This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

Strong and Weak Prompt Engineering for Remote Sensing Image-Text Cross-Modal Retrieval

Tianci Sun, Chengyu Zheng, Xiu Li, Yanli Gao,
Jie Nie*, *Member, IEEE* , Lei Huang, *Member, IEEE*, Zhiqiang Wei, *Member, IEEE*

***Abstract*—-modal retrieval is vital at the intersection of vision and language. Specifically, remote sensing image-text retrieval enhances our understanding of complex remote sensing content by combining multi-perspective visual information with concise textual descriptions and has increasingly become a hotspot for research. Existing prompts typically emphasize either global or local information, which fails to excavate or fully leverage the effective information of cross-modal data, resulting in the subpar performance of retrieval models. To address these limitations, we propose a novel method called Strong and Weak Prompt Engineering (SWPE) for remote sensing image-text retrieval.Specifically, SWPE employs the Strong and Weak Prompt Generation (SWPG) module to generate fine-grained and global category semantic prompts via an attention mechanism and a pre-trained classification model. The Prompt-guided Feature Fine-tuning (PFF) module then refines the prompt information using a Transformer architecture, integrating the refined prompts with high-level image and text features to enhance both fine-grained details and global semantics. Finally, the Adaptive Hard Sample Elimination (AHSE) module optimizes the triplet loss function by training the model with negative sample pairs of varying difficulty, assigning higher weights to simpler pairs.Extensive quantitative and qualitative experiments on four remote sensing benchmarks validate the superior effectiveness of SWPE.**

***Index Terms*—image-text cross-modal retrieval, remote sensing, prompt engineering**

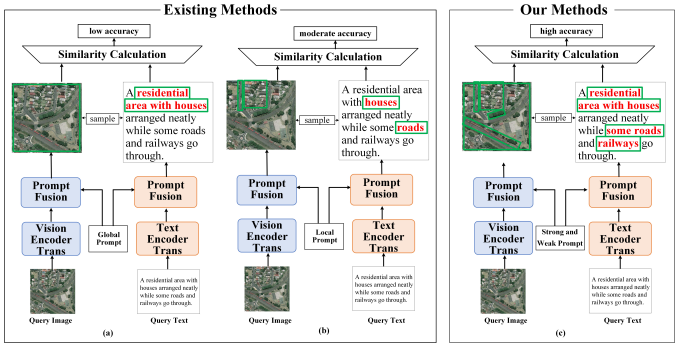

Fig. 1. The comparison of existing methods and ours. Previously, global prompt effectively guides global information but overlooks local details, while a local prompt does the opposite, offering good guidance for details but lacking in global information. Our method combines both by introducing a Strong and Weak Prompt, where the attention mechanism is responsible for significant fine-grained prompt information, and the pre-trained classification model provides global semantic prompt information. By allowing these two to interact and complement each other, our model achieves better performance.

I. Introduction

In recent years, the rapid development of remote sensing technology, coupled with advances in satellite data acquisition and storage, has enabled significant applications in disaster detection, resource management, national defense, and security [1]–[3]. Consequently, extracting valuable information from the vast volumes of remote sensing data has become a prominent research focus. Among data processing methods, cross-modal text-image retrieval is particularly noteworthy, as it integrates text descriptions with image content to facilitate efficient and accurate retrieval of specific information from large-scale datasets. Notably, Remote Sensing Text-Image Cross-Modal Retrieval (RSITR) plays a central role by intuitively linking remote sensing images with linguistic expressions, thereby attracting growing attention in the academic community.

This work was supported in part bythe National Natural Science Foundation of China(U23A20320),the Central Government Guided Local Science and Technology Development Fund(YDZX2022028).

Tianci Sun, Chengyu Zheng, Xiu Li, Yanli Gao, Jie Nie, Zhiqiang Wei, Lei Huang, are with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, 266100, China.*(Tianci Sun is the first author.) (Jie Nie is the corresponding author. email: niejie@ouc.edu.cn)*

The rapid development of deep learning methods has significantly advanced text-to-image cross-modal retrieval. Deep learning has demonstrated remarkable success in image and natural language processing, offering powerful tools for cross-modal retrieval tasks. Current methods for cross-modal retrieval typically involve two key steps: feature modeling and similarity matching, leveraging deep learning techniques such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to extract features from text and images. For instance,the work [4] utilizes pre-trained CNNs and GRUs to extract features from remote sensing images and semantic representations from text, proposing a hypersphere-based cross-modal retrieval method (HVSA), and employs curriculum learning strategies to optimize model performance. Similarly, work [5] utilizing ResNet-18 and GRU networks, introduced the KAMCL method, which focuses on addressing subtle differences in descriptions between remote sensing images and text often overlooked by existing methods. The fundamental objective of their work is to bridge the cross-modal information gap and address the semantic gap by constructing a representative text-image joint embedding space. The semantic gap poses significant challenges, hindering models from accurately capturing the relationships between text and images.To eliminate the semantic gap, early research mainly focused on unsupervised learning methods [6]–[8], which sought to explore the complex transformation relationships between two or more types of data, following the trajectory of linear canonical

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

correlation analysis and extending to its nonlinear forms, in hopes of capturing the deep-level associations between data. At the same time, some researchers advocate for supervised learning strategies [9]–[12] that use labeled information to learn more discriminative vector representations. Within this framework, the embedding space is optimized by bringing vectors of matching sample pairs closer together while pushing vectors of non-matching pairs farther apart, thereby enhancing the distinction between samples at the vector level.

In recent years, the emergence of pre-trained large models such as CLIP [13], BLIP [14], and MiniGPT-4 [15] has opened up new possibilities for cross-modal image-text retrieval.These large models leverage joint learning on extensive cross-modal datasets to establish semantic associations between images and text, enabling them to process data across different modalities. Pre-trained on diverse multi-modal datasets spanning various tasks and domains, these models acquire rich semantic information, enhancing their generalization and adaptability.Given the unique nature of remote sensing data, researchers have developed specialized large models for this field, such as RemoteCLIP [16] and GeoRSCLIP [17].Although the aforementioned multi-modal pre-trained large models do not explicitly mention "prompt" during training, they essentially utilize a templated text processing method, which is somewhat akin to "prompt". We can also observe that in some works and applications, prompt engineering has become a technical means to enhance model performance. There are also some methods that have adopted more direct prompts. In computer vision, works [18]–[20] enhance large-scale visual Transformer models through prompt engineering, with Visual Prompt Tuning (VPT), CoOp for adaptability, and a unified prompt method for visual-language representations. In retrieval, work [21] introduces a cross-modal image retrieval method using text titles, work [22] improves cross-modal interaction with local prompts, and work [23] adapts CLIP for text-video retrieval with VoP.

While the aforementioned methods have demonstrated notable progress, the prompts they employ remain relatively simplistic and poorly suited to remote sensing cross-modal retrieval tasks. Unlike natural images, remote sensing images are acquired by sensors positioned far above the Earth's surface, capturing diverse scenes and complex targets. Current prompts tend to focus on either global or local information, limiting their ability to extract the cross-modal data effectively. This shortcoming contributes to the suboptimal performance of retrieval models. For example, as illustrated in Fig.1 (a) and (b), global prompts can effectively capture overall information. However, this global information may not be reliable for cross-modal retrieval tasks involving complex data, thus outputting low matching accuracy. Similarly, local prompts can offer detailed guidance but lack interaction prompts from cross-modal data, leading to unsatisfactory outcomes.

In response to the aforementioned issues, we have fully considered the characteristics of remote sensing images such as large coverage, target diversity, small target size, and complex background noise. We have proposed a Strong and Weak Prompt Engineering for remote sensing image-text cross-modal retrieval (SWPE). We still use traditional frameworks to obtain feature representations of images and text.

In our model, we have innovatively designed three modules. The first is the Strong and Weak Prompt Generation Module (SWPG), which enhances initial learnable parameters by incorporating attention-based weak prompts and strong prompts derived from a pre-trained classification model. The self-attention mechanism in the attention-based weak prompts focuses on various internal segments of the data, effectively capturing long-range dependencies, while the cross-attention mechanism addresses the relationships between text and images, thereby augmenting their interaction. We optimize the prompts obtained through self-attention and cross-attention by appropriately distributing weights to achieve the desired fine-grained prompts. The strong prompts, based on the pre-trained classification model, first process the high-level representations of images and text through convolutional layers, subsequently inputting these representations into our pre-trained ResNet classification network to extract class prompt information, which is further optimized to yield global category semantic prompts.

