Parameter-Efficient Transfer Learning for Remote Sensing Image Captioning

Xuezhi Zhao, Zhigang Yang, Qiang Li, *Member, IEEE*, Qi Wang, *Senior Member, IEEE*

***Abstract*—Remote sensing image captioning (RSIC) aims to generate accurate and concise textual descriptions for remote sensing (RS) images. It plays a significant role in the analysis of earth observation data. The success of Vision-and-Language Pre-training (VLP) models provides the foundation for their transfer to the RSIC task. To reduce the cost of transferring VLP models to downstream tasks, numerous Parameter-Efficient Transfer Learning (PETL) techniques have been proposed. However, most of them focus on fine-tuning general-purpose foundation models without fully considering the unique characteristics of remote sensing data. In this paper, we introduce PE-RSIC, a novel PETL framework tailored for RSIC. Specifically, the framework builds on a pre-trained BLIP-2 model while further designing a lightweight Cross-modal RS adapter (CRS-Adapter) and a Class Prompt. During training, all parameters of the pre-trained model remain frozen, and the newly added CRS-Adapter modules are updated to efficiently transfer vision-and-language knowledge from the natural domain to the RS domain. The Class Prompt is obtained by projecting the vision-encoded [CLS] token into the decoder, guiding the model to generate more accurate captions. This approach enables the model to capture critical RS class features that might be lost during the query decoding process, with only a minimal increase in parameters. Extensive experiments show that our PE-RSIC framework outperforms full fine-tuning while utilizing only 5% of the trainable parameters.**

***Index Terms*—Remote sensing, image captioning, parameter-efficient transfer learning.**

I. Introduction

THE continuous advancement of Earth observation technology makes obtaining multi-source remote sensing (RS) data increasingly accessible. The tasks of extracting critical information from large-scale RS data remain a significant research focus in the RS field [1] [2] [3] [4]. Among these tasks, remote sensing image captioning (RSIC) has emerged as a prominent research hotspot. The goal of RSIC is to accurately interpret and describe contextual objects in remote sensing images using appropriate vocabulary, so that humans can intuitively grasp the key information contained within these images. This task holds significant application value across various fields, such as urban planning, disaster detection, military intelligence collection, and geographic information updating [5]. Traditional RSIC models typically use visual models for image feature extraction. During the text generation phase, language models are used to convert the visual feature vectors into corresponding textual descriptions. Moreover, the recent success of large-scale Vision-and-Language Pre-training (VLP) models [6] [7] has opened new possibilities for RSIC.

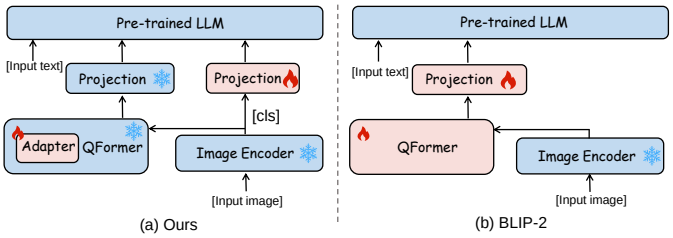


Fig. 1. Our approach builds upon BLIP-2 by augmenting QFormer with a set of trainable adapters. The learned query embeddings are merged with an additional encoded [CLS] token, providing the LLM with richer visual representations.

VLP combines the robust representation ability of visual pre-training models with the excellent reasoning capabilities of large language models (LLMs). By leveraging pre-training on large-scale unlabeled data [8], these models exhibit substantial potential for multi-modal representation and excellent transferability. Furthermore, parameter-efficient transfer learning (PETL) [9] has emerged as an effective strategy for adapting these models to specific vision-language downstream tasks. PETL typically achieves this by freezing the pre-trained model and introducing lightweight trainable parameters, enabling efficient adaptation to downstream tasks while significantly reducing memory consumption. However, existing methods primarily focus on downstream tasks in natural domains [10], which align more closely with pre-training data. This makes them suboptimal for the more complex RS domain, thereby limiting their application in RSIC. Therefore, there is a pressing need to develop PETL methods that can address the unique challenges of RSIC.

Specifically, PETL for RSIC faces the following challenges: Firstly, VLP in natural scenes exhibit a significant domain gap with the RS domain. To bridge this gap, it is necessary to design a method with fewer trainable parameters that can efficiently explore the unique knowledge of RS image-text pairs while fully inheriting the prior knowledge structure learned by VLP in natural scenes; Secondly, compared to natural images, RS images exhibit high similarity between different object categories and are often influenced by complex geographical backgrounds. These factors make it difficult to extract target features effectively, which subsequently impacts the generation of accurate captions; Finally, RS images often

This work was supported by the National Natural Science Foundation of China under Grant 62301385, 62471394, and U21B2041.

