find a 3.0% (53.36 vs. 55.01)

| Method | Ref | RSICD dataset | | | RSITMD dataset | | |
|---|---|---|---|---|---|---|---|
| | | Image-query-Text R@1 / R@5 / R@10 | Text-query-Image R@1 / R@5 / R@10 | mR | Image-query-Text R@1 / R@5 / R@10 | Text-query-Image R@1 / R@5 / R@10 | mR |
| AMFMN | TGRS'22 | 5.21 / 14.72 / 21.57 | 4.08 / 17.00 / 30.60 | 15.53 | 10.63 / 24.78 / 41.81 | 11.51 / 34.69 / 54.87 | 29.72 |
| GaLR | TGRS'22 | 6.59 / 19.85 / 31.04 | 4.69 / 19.48 / 32.13 | 18.96 | 14.82 / 31.64 / 42.48 | 11.15 / 36.68 / 51.68 | 31.41 |
| SWAN | ICMR'23 | 7.41 / 20.13 / 30.86 | 5.56 / 22.26 / 37.41 | 20.61 | 13.35 / 32.15 / 46.90 | 11.24 / 40.40 / 60.60 | 34.11 |
| HVSA | TGRS'23 | 7.47 / 20.62 / 32.11 | 5.51 / 21.13 / 34.13 | 20.16 | 13.20 / 32.08 / 45.58 | 11.43 / 39.20 / 57.45 | 33.16 |
| DOVE | TGRS'24 | 8.66 / 22.35 / 34.95 | 6.04 / 23.95 / 40.35 | 22.72 | 16.81 / 36.80 / 49.93 | 12.20 / 44.13 / 66.50 | 37.73 |
| AdaptFormer | NIPS'22 | 12.46 / 28.49 / 41.86 | 9.09 / 29.89 / 46.91 | 28.1 | 16.71 / 30.16 / 42.91 | 14.27 / 41.53 / 61.46 | 34.81 |
| PE-RSITR | TGRS'22 | 14.16 / 31.51 / 44.78 | 11.63 / 33.92 / 50.73 | 31.12 | 23.67 / 44.07 / 60.36 | 20.10 / 50.63 / 67.97 | 44.47 |
| GeoRSCLIP | TGRS'24 | 21.13 / 41.72 / 55.63 | 15.59 / 41.19 / 57.99 | 38.87 | 32.30 / 53.32 / 67.92 | 25.04 / 57.88 / 74.38 | 51.81 |
| HarMA | ICLRW'24 | 20.52 / 41.37 / 54.66 | 15.84 / 41.92 / 59.39 | 38.95 | 32.74 / 53.76 / 69.25 | 25.62 / 57.65 / 74.60 | 52.27 |
| **DFIM** | Ours | **23.42 / 45.09 / 62.63** | **17.70 / 46.61 / 63.42** | **43.14** | **34.97 / 57.36 / 71.25** | **28.72 / 59.83 / 77.92** | **55.01** |

Table 1: Comparisons of image-text retrieval results on RSICD and RSITMD.

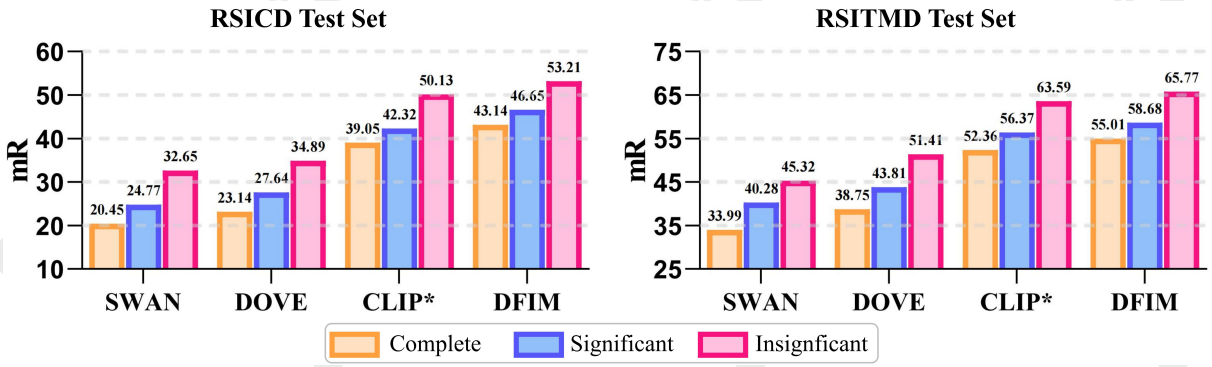



Figure 3: Results from the complete, significant, and insignificant test sets of the RSICD and RSITMD datasets are presented, assessing the performance of various retrieval methods in recognizing significant and insignificant objects. The *complete* dataset is divided into *significant* and *insignificant* subsets based on whether the image contains categories of common remote sensing objects or not.