The second module is the Prompt-guided Feature Fine-tuning (PFF), where we utilize a Transformer architecture to transform the prompt information. The transformed prompts are then fused with the extracted high-level image and text features through the Transformer architecture, enhancing critical fine-grained details and global semantics within the high-level features.

Finally, we have the Adaptive Hard Sample Elimination (AHSE) module. We optimize the triplet loss function to utilize negative sample pairs of varying difficulty to improve the model. We rank the negative sample pairs based on their difficulty and assign different weights to them. We group several pairs of negative samples that are the farthest apart in the sample space and treat them as a whole, proposing a method to learn network parameters from easy to difficult. During backpropagation, we assign greater weights to simpler sample pairs, aiming to first learn these simpler pairs to ensure effective and reliable image-text retrieval when dealing with complex remote sensing data.The innovations of our method are illustrated in Fig.1 (c), our approach simultaneously considers the salient details and global semantic information of remote sensing data, as well as cross-modal mutually guiding prompts, significantly enhancing the model's performance.

The main contributions of our work are as follows:

1) We propose an RSITR network called SWPE, which simultaneously considers the significant details and global semantic information of remote sensing data, and cross-modal mutual guidance prompts, to enhance the effectiveness and reliability of processing complex remote sensing data.To the best of our knowledge, applying prompt engineering in remote sensing image-text cross-modal retrieval can effectively address the challenges of target diversity, small target size, and complex background noise in remote sensing images.

2) We propose SWPG, which combines weak and strong prompts to obtain prompts that contain both significant details and global semantic information. Additionally, we designed the PFF module, which employs a transformer architecture to refine prompts and integrate them with image and text features,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

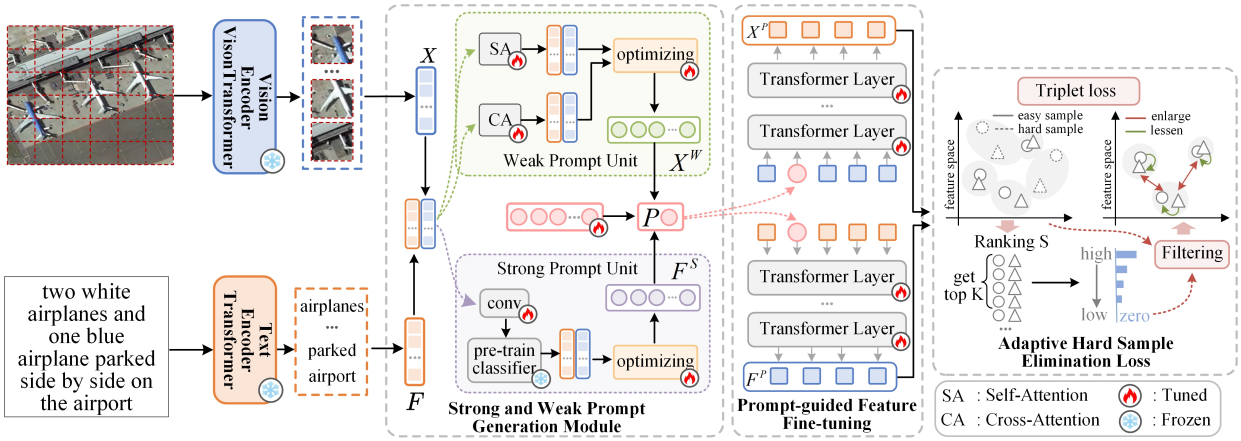

Fig. 2. The architecture of the proposed SWPE. First, we apply the pre-trained feature extractor RemoteClip to explore the high-level feature representations of images and texts. The generated features are fed into the Strong and Weak Prompt Generation Module, which includes both weak prompt units and strong prompt units. These units work together with a learnable feature initialized to one to generate the prompt information P. Next, we input the prompt information P along with the extracted high-level image features and high-level text features into the Prompt-guided Feature Fine-tuning Module. In this module, we use an attention mechanism to fuse the prompt information with the high-level image and text features, resulting in image and text features that incorporate the prompt information. Finally, we proceed to the Adaptive Hard Sample Elimination Loss Module, where we assign different weights to samples of varying difficulty levels, with the aim of prioritizing the learning of simpler samples. The details of SWPE are depicted inSection III.

producing the desired final features. Finally, we introduce AHSE loss for adaptive similarity matching, assigning variable weights to sample pairs based on difficulty, prioritizing the learning of simpler pairs.

3) Through quantitative and qualitative experiments, the effectiveness of our architecture has been validated on four RS text-image retrieval datasets.

II. RELATED WORK

A. Remote Sensing Text-Image Retrieval

In recent years, cross-modal remote sensing text-image research has gained prominence, focusing on uncovering the intrinsic connections between remote sensing images and associated texts to improve the comprehensive analysis and interpretation of remote sensing data. In the stage of image and text feature representation, most methods are still based on CNN and RNN implementations. For example,the work [24] proposed a semantic alignment module to fully explore the latent correspondences between images and texts. They used attention and gating mechanisms to filter and optimize data features, thereby obtaining more discriminative feature representations. And the work [25] introduced a new CMR-SITR model - the Multiscale Interactive Transformer (MSIT). MSIT employs a straightforward feature learning model for text and RS images, ensuring the model's lightweightness, and incorporates a Transformer encoder that enhances the usefulness of features by considering potential relationships between different representations. Of course, there are also work [26] proposed a method called Multiscale Salient Image-Guided Text Alignment (MSITA), which improves the performance of cross-modal RSIT retrieval by learning salient information and optimizing image-guided text alignment.Additionally, the work [27] developed a GaLR, which grants the model the capability of multilevel understanding of remote sensing images. The work [28] proposed a new RSCTIR framework based on Mask-guided Relation Modeling with Entity Loss (MGRM-EL), to thoroughly explore the unimodal feature learning of entities and relationships during the cross-modal model learning process, enhancing the unimodal learning capability of the RSCTIR model and eliminating unnecessary redundancy across modalities. The work [29] proposed an end-to-end GLISA model that captures the overall context and fine-grained details by combining global semantic features and local region-word information. There are also works addressing the semantic gap in remote sensing text-image cross-modal retrieval, for example, the work [30] proposed a method called Text-Guided Knowledge Transfer (TGKT) for remote sensing image-text retrieval. They utilize textual information to bridge the semantic gap between the domains of remote sensing images and natural images. Although some achievements have been made in this field, previous methods still struggle to address the relationships between different modalities, leading to unsatisfactory retrieval accuracy.

Early remote sensing image-text retrieval techniques relied on generating descriptive text for images based on titles and calculating the similarity between the generated text and the query text. For instance, work [31] introduced a sound-driven attention network to produce targeted captions reflecting the observer's interests, while work [32] developed a support vector machine-based decoder to mitigate the high computational demands of caption generators when training data is limited. Although caption-based methods for remote sensing image-text retrieval have achieved relative maturity, their inherent two-stage process often results in information loss, significantly limiting model performance and accuracy.

With the advent of deep learning, embedding-based one-stage methods have become mainstream. These approaches directly map images and text into the same high-dimensional space, calculate feature distances, and minimize information transformation loss. Among multi-modal approaches, the CLIP [13] model is particularly notable, leveraging a large number

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

of image-text pairs for pre-training and achieving strong image-text retrieval performance in natural image domains. Adapting this to the remote sensing field, work [16] proposed RemoteCLIP, which re-trains the CLIP model on remote sensing data, yielding significant improvements. However, re-training such models demands substantial computational resources, making fine-tuning the prevailing approach. The effectiveness of fine-tuning directly impacts model performance, and developing efficient fine-tuning strategies remains a significant challenge requiring further exploration.