Xuezhi Zhao is with the School of Computer Science, and with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China (e-mail: xuezhizhao@mail.nwpu.edu.cn).

Zhigang Yang, Qiang Li, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail: zgyang@mail.nwpu.edu.cn, liqmges@gmail.com, crabwq@gmail.com) (Corresponding author: Qi Wang, Qiang Li.)

display significant multi-scale feature differences, with limited regions containing rich geospatial objects. Therefore, visual and textual semantic alignment is more challenging than in natural scenarios. Resolving this issue requires further designing stronger modal alignment mechanisms to improve caption generation ability.

To address the above issues, we propose a novel adapter-based framework for the RSIC task. Given the excellent representational capability of BLIP-2 in natural image captioning, we adopt it as the baseline. On the basis, we design a Cross-modal RS Adapter (CRS-Adapter) tailored to the RSIC task, enabling it to adapt to the unique characteristics of RS data and improve the alignment between visual and textual modalities. Based on extensive explorations, the design of our adapter includes the following key decisions: 1) Unlike traditional method of adding adapters after attention or FFN, we insert adapters in parallel within each transformer block of the QFormer and remove skip connections. This allows the adapter to more effectively learn domain differences while preserving the natural-domain prior knowledge of the pre-trained transformer; 2) To better capture the complex features in RS, we insert two parallel branches after multi-head self-attention module. This maintains the integrity of query information in the QFormer while learning cross-attention mixed information; 3) In order to enhance modal alignment for RSIC, we incorporate a gating mechanism in the upsampling layer of the adapter. This mechanism adjusts the information flow, facilitating the interaction between different modal information and adaptively emphasizes the most relevant and useful cross-modal relationships for image caption.

We obtain a pre-trained QFormer with an adapter that provides well-aligned RS visual embeddings for the subsequent frozen LLM text generator. However, these embeddings are constrained by the token count, which may lead to the loss of certain RS class-specific features. Classes form unique RS scenes and are directly relevant to image caption generation. To address this issue, we propose a Class Prompt generated from vision-encoded [CLS] tokens to better guide the LLM in generating RS captions, as shown in Fig. 1. Meanwhile, this approach requires minimal computation and memory overhead, which perfectly aligns with the lightweight nature of PETL methods. In conclusion, our contributions are as follows:

- The CRS-Adapter is proposed, which efficiently learns RS domain knowledge while adaptively enhancing modal alignment and inheriting pretrained prior knowledge.
- A Class Prompt is developed to help the model capture critical RS classes features, guiding the text decoder to produce more precise and effective RS captions.
- We propose PE-RSIC, a novel and effective PETL framework for RSIC. Extensive experiments on mainstream datasets validate the effectiveness of our approach.

The remainder of this paper is organized as follows: Sec. II reviews relevant studies. The proposed PE-RSIC approach is detailed in Sec. III. Sec. IV presents and analyzes the experimental results. Finally, the conclusion is in Sec. V.

II. Related Work

A. Remote Sensing Image Captioning

Since Qu et al. [11] introduce a pioneering multimodal architecture based on CNN-RNN frameworks for RSIC, numerous RSIC methods have emerged. Currently, mainstream approaches predominantly employ a unified encoder-decoder framework [12]. This framework typically utilizes CNN-based architectures like ResNet or transformer-based architectures like Vision Transformer(ViT) as image encoders to extract rich remote sensing visual features. The decoder usually employs structures such as RNN to translate visual embeddings into textual representations, thereby generating corresponding descriptive sentences. Moreover, some studies utilize transformers [13] as decoders to enhance the text generation capabilities of RSIC.

Remote sensing images contain a multitude of entities at various scales. To address this challenge, most methods apply attention mechanisms and feature aggregation strategies to enhance image feature representation [14]. Huang et al. [15] propose a multi-scale feature aggregation network and designed denoising operations to eliminate redundant information from the aggregated features. Yuan et al. [16] propose a framework that combines multi-level attention with a multi-label attribute graph convolution to handle scale variations and visual relationships more effectively. Zhang et al. [17] introduce a Label-Attention Mechanism to effectively leverage label information for guiding the computation of attention masks. Additionally, some studies focus on improving language decoders to generate higher-quality sentences. Li et al. [18] propose a multi-level attention model, adding three attention structures in the decoder to better mimic human attention, thereby enhancing the focus on both images and sentences. Yang et al. [19] introduce a noun sequence generation decoder to enhance the capability of the model to understand key image targets. However, these methods do not fully leverage the advantages offered by the VLP paradigm.