performance drop in the *mR* metric. This shows DVE can substantially improve retrieval ability. **Effectiveness of FSM.** Without using FSM (*i.e.*, w/o FSM), it can be found that a 4.1% (52.73 vs. 55.01) performance drop in the *mR* metric. This indicates that FSM can significantly improve the fine-grained matching performance. Overall, the results show that the combination of DVE and FSM can effectively mitigate the semantic confusion.

| Method | Sentence Retrieval | Image Retrieval | |
|---|---|---|---|
| | R@1 / R@5 / R@10 | R@1 / R@5 / R@10 | mR |
| baseline | 31.24 / 53.68 / 68.31 | 25.36 / 57.04 / 73.61 | 51.54 |
| w/o DVE | 33.27 / 55.19 / 70.06 | 27.38 / 58.09 / 76.15 | 53.36 |
| w/o FSM | 33.01 / 54.43 / 69.58 | 26.23 / 57.96 / 75.17 | 52.73 |
| **DFIM** | **34.97 / 57.36 / 71.25** | **28.72 / 59.83 / 77.92** | **55.01** |

Table 2: Ablation studies for DFIM's components on RSITMD.

**Analysis of Hyper-parameters.** The sensitivity analysis of hyper-parameters $\lambda$, $\alpha$ and $\beta$ is shown in Fig.4. $\lambda$ represents the intensity of visual feature enhancement. $\alpha$ and $\beta$ balance the contributions of global and local matching, respectively. Over a wide range of hyper-parameters, our model shows very small fluctuations. In general, variations in the hyper-parameters do not lead to significant degradation in performance, which demonstrates the stability of our DFIM.

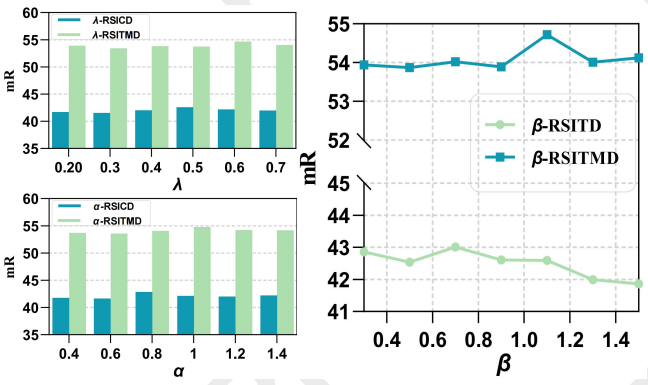



Figure 4: Ablation studies of Hyper-parameters.

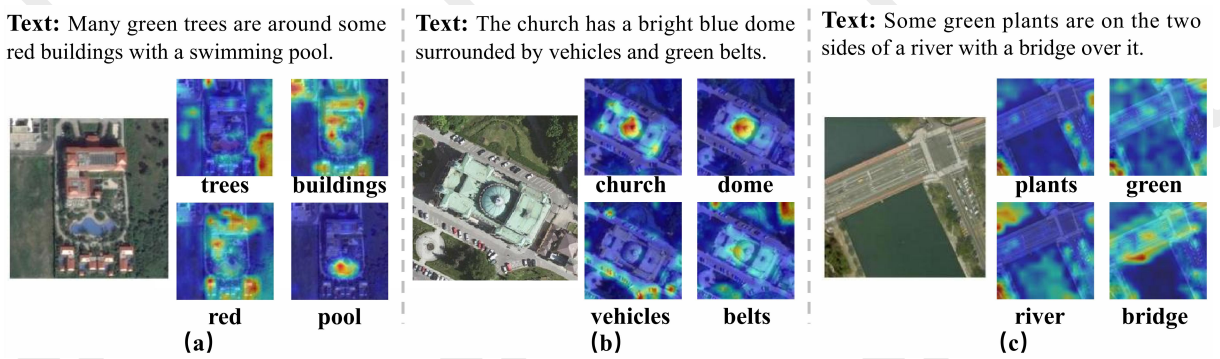



Figure 5: Several examples of visualizations illustrating the region-word alignment. Highlighted regions correspond to the relevant words.