B. Prompt Engineering

The concept of prompt engineering was initially introduced to bridge the gap between pre-training and fine-tuning. By integrating prompts into the features learned during pre-training, studies [33], [34] were able to enhance the performance of downstream tasks without the need for fine-tuning steps. As pretrained language models (PLMs) continue to grow in size and capability, rapid learning has become a powerful paradigm for improving model capabilities. Early approaches [35], [36] primarily used manually generated discrete prompts to guide the model. However, due to the time-consuming and tedious nature of manually generated prompts, later research [37], [38] shifted towards automatically searching for discrete prompts tailored to specific tasks. Yet, these methods largely depended on the quality of the generated prompts. Recently, some techniques have begun using continuous, learnable embeddings as prompts [39]–[41] to achieve state-of-the-art performance. Additionally, in natural language processing, the work [42] used natural language prompts in zero-shot and few-shot scenarios. The work [37] treated prompt generation as a text generation task using T5 [36]. Some methods explored the use of continuous templates, operating directly in the model's embedding space. The work [43] employed gradient descent to tune the mix of soft templates in continuous word vectors. P-tuning,proposed by work [44], involves inserting continuous free parameters into the embeddings' input to automatically search for prompts. Furthermore, efforts have been made to inject knowledge into templates to enhance their performance. Another work [45] used sentiment knowledge to enhance prompts within a unified network. And work [46] integrated external knowledge into speech to improve the performance of prompt-based learning.

C. Remote Sensing Image-Text Retrieval and Prompt

In the context of remote sensing image-text retrieval, a "prompt" refers to an input designed to guide the model in generating meaningful associations between images and text. Prompts are crucial for aligning the representations of multimodal data, facilitating efficient and accurate retrieval by leveraging the model's pre-trained capabilities. Methods similar to "prompts" are employed in training cross-modal large models, including CLIP, BLIP, RemoteCLIP, and GeoRSCLIP. Specifically, CLIP, a large cross-modal pre-trained model designed for retrieval tasks, utilizes template-based text processing to handle image-text pairs, akin to the prompts discussed in this work. The text input for the CLIP model is typically formatted into a simple template, such as: "A photo of a [MASK]," where [MASK] is replaced by the relevant descriptive term. Models trained in this manner have demonstrated promising results in remote sensing text-image cross-modal retrieval. Drawing inspiration from these approaches, this paper integrates attention-based weak prompts with strong prompts derived from pre-trained classification models in the CLIP framework to enhance remote sensing text-image cross-modal retrieval. The proposed method effectively models and improves retrieval performance.

III. METHOD

In this section, we present the SWPE, which is tailored to bolster the performance of RS image-text retrieval tasks. Initially, we provide a comprehensive overview of the network, elucidating its underlying motivations and architectural framework. Subsequently, we delve into an in-depth examination of the detailed descriptions of the modules proposed within the architecture. Concluding this discussion, we articulate the loss functions that are pivotal for the training regimen of the network.

A. Overview of the Proposed SWPE

The structure of the SWPE is depicted in Fig.2. Initially, we employ a pre-trained feature extractor to explore the high-level feature representations of both images and text. The generated features then sequentially pass through three modules. The SWPG module is designed with two units comprising weak and strong prompts to achieve comprehensive prompting of global class semantic prompts and fine-grained prompts. The PFF module utilizes the attention mechanism within the Transformer architecture to transform the prompt information P and facilitate the integration of the prompt information. Finally, the AHSE loss is designed based on the similarity of cross-modal data, enabling the model to learn from easy to difficult samples.

The objective of cross-modal text-image retrieval is to establish a bidirectional retrieval process between text and images. This involves retrieving semantically similar samples from one modality based on a query sample from the other. Here, we formulate the aforementioned process for better explanation, specifically: Firstly, a large-scale dataset $D = (I, T)$ containing semantically relevant image text pairs is constructed. Then, a deep learning model is used to extract visual features $X \in R^{N \times D}$from images, semantic features $F \in R^{L \times D}$ from texts, and these features are unified into a shared semantic space for similarity comparison through feature mapping algorithms. During the model training phase, various methods are used to optimize the model, such as the loss function $\mathcal{L}$, to make its predicted output as close as possible to the labeled data, and a similarity measure is defined to evaluate the similarity of cross modal data pairs. Finally, performance evaluation is conducted through standardized indicators such as accuracy and recall to ensure that the model has good generalization ability and practical application effectiveness.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

B. Feature Extraction

We will provide a detailed explanation of feature representation from two aspects: image feature extraction and text feature extraction.

1) Image Feature Extraction : For the input image data $I \in R^{H\times W\times 3}$, we utilize a pre-trained RemoteClip to extract the high-level features of remote sensing images, as shown in equation (1):

$$X = \mathcal{VIT}\left(I\right) \tag{1}$$

where $I \in R^{H\times W\times 3}$ is the input remote sensing image, X is the set of visual feature from remote sensing image, and $\mathcal{VIT}\left(\cdot\right)$ is the operation of Vision Transformer.

2) Text Feature Extraction : For the input text data $T \in R^{L}$, we also used pre trained RemoteClip to extract the high-level features of remote sensing text, as shown in equation (2)

$$F = \mathcal{T}\left(T\right) \tag{2}$$

where $T \in R^{L}$ is the input remote sensing text, F_g is a set of remote sensing text features, and $\mathcal{T}\left(\cdot\right)$ is the operation of Transformer.

C. Strong and Weak Prompt Generation Module

The strong and weak prompt generation module mainly consists of two sub-units, each responsible for handling different tasks.

1) Weak Prompt Unit Based on Attention: The core idea of this unit is to further mine and extract the critical information in remote sensing image and text features through the attention mechanism. In remote sensing image-text retrieval tasks, the data often contains a large amount of complex and diverse background information, and may even include noise interference. Through the attention mechanism, the model can automatically learn the importance of different regions or parts of the image and text, thereby highlighting more representative and distinctive feature information. Specifically, we harness both self-attention and cross-attention mechanisms to distill more critical fine-grained information from the image and text features, while simultaneously mitigating the noise inherent in remote sensing imagery. The self-attention mechanism, an internal attention strategy, enables the model to focus on disparate segments of the data during the processing of remote sensing data, thereby capturing the long-range dependencies intrinsic to the data. Conversely, the cross-attention mechanism is particularly adept at handling the relationship between two sequences that are related yet distinct, which is especially beneficial in the context of text-image retrieval tasks. In our work on cross-modal retrieval in remote sensing text-image scenarios, the cross-attention mechanism facilitates the model's ability to reference information from a text when processing an image, thereby enabling the capture of the interplay between them. Subsequently, we optimize the prompts derived from both self-attention and cross-attention by assigning appropriate weights, which allows us to obtain the desired prompts.Below, we present the schematic formula for attention:

$$\mathcal{SA}\left(Q,K,V\right) = softmax\left(\frac{Q_s K_s^T}{\sqrt{d_k}}\right)V_s \tag{3}$$

$$\mathcal{CA}\left(Q,K,V\right) = softmax\left(\frac{Q_c K_c^T}{\sqrt{d_k}}\right)V_c \tag{4}$$

$$X^W = \mathcal{O}\left(\mathcal{SA},\mathcal{CA}\right) \tag{5}$$

where $\mathcal{SA}\left(\cdot\right)$ refers to self-attention, $\mathcal{CA}\left(\cdot\right)$ refers to cross-attention, $\mathcal{O}\left(\cdot\right)$ refers to the operation optimizing both $\mathcal{SA}$ and $\mathcal{CA}$, and X^W represents the prompt information output. In $\mathcal{SA}$, the input Q_s, K_s, and V_s are features from the same modality, such as image features or text features. In $\mathcal{CA}$, the input Q_c, K_c, and V_c from two different modalities, namely text features and image features, with Q_c coming from the text feature vector, while K_c and V_c come from the image feature vectors.

2) Strong Prompt Unit Based on Pre-Trained Classification Model : In this unit,we used a pre-trained image classification model as a strong prompt unit to obtain category prompt information. First, we trained the ResNet network as a classification network, then processed the input high-level image features and high-level text features, and finally output their category features. We use these output category features as global class semantic prompt.During the training process of the model, the parameters of the pre-trained classification model are frozen and do not participate in updates,the formula is shown as follows.

$$X^c = Conv\left(X,\theta_i\right) \tag{6}$$

$$F^c = Conv\left(F,\theta_i\right) \tag{7}$$

$$X^s = \mathcal{C}\left(X^c\right) \tag{8}$$

$$F^s = \mathcal{C}\left(F^c\right) \tag{9}$$

$$F^S = \mathcal{O}\left(X^s,F^s\right) \tag{10}$$

where X and F represent the high-level image features and high-level text features as inputs, respectively, θ_i is the parameter of the convolutional layer, and $\mathcal{C}\left(\cdot\right)$ represents the operation of the classifier for extracting category features.X^c and F^c are the image features and text features after being processed by the convolutional layer, while X^s and F^s are the category features of the image and text obtained from the classifier. F^S is the global category semantic prompt obtained after optimization, $\mathcal{O}\left(\cdot\right)$ refers to the operation optimizing both X^s and F^s, we use feature addition to optimize X^s and F^s.