Recently, some studies have applied VLP to RSIC. BITA [20] introduces an Interactive Fourier Transformer to replace the conventional Transformer architecture, enabling more effective alignment of RS image-text features. RS-MoE [21] employs a Mixture of Experts framework, which uses dynamic task-specific prompts and LLMs to achieve higher scalability. However, these methods still require substantial computational resources, while PETL alleviate this limitation.

B. Vision-Language Pre-training

In recent years, advances in NLP have driven the progress of VLP research. VLP has shown encouraging results in various downstream tasks [22], mainly including end-to-end vision-language and modular vision-language pre-training.

The end-to-end pre-training method is an early VLP paradigm. According to the encoder type, it can be categorized into dual encoder architectures and fusion encoder architectures [23]. The fusion encoder processes image and text features as input and uses additional network branches for modality interaction. Among them, two-stream structures such as ViLBERT [24], ALBEF [25] and LXMERT [26] use two

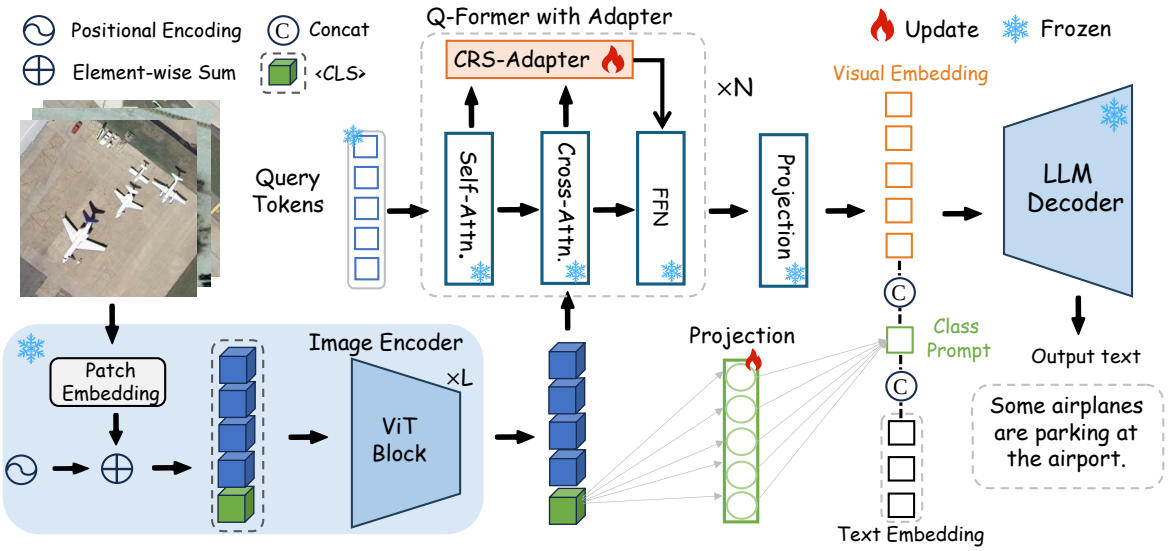


Fig. 2. The overall architecture of the proposed novel PE-RSIC framework, and it mainly consists of some key parts: the Vision encoder, QFormer, CRS-Adapter, Class Prompt and LLM decoder. The input image is processed by a pretrained Vision Transformer to extract visual features, which interact with the Q-Former integrated with a CRS-Adapter to generate visual embeddings. Subsequently, the vision-encoded [CLS] token is projected as a class prompt into the LLM. Finally, the visual embeddings are fed into the LLM to produce the final image-text description.

independent Transformer-based branches for extracting region and text features, along with another Transformer module for VL interaction. In contrast, OSCAR [27] and SimVLM [28] directly concatenate multimodal features and then input them into a single encoder. Dual encoder (CLIP [29] and ALIGN [30]) after getting the image text embedding from the two encoders, employ some simpler and more direct VL alignment methods such as contrastive learning.

Due to the excellent feature extraction ability of large visual models (LVMs) [29] [31] and superior text generation ability of LLMs [32], the Modular Vision-Language Pre-training method freezes both components and adds trainable modules to align visual features with the text space. The most representative ones are Flamingo [7] and BLIP-2 [6]. Specifically, Flamingo [7] injects visual features into the LLM generation process by cross-attention between the visual query and the layer newly inserted into the LLM, aligning arbitrarily interleaved visual and textual data sequences and achieving few-shot learning capabilities. BLIP-2 [6] uses a Query Transformer (QFormer) between frozen vision and language pre-trained models to bridge the modality gap. InstructBLIP [33] introduces instruction tuning to BLIP-2, improving its zero-shot generalization capabilities. MiniGPT-v2 [34] further advances this paradigm through task-specific identifiers.