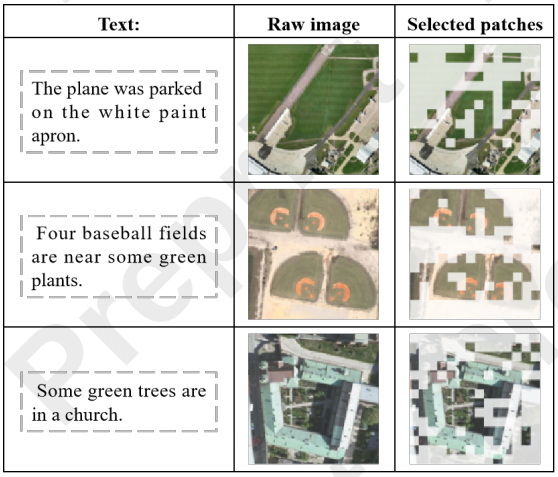



Figure 6: Visualization of token filtering results. The white mask in the image represents the pruned visual tokens.

### 4.4 Qualitative Analysis (RQ3)

To capture the learning insights of DFIM, we perform a series of detailed visual analyses. *(1) Visual Token Selection Visualization.* In Fig.6, we display the visualization results of token filtering on the RSICD dataset. The figure consists of textual descriptions, the original images, and the corresponding processed outputs, where the white mask indicates redundant patches. It can be observed that based on the textual descriptions, our DFIM retains key tokens and prunes irrelevant ones. For instance, DFIM preserves the "baseball field" in the second row of the example shown in Fig.6. Besides, we observe that for different scenes (*i.e.*, airport, baseball field, and church), DFIM can adaptively distinguish between primary and redundant tokens based on the size of different targets. These visualization results highlight DFIM's capacity to accredit diverse scenes and also validate its interpretability in dealing with visual redundant representations. *(2) Region-Word Alignment Visualization.* In order to further evaluate the DFIM's capacity to capture fine-grained relationships, we leverage gradient-weighted attention [Chefer *et al.*, 2021] to generate heat maps, where colors closer to red indicate more semantic relevance. As shown in Fig.5 (a), "trees", "buildings", and "pool" are accurately mapped to their corresponding image regions. It is worth noting that, although the "pool" area is very small, DFIM is still able to align the region and word accurately. In addition, there are multiple dispersed small targets (*i.e.*, vehicles) in Fig.5 (b), we observe that DFIM can accurately match the word "vehicles" with the dispersed targets in the image. Meanwhile, DFIM can accurately identify the entire region associated with color-related words (*e.g.*, "red" and "green"), rather than merely the target object that exhibits that color. As shown in Fig.5 (c), the region corresponding to "green" encompasses not only areas associated with green plants but also all green regions in the image. This indicates that DFIM can achieve fine-grained matching and resist overfitting.

## 5 Conclusion

In this paper, we propose a novel Discriminative and Fine-grained Information Mining (DFIM) approach to alleviate semantic confusion in the RSITR task. We first analyze two causes of semantic confusion: visual representation redundancy and excessive inter-class similarity. To address these issues, we design a Dynamic Visual Enhancement (DVE) module and a Fine-grained Semantic Matching (FSM) module, respectively. Specifically, the DVE module can adaptively strengthen the visual discriminative features that are critical to both modalities. Meanwhile, the FSM module cleverly formulates the fine-grained matching relationship between image regions and text words as an optimal transport problem, amplifying inter-class distinctions. Comprehensive experiments validate DFIM's effectiveness in both retrieval accuracy and visual interpretability. In future work, we plan to extend DFIM to more tasks [Yang *et al.*, 2023; Wu *et al.*, 2025; Xu *et al.*, 2025; Yang *et al.*, 2021] and assess its effectiveness in diverse scenarios.

## Acknowledgments

This research is supported by the National Natural Science Foundation of China (62276112, 62272435, and U22A2094).

## References

[Alexey, 2020] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[Bolya *et al.*, 2022] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

[Cao *et al.*, 2024] Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15710–15719, 2024.

[Chapel *et al.*, 2021] Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. *Advances in Neural Information Processing Systems*, 34:23270–23282, 2021.

[Chefer *et al.*, 2021] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

[Chen *et al.*, 2022] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[Chen *et al.*, 2025] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2025.

[Courty *et al.*, 2016] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[Huang, 2024] Tengjun Huang. Efficient remote sensing with harmonized transfer learning and modality alignment. *arXiv preprint arXiv:2404.18253*, 2024.

[Jiang and Ye, 2023] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023.

[Joyce *et al.*, 2009] Karen E Joyce, Stella E Belliss, Sergey V Samsonov, Stephen J McNeill, and Phil J Glassey. A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Progress in physical geography*, 33(2):183–207, 2009.

[Kingma, 2014] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Liu *et al.*, 2020] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020.