Afterward, we perform a weighted combination of the prompt information obtained from the attention mechanism and the pre-trained model to derive the desired prompt information P, as shown in equation (11).

$$P = \frac{\omega_1 F^S + \omega_2 X^W + \omega_3 G}{\omega_1 + \omega_2 + \omega_3} \tag{11}$$

where ω_1, ω_2 and $\omega_3$represent the weights of F^S , X^W and G. G respectively.represents a learnable prompt information that we initialize with all parameters set to zero.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

D. Prompt-guided Feature Fine-tuning

In this module, the prompt information is fused with the high-level image features and high-level text features through Prompt-guided Feature Fine-tuning. We use the attention mechanism in the Transformer architecture to transform the prompt information P as follows:

$$P' = \mathcal{T}\left(\mathcal{SA}\left(P\right) + \mathcal{LN}\left(P\right)\right) \tag{12}$$

$$\hat{P} = \mathcal{T}\left(\mathcal{FFN}\left(\mathcal{LN}\left(P'\right)\right) + \mathcal{LN}\left(P'\right)\right) \tag{13}$$

where the self-attention operator $\mathcal{SA}\left(\cdot\right)$, feed-forward network $\mathcal{FFN}\left(\cdot\right)$ and layer normalization $\mathcal{LN}\left(\cdot\right)$ are applied to obtain the transformed prompts $\hat{P}$ .Specifically, for the text and image encoders at each i-th layer, we consider learning a set of layer-wise prompts $\hat{P}$, and then feed the transformed $\hat{P}$ into both the text and visual encoders.Since we utilize the attention mechanism in Transformer, the outermost function in equations (12) and (13) is $\mathcal{T}\left(\cdot\right)$. During model training, we simultaneously freeze the text and visual encoders, optimizing only the prompts P and the Transformer layers,as shown in equations (14) and (15).

$$X^P = \mathcal{T}\left(\hat{P}^i\right) \tag{14}$$

$$F^P = \mathcal{T}\left(\hat{P}^t\right) \tag{15}$$

where X^P, F^P represent the image features and text features fused through the Transformer approach after incorporating the prompt information $\hat{P}$,$x_1, x_2, x_3...x_i$ represents the high-level image features, $f_1, f_2, f_3...f_i$represents the high-level text features, and $\hat{P}$ represents the transformed prompt information.$\hat{P}^i$ is a new feature formed by the addition of $\hat{P}$ and $x_1, x_2, x_3...x_i$, $\hat{P}^t$ is a new feature formed by the addition of $\hat{P}$ and $f_1, f_2, f_3...f_i$.

E. Adaptive Hard Sample Elimination Loss

In this module,the image features X_P and text features F_P are used for similarity calculation. When input as a triplet sample, the positive pairs should be as close as possible, and the negative pairs should be as far apart as possible. More importantly, the distance between negative pairs should be greater than that of positive pairs with a fixed margin. Currently, the main method in Remote Sensing Cross-Modal Text and Image Retrieval(RSCTIR) is to use hinge loss (SH) for similarity matching, as shown in equations (16), (17), and (18).

$$T_{\hat{I}_n} = \left[\alpha - sim\left(I, S\right) + sim\left(\hat{I}_n, S\right)\right]_+ \tag{16}$$

$$T_{\hat{S}_n} = \left[\alpha - sim\left(I, S\right) + sim\left(I, \hat{S}_n\right)\right]_+ \tag{17}$$

$$L_{sh} = \sum_{\hat{I}_n} T_{\hat{I}_n} + \sum_{\hat{S}_n} T_{\hat{S}_n} \tag{18}$$

where S is the query statement and I is the paired image. ($\hat{I}_n$ is the nth unpaired image). θ is the margin parameter. $sim\left(I, S\right)$ represents the similarity of a positive pair, and $sim\left(\hat{I}_n, S\right)$ represents the similarity of a negative pair. $\left[x\right]_+ = max\left(x, 0\right)$. SH loss contains two symmetrical components. The first component iterates over all unpaired images ($\hat{I}_n$, given the query statement S), and the second component iterates over all unpaired sentences ($\hat{S}_n$, given the query image I). If the similarity between I and S is greater than any negative pair, and the margin is α, then the SH loss is zero. Previous studies have also proposed a Maximal Hinge (MH) loss based on SH loss, as shown in equation (19).

$$L_{mh} = max\left(T_{\hat{I}_n}\right) + max\left(T_{\hat{S}_n}\right) \tag{19}$$

Similar to SH loss, MH loss has two symmetric terms. Unlike SH loss, MH loss only considers the hardest negative pairs. When using SH loss, the process of gradient backpropagation can be described as follows:

$$\frac{\partial L_{sh}}{\partial \theta} = \frac{\partial\left(\sum_{\hat{I}_n} T_{\hat{I}_n} + \sum_{\hat{S}_n} T_{\hat{S}_n}\right)}{\partial \theta} \tag{20}$$

where θ is the parameter that needs gradient updates. $T_{\hat{I}_n}$ and $T_{\hat{S}_n}$ represent the difficulty of the triplet samples. The higher the value of $T_{\hat{I}_n}$, the more challenging the triplet data, meaning the lower the similarity of the positive pair (S, I) and the higher the similarity of the negative pair $\left(S, \hat{I}_n\right)$. For the query statement S, $\hat{I}_m = argmaxsim\left(\hat{I}_n, S\right)$ is the hardest unpaired image. SH loss considers all sample pairs equally. Regardless of the values of $T_{\hat{I}_n}$ and $T_{\hat{S}_n}$, the contribution of all sample pairs to gradient updates is the same. When MH loss is used, the process of gradient backpropagation can be represented as follows:

$$\frac{\partial L_{mh}}{\partial \theta} = 1 \times \frac{\partial T_{\hat{I}_m}}{\partial \theta} + 1 \times \frac{\partial T_{\hat{S}_m}}{\partial \theta} \tag{21}$$

However, for RSCTIR, the wide variation in RS image distribution and the high similarity in representing text distribution increase the alignment difficulty. Therefore, directly adopting traditional alignment strategies might confuse the network and hinder its generalization performance. To learn more effective parameters and achieve better generalization performance, this paper introduces an adaptive alignment strategy [47], specifically, the core idea is that sample pairs of different difficulties have different weights, and the network parameters θ are learned using an easy-to-hard strategy. Therefore, in gradient backpropagation, we assign greater weights to simpler sample pairs, aiming to learn simpler pairs first. We measure the difficulty of triplet samples through $T_{\hat{I}_{nK}}$ and $T_{\hat{S}_{nK}}$. First, we modify formula (19), where we arrange image features and text features in the form of a matrix, with the data on the main diagonal being the most matching pair of samples, and the rest being negative sample pairs. Here, we also rank the negative sample pairs and select the top K farthest-ranked negative sample pairs, as shown in formula (22):

$$L_{fmh} = \sum max\left(T_{\hat{I}_{nK}}\right) + \sum max\left(T_{\hat{S}_{nK}}\right) \tag{22}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