C. Parameter-Efficient Transfer Learning

Parameter-Efficient Transfer Learning aims to adapt frozen pre-trained large models to specific downstream tasks by updating a small number of parameters, thereby improving the memory consumption problem in pre-trained models. It is first proposed in the field of NLP. Methods based on prompts [35] integrate vectors as prompts into the input to optimize the model. Based on CLIP, CoOp [36] first apply the learnable prompt to the CV field. However, since the generalization of the CoOp learned context is not insufficient, Zhou et al. [37] propose CoCoOp to extend CoOp by creating image-specific labels conditioned on the input. In order to combine the advantages of textual cue learning and visual cue learning, Wang et al. [38] propose a unified cue tuning (UPT) method by learning a micro neural network to jointly optimize prompts from different modalities.

Adapter-based methods [9] insert multiple adapters, each containing a small number of trainable parameters, into the model for fine-tuning. Recently, adapters have been widely used in CV and Cross-modal fields. AdapterFormer [39] is an adapter of the original linear layer based on the Vision Transformer, which can be effectively adapted to vision tasks. Building on CLIP-Adapter [40], Zhang et al. [41] propose Tip-Adapter, an untrained approach to enhance CLIP's few-shot capability. Sung et al. [10] propose a cross-modal adapter for multimodal domains, which provides an early transfer learning framework for VL tasks. Recent studies [42] enhance cross-modal interaction by designing better adapter structures.

In addition to the above methods, Low-Rank Adaptation (LoRA) [43] injects low-rank matrices into model weights, which reparameterizes the weight update matrix by the simple low-rank matrix. Ben Zaken et al. [44] propose BitFit, a fine-tuning method that updates only the bias terms of the model. However, multi-modal PETL of across domains remains underexplored. Existing PETL methods are suboptimal in the more challenging RS domain, as they are mainly tailored for downstream tasks with the same domain as pre-training.

III. METHOD

This section introduces the proposed PE-RSIC framework with its pipeline illustrated in Fig. 2. Firstly, we describe the model architecture in Sec. III-A, followed by the CRS-Adapter in Sec. III-B. Finally, we introduce the Class Prompt, Model training optimization and caption generation in Sec. III-C and Sec. III-D.

A. Model Architecture

Our model builds upon the fully pre-trained BLIP-2 framework, comprising a pre-trained vision encoder, a large language model (LLM), and a QFormer, further enhanced with the trainable CRS-Adapters and Class Prompt.

1) ***Vision Encoder***: We utilize the ViT as the visual encoder to extract the visual representation of the input image and initialize its parameters with CLIP pre-trained weights. Given an input image, we denote it as $I \in \mathbb{R}^{H\times W\times 3}$. Firstly, I is divided into $N \times N$ patches V_{patch}, and then an additional classification token V_{cls} is added to the token sequence. In addition, ViT adds the positional embedding V_{pos} to each token to get the initial visual embedding sequence V_0, which can be expressed as

$$V_0 = \left[V_{\text{cls}}; V_{\text{patch}}\right] + V_{\text{pos}}, \tag{1}$$

where V_{cls} denotes the added [CLS] token, V_{patch} denotes the patch tokens, and V_{pos} denotes the position embedding vector. The initial visual embedding sequence V_0 passes through L transformer blocks. The process of the L-th layer can be expressed as

$$\tilde{V}_l = \text{MSA}(\text{LN}(V_{l-1})) + V_{l-1}, \tag{2}$$

$$V_l = \text{FFN}(\text{LN}(\tilde{V}_l)) + \tilde{V}_l, \tag{3}$$

where $\text{MSA}(\cdot)$ represents the multi-head self-attention module, $\text{LN}(\cdot)$ represents Layer Normalization, and $\text{FFN}(\cdot)$ represents the Feed-Forward Network. $\tilde{V}_l$ denotes intermediate hidden features extracted by the $\text{MSA}(\cdot)$ module in the l-th ViT block, while V_l is the final output of the L-th block in the ViT. The [CLS] token from V_l is regarded as the global representation of the input image. Finally, the visual embedding V_e is obtained by applying normalization and projection to the V_l.