[Liu *et al.*, 2022] Xiangcheng Liu, Tianyi Wu, and Guodong Guo. Adaptive sparse vit: Towards learnable adaptive token pruning by fully exploiting self-attention. *arXiv preprint arXiv:2209.13802*, 2022.

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

[Liu *et al.*, 2024b] Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. Causality-inspired invariant representation learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14052–14060, 2024.

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

[Lv *et al.*, 2021] Yafei Lv, Wei Xiong, Xiaohan Zhang, and Yaqi Cui. Fusion-based correlation learning model for cross-modal remote sensing image retrieval. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[Ma *et al.*, 2024] Qing Ma, Jiancheng Pan, and Cong Bai. Direction-oriented visual-semantic embedding model for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Mao *et al.*, 2018] Guo Mao, Yuan Yuan, and Lu Xiaoqiang. Deep cross-modal retrieval for remote sensing image and audio. In *2018 10th IAPR workshop on pattern recognition in remote sensing (PRRS)*, pages 1–7. IEEE, 2018.

[Martins and Astudillo, 2016] Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR, 2016.

[Mi *et al.*, 2022] Li Mi, Siran Li, Christel Chappuis, and Devis Tuia. Knowledge-aware cross-modal text-image retrieval for remote sensing images. In *Proceedings of the Second Workshop on Complex Data Challenges in Earth Observation (CDCEO 2022)*, 2022.

[Pan *et al.*, 2023] 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.

[Pan *et al.*, 2025] Haowen Pan, Xiaozhi Wang, Yixin Cao, Zenglin Shi, Xun Yang, Juanzi Li, and Meng Wang. Precise localization of memories: A fine-grained neuron-level knowledge editing technique for llms. *arXiv preprint arXiv:2503.01090*, 2025.

[Qin *et al.*, 2024] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024.

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

[Wang *et al.*, 2022] Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 93–109. Springer, 2022.

[Wang *et al.*, 2024] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5805–5813, 2024.

[Wei *et al.*, 2023] Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. Joint token pruning and squeezing towards more aggressive compression of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2092–2101, 2023.

[Wu *et al.*, 2025] Sifan Wu, Hongzhe Zhang, Zhenguang Liu, Haipeng Chen, and Yingying Jiao. Enhancing human pose estimation in the internet of things via diffusion generative models. *IEEE Internet of Things Journal*, 2025.

[Xie *et al.*, 2020] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR, 2020.

[Xu *et al.*, 2025] Yanlong Xu, Haoxuan Qu, Jun Liu, Wenxiao Zhang, and Xun Yang. Cmmloc: Advancing text-to-pointcloud localization with cauchy-mixture-model based framework. *arXiv preprint arXiv:2503.02593*, 2025.

[Yang *et al.*, 2021] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1–10, 2021.

[Yang *et al.*, 2023] Yuheng Yang, Haipeng Chen, Zhenguang Liu, Yingda Lyu, Beibei Zhang, Shuang Wu, Zhibo Wang, and Kui Ren. Action recognition with multi-stream motion modeling and mutual information maximization. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1658–1666, 2023.

[Yang *et al.*, 2024a] Cong Yang, Zuchao Li, and Lefei Zhang. Bootstrapping interactive image-text alignment for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Yang *et al.*, 2024b] Lingling Yang, Tongqing Zhou, Wentao Ma, Mengze Du, Lu Liu, Feng Li, Shan Zhao, and Yuwei Wang. Remote sensing image-text retrieval with implicit-explicit relation reasoning. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Yang *et al.*, 2024c] Xun Yang, Tianyu Chang, Tianzhu Zhang, Shanshan Wang, Richang Hong, and Meng Wang. Learning hierarchical visual transformation for domain generalizable visual matching and recognition. *International Journal of Computer Vision*, 132(11):4823–4849, 2024.

[Yuan *et al.*, 2022a] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv preprint arXiv:2204.09868*, 2022.

[Yuan *et al.*, 2022b] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

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

[Zhang *et al.*, 2022] Huan Zhang, Yingzhi Sun, Yu Liao, SiYuan Xu, Rui Yang, Shuang Wang, Biao Hou, and Licheng Jiao. A transformer-based cross-modal image-text retrieval method using feature decoupling and reconstruction. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 1796–1799. IEEE, 2022.

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

[Zhang *et al.*, 2024] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Zhou *et al.*, 2024] Zihui Zhou, Yong Feng, Agen Qiu, Guofan Duan, and Mingliang Zhou. Fine-grained information supplementation and value-guided learning for remote sensing image-text retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:19194–19210, 2024.