TABLE I
COMPARISONS WITH THE STATE-OF-THE-ART APPROACHES

Type	Method	Backbone	RSICD dataset							RSITMD dataset						
		vision encoding/text encoding	Image to Text			Text to Image			mR	Image to Text			Text to Image			mR
			R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
N	SCAN t2i(2018)	RoI Transformer/GRU	4.42	11.20	17.68	4.02	11.54	18.60	11.24	10.59	28.72	38.41	10.04	29.54	42.91	26.70
	SCAN i2t(2018)	RoI Transformer/GRU	5.90	13.21	19.96	3.86	16.83	26.49	14.38	10.84	25.86	37.26	9.62	29.80	41.06	25.74
	CAMP-triplet(2019)	RoI Transformer/GRU	5.22	13.05	21.02	4.30	17.10	27.57	14.71	11.82	27.4	38.12	8.69	27.17	43.60	26.13
	CAMP-bce(2019)	RoI Transformer/GRU	4.50	10.08	16.48	3.03	15.12	23.05	12.04	9.21	22.56	35.73	6.81	25.65	40.05	23.34
	MTFN(2019)	ResNet-18/GRU	5.01	12.86	21.30	4.96	12.14	29.12	14.23	10.80	27.68	36.4	9.82	30.28	48.27	27.21
	SGRAF(2021)	Faster R-CNN/Bi-GRU	3.93	13.36	22.78	3.42	14.53	26.72	14.12	5.97	18.81	30.09	6.02	24.25	41.81	21.16
	NAAF(2022)	ResNet-18/GRU	5.92	16.31	27.23	4.21	17.20	26.33	16.20	10.50	30.745	41.86	9.13	30.38	44.25	27.81
R	SAM i2t(2021)	CNN/RNN	6.41	20.10	32.89	5.91	18.08	29.73	18.85	13.27	30.09	44.91	9.87	33.89	50.66	30.45
	SAM t2i(2021)	CNN/RNN	5.91	18.59	30.63	4.82	18.10	33.44	18.58	11.73	27.25	40.06	11.07	35.03	50.94	29.35
	LW-MCR(2022)	SqueezeNet/RNN	3.66	13.63	21.68	3.77	15.32	28.20	14.38	6.64	17.92	30.31	5.53	25.51	44.34	21.71
	AMFMN(2022)	ResNet-18/GRU	5.39	15.32	28.82	5.05	18.19	29.07	16.97	11.37	27.70	39.29	9.04	32.46	49.68	28.26
	GaLR(2022)	ResNet-18+ppyolo/GRU	4.67	17.47	27.63	4.65	20.20	34.80	18.24	13.27	31.85	41.15	10.92	35.30	52.96	30.91
	MCRN(2022)	CNN/GRU	5.58	18.38	30.56	4.87	18.92	33.25	18.59	8.85	23.67	35.84	8.05	30.97	49.60	26.16
	CMFM(2023)	ResNet-18/GRU	5.40	16.47	27.54	4.15	17.99	31.25	17.13	8.85	23.89	37.17	7.12	31.95	50.88	26.64
	IEFT(2023)	Transformer/Bert	4.86	19.76	34.48	6.77	23.88	38.15	21.32	14.82	33.63	49.56	10.98	37.61	58.25	34.14
	PIR(2023)	Swin Transformer/Bert	10.52	26.44	39.52	6.83	23.97	39.09	24.40	19.03	41.15	53.32	13.72	42.48	62.88	38.76
	HVSA(2023)	CNN/RNN	8.60	22.78	33.21	5.95	22.53	34.57	21.27	14.6	32.96	46.46	10.84	38.01	57.35	33.37
	KAMCL(2023)	ResNet-18/GRU	8.14	22.14	33.39	6.35	23.70	38.87	22.10	14.82	29.65	43.58	10.53	35.80	55.84	31.70
C	Clip	Transformer/Transformer	10.52	25.25	39.16	9.33	30.01	46.04	26.72	14.60	36.28	49.56	14.42	43.65	64.65	37.19
	GLISA(2024)	Transformer/Transformer	**19.52**	**40.44**	52.28	14.75	39.50	55.46	36.99	**28.21**	50.94	62.50	20.50	54.80	72.19	48.14
	RemoteClip(2024)	Transformer/Transformer	16.65	35.13	47.39	14.73	40.14	56.96	35.17	27.21	50.22	63.93	24.20	57.21	74.96	49.62
	SWPE(our)	Transformer/Transformer	18.66	39.52	**53.61**	**15.33**	**40.86**	**57.73**	**37.62**	27.88	**51.76**	**64.82**	**25.27**	**58.23**	**75.27**	**50.54**

where $\hat{I}_{nK}$ refers to the K unpaired images from n unpaired images, and $\hat{S}_{nK}$ refers to K unpaired sentences from n unpaired sentences. $T_{\hat{I}_{nK}}$ and $T_{\hat{S}_{nK}}$represent the difficulty of these triplet samples. The process of gradient backpropagation can be described as follows:

$$\begin{aligned}\frac{\partial L_{dwh}}{\partial \theta} = & \sum_{\hat{I}_{nK}\neq\hat{I}_{mK}} \left(1-\beta\sum\max\left(T_{\hat{I}_{nK}}\right)\right)e^{-\beta\sum\max\left(T_{\hat{I}_{nK}}\right)} \\ & \times \frac{\partial\sum\max\left(T_{\hat{I}_{nK}}\right)}{\partial\theta} \\ & + \sum_{\hat{S}_{nK}\neq\hat{S}_{mK}} \left(1-\beta\sum\max\left(T_{\hat{S}_{nK}}\right)\right)e^{-\beta\sum\max\left(T_{\hat{S}_{nK}}\right)} \\ & \times \frac{\partial\sum\max\left(T_{\hat{S}_{nK}}\right)}{\partial\theta} \\ & + \left(1-\beta\sum\max\left(T_{\hat{S}_{nK}}\right)\right)e^{-\beta\sum\max\left(T_{\hat{I}_{nK}}\right)} \\ & \times \frac{\partial\sum\max\left(T_{\hat{S}_{nK}}\right)}{\partial\theta} \\ & + \left(1-\beta\sum\max\left(T_{\hat{S}_{nK}}\right)\right)e^{-\beta\sum\max\left(T_{\hat{S}_{nK}}\right)} \\ & \times \frac{\partial\sum\max\left(T_{\hat{S}_{nK}}\right)}{\partial\theta}\end{aligned} \tag{23}$$

In this context, the hyperparameter β represents the model's emphasis on simpler samples. Simpler sample pairs have greater weights, thus contributing more to gradient backpropagation. Therefore, the K-fold Difficulty Weighted Hinge (K-DWH) loss is proposed:

$$\begin{aligned}L_{K-dwh} = & \sum_{\hat{I}_{nK}}\left(\sum\max\left(T_{\hat{I}_{nK}}\right)e^{-\beta\sum\max\left(T_{\hat{I}_{nK}}\right)}\right) \\ & + \sum_{\hat{S}_{nK}}\left(\sum\max\left(T_{\hat{S}_{nK}}\right)e^{-\beta\sum\max\left(T_{\hat{S}_{nK}}\right)}\right)\end{aligned} \tag{24}$$

Finally, the loss function of the model is used to train the model combined with the commonly used triplet loss and difficulty-weighted hinge function.The triplet loss aims to lessen the distance between a sample and its positive counterpart, while enlarging the distance between the sample and its negative counterpart. The difficulty-weighted hinge function filters simpler sample pairs for learning first, optimizing the distance between the sample and its positive counterpart.The image-text triplet loss is as follows:

$$\begin{aligned}L_{tri} = & \sum_{T_n} relu\left(\alpha - S(I,T) + S\left(I,T_n\right)\right) \\ & + \sum_{I_n} relu\left(\alpha - S(I,T) + S\left(I_n,T\right)\right)\end{aligned} \tag{25}$$

Therefore, the total loss function of the model is as follows:

$$L = \lambda L_{tri} + \mu L_{K-dwh} \tag{26}$$

where λ and μ are the dynamic equilibrium factors of the loss functions L_{tri} and L_{K-dwh}.

IV. EXPERIMENTS

To evaluate the optimal performance of SWPE, we conducted extensive experiments on four remote sensing image-text datasets. We also conducted an ablation study to verify the effectiveness of each module.

A. Experimental Settings

Datasets. The experiment were carried out on four benchmark remote sensing datasets: RSICD, RSITMD, UCM, and Sydney. The RSICD dataset is the largest, containing 10,921 low-quality remote sensing images of size 224 × 224 pixels. The RSITMD dataset has 4,743 low-quality remote sensing images divided into 32 different categories. The UCM dataset includes 2,100 low-quality remote sensing images of size 500 × 500 pixels, with 21 different categories. Sydney is the smallest dataset, comprising 613 low-quality remote sensing images of size 500 × 500 pixels, divided into 7 different categories. Each image is described with five sentences.

Evaluation Metrics. We selected two metrics, R@K and mR, to evaluate our model. R@K, which includes R@1, R@5, and R@10, represents the proportion of ground truths appearing in the top K results. mR calculates the average of all R@K values, providing a more reasonable assessment of our model's performance.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

TABLE II
COMPARISONS WITH THE STATE-OF-THE-ART APPROACHES

Type	Method	Backbone	UCM							SYDNEY						
		vision encoding/text encoding	Image to Text			Text to Image			mR	Image to Text			Text to Image			mR
			R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
N	SCAN t2i(2018)	ResNet/GRU	13.89	45.81	68.59	13.11	50.69	78.21	45.05	19.14	50.71	74.12	17.81	55.94	76.57	49.05
	SCAN i2t(2018)	ResNet/GRU	12.81	46.99	69.11	12.59	46.89	72.71	43.52	20.42	54.16	67.62	16.13	57.61	75.98	48.65
	CAMP-triplet(2019)	ResNet/GRU	11.20	44.29	65.72	9.89	46.11	77.29	42.42	22.83	50.46	75.88	15.31	43.14	70.39	46.34
	CAMP-bce(2019)	ResNet/GRU	15.09	47.18	68.59	11.93	47.22	76.98	44.50	15.64	49.65	71.33	11.59	51.34	76.16	45.95
	MTFN(2019)	ResNet/GRU	10.68	47.62	64.31	14.74	52.65	81.13	45.19	21.71	51.58	69.03	14.13	55.96	78.62	48.51
	SGRAF(2021)	ResNet/Bi-GRU	4.76	36.19	63.81	10.76	48.67	78.38	40.43	15.52	43.10	62.07	7.93	46.55	73.10	41.38
	NAAF(2022)	ResNet/GRU	12.38	49.52	64.76	11.52	44.38	66.29	41.48	24.14	56.90	68.97	17.24	46.90	75.86	48.34
R	SAM i2t(2021)	CNN/RNN	13.78	46.77	67.15	11.32	49.27	81.81	45.02	24.12	55.17	75.93	13.41	54.46	76.62	49.95
	SAM t2i(2021)	CNN/RNN	11.92	45.23	70.95	11.42	53.41	83.12	46.01	27.64	56.87	72.35	19.01	57.93	79.96	52.29
	LW-MCR(2022)	SqueezeNet/RNN	13.81	40.10	59.05	10.38	44.86	74.95	40.53	17.24	43.10	63.79	16.21	56.21	75.17	45.29
	AMFMN(2022)	ResNet/GRU	14.48	51.17	67.35	14.49	51.82	80.71	46.67	20.69	50.73	75.23	13.45	60.00	82.70	50.47
	MCRN(2022)	CNN/GRU	12.38	44.29	68.09	11.71	50.48	92.76	46.62	25.86	55.17	65.52	15.52	54.83	80.34	49.54
	CMFM(2023)	ResNet/GRU	13.33	50.00	66.67	11.43	49.24	74.57	44.21	17.24	51.72	62.07	15.17	55.52	81.38	47.18
	IEFT(2023)	Transformer/Bert	11.90	47.62	75.71	13.63	53.03	90.45	48.72	22.41	63.79	72.41	18.59	54.65	79.55	51.90
	HVSA(2023)	CNN/RNN	14.29	39.52	59.52	13.71	48.48	73.33	41.48	13.79	48.28	62.07	14.48	55.52	77.24	45.23
C	Clip	Transformer/Transformer	17.63	51.43	74.29	14.57	57.05	94.95	51.65	22.41	55.17	67.24	16.90	56.55	80.34	49.77
	RemoteClip(2024)	Transformer/Transformer	15.71	50.95	80.47	17.71	63.33	**98.00**	54.36	20.69	**56.89**	**74.14**	21.03	64.83	82.41	53.33
	(SWPE)our	Transformer/Transformer	**26.19**	**68.10**	**86.19**	**22.29**	**66.48**	93.24	**60.41**	**24.13**	53.45	70.69	**22.41**	**65.51**	**86.66**	**53.79**

Implementation Details. All experiments were conducted on an RTX 3090 graphics card. For image feature extraction, we set the input image size to 256 × 256 pixels and used the Swin Transformer as the backbone to extract image features. For the remote sensing image feature extraction module, the dimension of the visual embeddings was set to 512. For text feature extraction, the dimension of the word embeddings was set to 300. We trained our model for 50 epochs using the Adam optimizer with a learning rate of 0.00001 and a batch size of 25.The value of K in the previous loss function is set to 5.

TABLE III
ABLATION EXPERIMENTAL RESULTS

Approach	Image-to-Text			Text-to-Image			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
SWPE-Base	16.65	35.13	47.39	14.73	40.14	56.96	35.17
SWPE-WPU	17.84	38.33	52.06	14.27	36.93	52.61	35.34
SWPE-SPU	18.48	38.88	51.51	14.53	38.12	53.91	35.90
SWPE-AHSE	16.93	36.23	49.86	15.01	40.26	57.93	36.04
SWPE-WPU&SPU	16.47	35.41	51.33	15.70	41.15	56.98	36.17
SWPE-WPU&AHSE	17.47	37.79	50.41	15.74	40.55	57.42	36.56
SWPE-SPU&AHSE	18.57	37.88	52.06	15.86	40.82	58.18	37.23
SWPE	**18.66**	**39.52**	**53.61**	**15.33**	**40.86**	**57.73**	**37.62**

B. Comparisons With the State-of-the-Art Approaches

In this section, we compare our method with several baselines, including SCAN [48], CAMP [49], MTFN [50], SGRAF [51], NAAF [52], SAM [24], LW-MCR [53], AMFMN [54], GaLR [27], MCRN [55], CMFM [56], IEFT [57], PIR [5], HVSA [47], KAMCL [4] ,GLISA [29] and RemoteClip [16]. Additionally, we have labeled the application domains of the baseline methods: "N" denotes that the method is applied to natural image-text retrieval, "R" indicates that the method is applied to remote sensing image-text retrieval, and "C" represents methods using CLIP. The experimental results are shown in Tables I and II.

1)Results on RSICD: The experimental results on the RSICD dataset are located in the left part of Table I. The results indicate that although the R@1 and R@5 in Image-to-Text are slightly lower than GLISA, SWPE performs the best in all other metrics, including mR and SWPE. Compared to the previously best-performing retrieval network GLISA, SWPE's mR improved by **0.63**, which is a 1.70% increase. In a detailed analysis, our model achieved the most significant improvement in the R@10 metric for the Image-to-Text task. Compared to GLISA, the R@10 score increased by **1.33** points, which is a 2.54% improvement, demonstrating the superiority of the model.

TABLE IV
ABLATION STUDY OF HYPER-PARAMETER β

β	Image-to-Text			Text-to-Image			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
0.2	18.02	37.87	51.32	15.71	40.31	55.68	36.48
0.4	18.93	39.24	52.79	15.80	41.20	57.43	37.57
0.6	18.66	39.52	53.61	15.33	40.86	57.73	**37.62**
0.8	19.21	39.06	52.69	15.33	40.91	57.25	37.41

2)Results on RSITMD: The results on the RSITMD dataset are located in the right section of Table I. Apart from the R@1 metric for Image-to-Text, our model has achieved the best results in all other metrics.Compared to the previously best-performing retrieval network RemoteClip, SWPE's mR has improved by **0.92**.Although there were no significant improvements in the metrics of R@1, R@5, and R@10, our model still managed to achieve the best results among all current state-of-the-art methods, demonstrating the superiority of this model.

3)Results on UCM: The experimental results on the UCM dataset are located in the left section of Table 2. Although the RemoteClip method achieved the best R @ 10 metric in text to image, our method outperforms all baselines in other evaluation metrics. Importantly, compared to the previous best performing cross modal retrieval method, mR has improved by **6.05**, which provides a significant indication and strong evidence of the effectiveness of SWPE.Make a concrete analysis, our model's results on this dataset have improved by **11.13**$\%$compared to the best RemoteCLIP results in the comparative experiments. We have conducted multiple experiments on this dataset, with an mR consistently above **60**. We attribute this to the effectiveness and stability of the SWPE model we proposed, especially considering that this dataset contains fewer images and more categories, the strong prompt unit based on the pre-trained classification model in

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

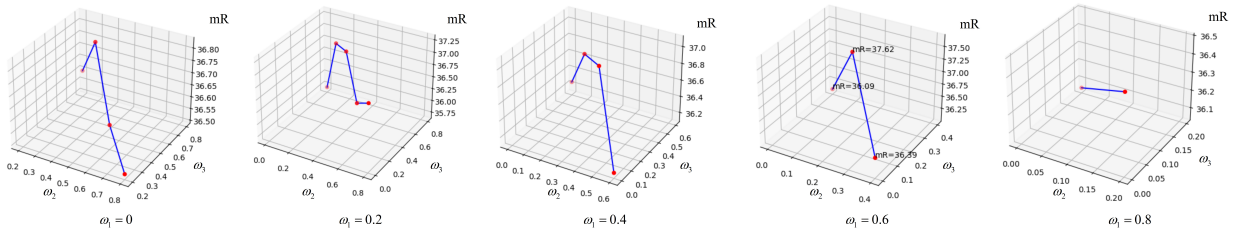

Fig. 3. This figure shows an ablation study on the weight parameter ω. It can be observed from the figure that the best experimental results are achieved when ω_2=0.6, ω_2=0.2, and ω_3=0.2.

our model has played a significant role in this dataset.

4)Results on Sydney: The results on the Sydney dataset are located in the right section of Table 2.Although the Remote Clip method achieved the best R@5 and R@10 metrics in Image-to-Text, our approach significantly outperformed Remote Clip in the R@1 metric for Image-to-Text, with a **3.44** higher score, which is a **16.62**% increase. In Text-to-Image, our method's R@10 score was **4.25** higher than Remote Clip's, representing a **5.15**% increase. Moreover, compared to the best-performing retrieval network Remote Clip, SWPE's mean Reciprocal Rank (mR) improved by **0.46**, demonstrating the superiority of our model.

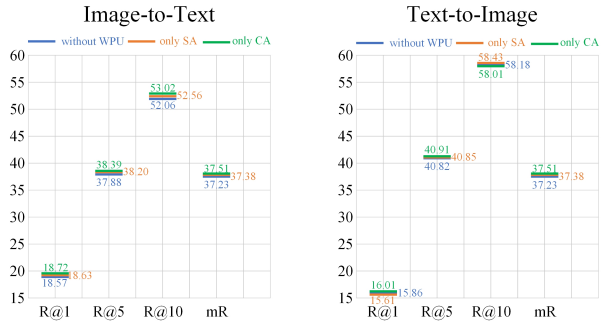

Fig. 4. Ablation of Self-Attention and Cross-Attention in WPU

C. Ablation Studies

In this section, to investigate the performance of each module, we conducted ablation experiments on SWPE. The results are shown in Table III, where SWPE-Base is the baseline model without any added modules. SWPE-WPU represents the addition of only weak prompt units based on attention, SWPE-SPU represents the addition of only strong prompt units based on pre-trained classification models, SWPE-AHSE represents the addition of only the adaptive hard sample elimination loss, SWPE-WPU&SPU represents the addition of both weak prompt units based on attention and strong prompt units based on pre-trained classification models, SWPE-WPU&AHSE represents the addition of both weak prompt units based on attention and the adaptive hard sample elimination loss, SWPE-SPU&AHSE represents the addition of both strong prompt units based on pre-trained classification models and the adaptive hard sample elimination loss, and SWPE represents the complete model.The results presented in Table III demonstrate that each module contributes to the overall performance of the proposed model, confirming their effectiveness and superiority. A detailed analysis indicates that the WPU module provides the most significant performance improvement. This improvement is attributed to the WPU module's ability to extract semantic category features from complex remote sensing images and utilize their category information as global prompts to optimize the model.

D. Ablation study of hyper-parameter β

We conducted an ablation study on the hyper-parameter β, which is used to control the model's focus on simple samples. When the value of β is high, the model pays more attention to simple samples and assigns less weight to difficult ones. The purpose of this design is to allow the model to start learning from simple samples and gradually transition to complex ones, thereby enhancing the model's generalization ability. The results of our ablation experiments are presented in Table IV, from which we can see that the best experimental results are achieved when β equals 0.6. Based on this, we can conclude that focusing on simple samples can effectively improve the model's generalization ability.

E. Ablation study of hyper-parameter ω

We conducted an ablation study on ω_1, ω_2, and ω_3, which are parameters controlling the weights of F^S, X^W, and G, respectively, to verify how to effectively allocate the weights of prompt information in our model. The experimental results are shown in Fig.3 We can see that the best experimental results are achieved when ω_1=0.6, ω_2=0.2, and ω_3=0.2. This outcome also validates the effectiveness of our strong prompt unit.

F. Ablation of Self-Attention and Cross-Attention in WPU

We conducted ablation studies on self-attention and cross-attention in weak prompt unit, and the experimental results are shown in Fig. 4. The blue represents the model without weak prompt units, and its ablation study results are also the 7th row of data in Table III. To more clearly demonstrate the improvement effects of self-attention and cross-attention on the model, we have also presented them. Orange represents the use of only self-attention in the weak prompt unit, and green represents the use of only cross-attention in the weak prompt unit. From the experimental results, we can see that cross-attention has a more significant improvement on the model. We analyze that the interaction of different modality information

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

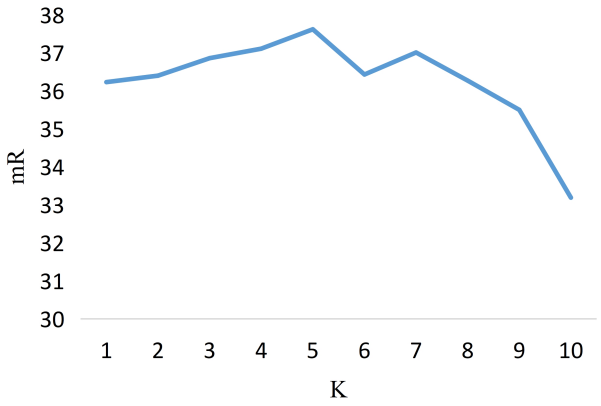

Fig. 5. This line chart shows the experimental results for different values of K. From the chart, we can see that the best experimental result is achieved when K is set to 5.

is more conducive to improving the model's performance. However, as can be seen from the last row of data in Table III, the complete weak prompt unit with both effects yield the best experimental results. We analyze that the reason is that self-attention more deeply mines the salient information within the modality, and cross-attention enhances the perception of information between modalities. The combined effect of both leads to an effective enhancement of the model's performance.

G. The validity of top K taking different values

In the adaptive hard sample elimination loss section, we propose selecting the top K farthest negative sample pairs from the ranking of negative samples. We conducted experiments regarding the specific value of top K to verify the validity of different top K values. The experimental results are shown in Fig.5.

H. Complexity Analysis

We also conducted complexity analysis experiments, as shown in Table V. "Trainable Parameters" refers to the total number of parameters that can be trained in the network, and "Time" refers to the average time required to train the network on a sample pair. We analyzed several natural domain-oriented methods ("N"), remote sensing domain-oriented methods ("R"), and methods using pretrained large models ("C"). It can be seen from the table that the method without pre training large model has fewer parameters and requires less time to train the network. However, the method of pre training large model takes a relatively long time to train, but in general, our trainable parameters are not large, and the model still has good effect. However, based on these data alone, it is difficult to clearly observe the performance of the model. Therefore, we use time/parameters to measure the model. It can also be seen from the specific data in the table that although our model is based on remoteclip and has 57M trainable parameters, it achieves the best effect in terms of time/parameters. In addition, although our model adds 57M trainable parameters, the increase of training time is still in an acceptable range. Considering the improvement of model performance, the increase of model training time becomes negligible. Therefore, we can confirm that our model has achieved a balance between effectiveness and complexity

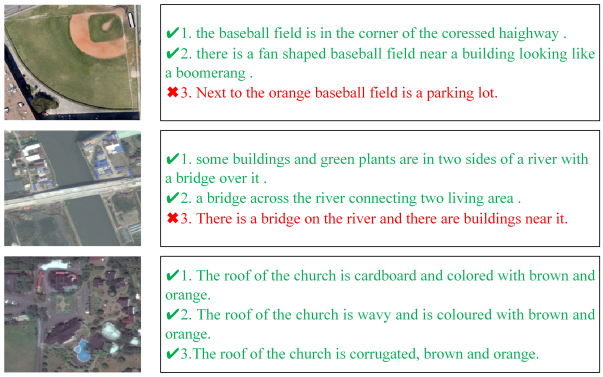

Fig. 6. Top-3 image-to-text retrieval results on RSITMD dataset. The groundtruth texts are marked with green checks, and the wrong results are indicated by cross marks.

Fig. 7. Top-3 text-to-image retrieval results on the RSITMD dataset. We have annotated the matched images with green bounding boxes.

I. Qualitative Analysis Results

In this subsection of the experiment, we visualize partial results of remote sensing text-to-image retrieval to analyze the effectiveness of SWPE. The results of image-to-text retrieval are shown in Fig.6. From the figure, we can observe that SWPE can accurately retrieve short or complex sentences corresponding to the images, mainly attributed to the incorporation of the classification prompting module in SWPE. Additionally, Fig.7 presents the results of text-to-image retrieval, indicating that corresponding images can be retrieved based on given text, demonstrating that SWPE can adapt well to images of different categories.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

TABLE V
COMPLEXITY ANALYSIS

Type	Method	Trainable Parameters	Time(s)	Time(s)/Parameters
N	SCAN	37M	0.0081	0.00021892
	CAMP	38M	0.0061	0.00016053
	MTFN	69M	0.0057	0.00008260
	SGRAF	28M	0.0085	0.00030357
	NAAF	23M	0.0058	0.00025217
R	SAM	17M	0.0474	0.00278824
	LW-MCR	12M	0.0024	0.00020000
	AMFMN	37M	0.0025	0.00006757
	GaLR	47M	0.0043	0.00009149
	MCRN	52M	0.0037	0.00007115
	CMFM	41M	0.0026	0.00006341
	HVSA	34M	0.0025	0.00007353
C	RemoteCLIP	0M	0.0114	0.00002670
	SWPE	57M	0.0123	**0.00002480**

V. CONCLUSION

Our study proposed a Strong and Weak Prompt Engineering (SWPE) for remote sensing image-text cross-modal retrieval, aimed at addressing the challenges posed by existing methods that emphasize either global or local information and fail to effectively mine or fully utilize valuable information in cross-modal data. By introducing weak prompts based on attention mechanisms, we have successfully achieved the extraction of fine-grained information from remote sensing text and images, thereby reducing the impact of information loss and background noise in the images. At the same time, using strong prompts based on pre-trained classification models to obtain global category semantic prompts effectively enhances the model's perception. Then, the prompt guided feature fine tuning module was designed, using transformer architecture to transform the prompt information and fuse the transformed prompt information with the extracted high-order image features and high-order text features to enhance the more important fine-grained information and global semantics in high-order features. Finally, we designed the adaptive hard sample elimination loss module, where we optimized the triplet loss function to train the model using negative sample pairs of different difficulty levels, assigning greater weights to simpler sample pairs. Experimental results show that our proposed method outperforms current state-of-the-art methods across multiple remote sensing datasets. This not only validates the effectiveness of our method but also provides new ideas and approaches for research in the field of cross-modal text-image retrieval. Future work will focus on addressing more complex cross-modal matching problems and continually refining the proposed method to meet evolving real-world demands.

REFERENCES

[1] Yingjie Liu, Xiaofeng Li, and Yibin Ren, "A deep learning model for oceanic mesoscale eddy detection based on multi-source remote sensing imagery," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 6762–6765.

[2] Keiller Nogueira, Samuel G Fadel, Ícaro C Dourado, Rafael de O Werneck, Javier AV Muñoz, Otávio AB Penatti, Rodrigo T Calumby, Lin Tzy Li, Jefersson A dos Santos, and Ricardo da S Torres, "Exploiting convnet diversity for flooding identification," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 9, pp. 1446–1450, 2018.

[3] Qingling Zhang and Karen C Seto, "Mapping urbanization dynamics at regional and global scales using multi-temporal dmsp/ols nighttime light data," *Remote Sensing of Environment*, vol. 115, no. 9, pp. 2320–2329, 2011.

[4] 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*, vol. 61, pp. 1–13, 2023.

[5] 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*, 2023, pp. 611–620.

[6] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[7] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*. PMLR, 2015, pp. 1083–1092.

[8] Fangxiang Feng, Xiaojie Wang, and Ruifan Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 7–16.

[9] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint arXiv:1607.06215*, 2016.

[10] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 154–162.

[11] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng, "Deep supervised cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10394–10403.

[12] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs, "Generalized multiview analysis: A discriminative latent space," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2160–2167.

[13] 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*. PMLR, 2021, pp. 8748–8763.

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12888–12900.

[15] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[16] 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.

[17] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin, "Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model," *arXiv preprint arXiv:2306.11300*, 2023.

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[19] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[20] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy, "Unified vision and language prompt learning," *arXiv preprint arXiv:2210.07225*, 2022.

[21] Huaying Zhang, Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama, "Parameter-efficient tuning of cross-modal retrieval for a specific database via trainable textual and visual prompts," *International Journal of Multimedia Information Retrieval*, vol. 13, no. 1, pp. 14, 2024.

[22] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang, "Dgl: Dynamic global-local prompt tuning for text-video retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 6540–6548.

[23] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang, "Vop: Text-video co-operative prompt tuning for cross-modal retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6565–6574.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2025.3534474

[24] Qimin Cheng, Yuzhuo Zhou, Peng Fu, Yuan Xu, and Liang Zhang, "A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4284–4297, 2021.

[25] Yijing Wang, Jingjing Ma, Mingteng Li, Xu Tang, Xiao Han, and Licheng Jiao, "Multi-scale interactive transformer for remote sensing cross-modal image-text retrieval," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 839–842.

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

[27] 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*, vol. 60, pp. 1–16, 2022.

[28] Shun Zhang, Yupeng Li, and Shaohui Mei, "Exploring uni-modal feature learning on entities and relations for remote sensing cross-modal text-image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[29] Gang Hu, Zaidao Wen, Yafei Lv, Jianting Zhang, and Qian Wu, "Global-local information soft-alignment for cross-modal remote-sensing image-text retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[30] An-An Liu, Bo Yang, Wenhui Li, Dan Song, Zhengya Sun, Tongwei Ren, and Zhiqiang Wei, "Text-guided knowledge transfer for remote sensing image-text retrieval," *IEEE Geoscience and Remote Sensing Letters*, 2024.

[31] Xiaoqiang Lu, Binqiang Wang, and Xiangtao Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1985–2000, 2019.

[32] Genc Hoxha and Farid Melgani, "A novel svm-based decoder for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[33] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang, "Ppt: Pre-trained prompt tuning for few-shot learning," *arXiv preprint arXiv:2109.04332*, 2021.

[34] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[35] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[37] Tianyu Gao, Adam Fisch, and Danqi Chen, "Making pre-trained language models better few-shot learners," *arXiv preprint arXiv:2012.15723*, 2020.

[38] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," *arXiv preprint arXiv:2010.15980*, 2020.

[39] Brian Lester, Rami Al-Rfou, and Noah Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[40] Zujie Liang, Huang Hu, Can Xu, Jian Miao, Yingying He, Yining Chen, Xiubo Geng, Fan Liang, and Daxin Jiang, "Learning neural templates for recommender dialogue system," *arXiv preprint arXiv:2109.12302*, 2021.

[41] Xu Guo, Boyang Li, and Han Yu, "Improving the sample efficiency of prompt tuning with domain adaptation," *arXiv preprint arXiv:2210.02952*, 2022.

[42] Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough, "Open aspect target sentiment classification with natural language prompts," *arXiv preprint arXiv:2109.03685*, 2021.

[43] Guanghui Qin and Jason Eisner, "Learning how to ask: Querying lms with mixtures of soft prompts," *arXiv preprint arXiv:2104.06599*, 2021.

[44] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang, "Gpt understands, too," *AI Open*, 2023.

[45] Chengxi Li, Feiyu Gao, Jiajun Bu, Lu Xu, Xiang Chen, Yu Gu, Zirui Shao, Qi Zheng, Ningyu Zhang, Yongpan Wang, et al., "Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis," *arXiv preprint arXiv:2109.08306*, 2021.

[46] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun, "Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification," *arXiv preprint arXiv:2108.02035*, 2021.

[47] 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.

[48] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201–216.

[49] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5764–5773.

[50] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song, "Matching images and text with multi-modal tensor fusion and re-ranking," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 12–20.

[51] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 1218–1226.

[52] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang, "Negative-aware attention framework for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15661–15670.

[53] Zhiqiang Yuan, Wenkai Zhang, Xuee Rong, Xuan Li, Jialiang Chen, Hongqi Wang, Kun Fu, and Xian Sun, "A lightweight multi-scale crossmodal text-image retrieval method in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[54] 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*, 2022.

[55] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Yongqiang Mao, Ruixue Zhou, Hongqi Wang, Kun Fu, and Xian Sun, "Mcrn: A multi-source cross-modal retrieval network for remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 115, pp. 103071, 2022.

[56] Hongfeng Yu, Fanglong Yao, Wanxuan Lu, Nayu Liu, Peiguang Li, Hongjian You, and Xian Sun, "Text-image matching for cross-modal remote sensing image retrieval via graph neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 812–824, 2022.

[57] Xu Tang, Yijing Wang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao, "Interacting-enhancing feature transformer for cross-modal remote-sensing image and text retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/