2) ***Query Transformer with Adapter***: QFormer is a Transformer-based structure to bridge the gap between image encoders and LLMs. We initialize it with the QFormer parameters from the pre-trained BLIP-2 model. The input of QFormer is a query embeddings, which interact with each other through self-attention layers. After stacking N-1 layers, the self-attention output feature F_q can be calculated, i.e.,

$$F_q = \text{MSA}(\text{LN}(F_{n-1})) + F_{n-1}. \tag{4}$$

The F_q then interacts with the visual embedding V_e obtained in the image encoder through the cross-attention layer (inserted into all other transformer blocks). This process can be expressed as

$$\tilde{F}_n = \text{MCA}(Q = F_q, K = V_e, V = V_e) + F_q, \tag{5}$$

where $\text{MCA}(\cdot)$ represents the multi-head cross-attention module. Specifically, in each cross-attention layer, the self-attention output features F_q serve as the Q, while the visual embedding V_e serves as both the K and V. The hidden state $\tilde{F}_n$ finally passes through a $\text{FFN}(\cdot)$, producing the final output F_n, i.e.,

$$F_n = \text{FFN}(\text{LN}(\tilde{F}_n)) + \tilde{F}_n. \tag{6}$$

This output F_n then passes through a projection layer to generate a more refined visual representation of the input image.

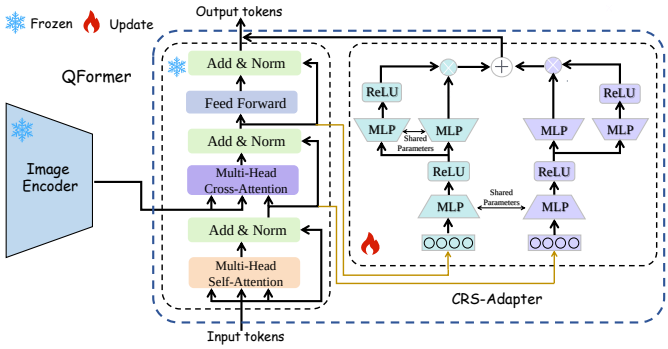


Fig. 3. The implementation details of our CRS-Adapter.It takes inputs from the $\text{MCA}(\cdot)$ and $\text{MSA}(\cdot)$ modules and processes them through parallel branches. After processing, the features are combined and then output.

This representation provides the most relevant information and ensures better alignment with the frozen LLM text generator, facilitating improved caption generation.

Moreover, considering the differences between RS image-text data and natural domain data, we propose the **CRS-Adapter**, which learns the specialized knowledge of RS image-text. The specific details are introduced in the Sec. III-B.

3) ***Large Language Model***: We connect QFormer with adapter to the frozen LLM to obtain the generative language ability of LLM. Specifically, the output query F_n is linearly projected to match the dimension of the LLM's input embedding. The projected vision embedding V_E is then added to the input text embedding. V_E acts as a soft prompt to condition LLM to generate image caption according to the visual representation extracted by QFormer. Meanwhile, to enrich the decoder with more comprehensive visual knowledge while considering computational cost, we introduce a Class Prompt, which will be detailed in the Sec. III-C.

B. CRS-Adapter

Building on the success of adapters [9] in NLP, an increasing number of adapter-based PETL approaches have demonstrated strong performance in downstream tasks. The adapter uses a down-projection linear layer to map the input features to a lower-dimensional representation, followed by a nonlinear activation function (typically ReLU), and then an up-projection linear layer to restore the dimension. Given the input $X \in \mathbb{R}^{n\times d}$, this process can be expressed as

$$\hat{X} = f\left(X \cdot W_{\text{down}}\right) W_{\text{up}} + X, \tag{7}$$

where $W_{\text{down}} \in \mathbb{R}^{d\times \hat{d}}$ represents the downsampling matrix, $W_{\text{up}} \in \mathbb{R}^{\hat{d}\times d}$ represents the upsampling matrix, $f(\cdot)$ represents the non-linear activation function, and $\hat{X} \in \mathbb{R}^{n\times d}$ represents the adapter features. To more effectively learn RS visual-text features without altering the pre-trained model structure, we propose a dual-branch parallel CRS-Adapter. It utilizes weight sharing and gating mechanism to control information flow, enhancing the ability of model to learn relevant multimodal relationships, as shown in Fig. 3.

*1) **Residual Learning for Adapter***: In most scenarios, the adapter is inserted after FFN, and Equation 6 can be modified as follows

$$F_n = Adapter(\mathrm{FFN}(\mathrm{LN}(\tilde{F}_n)) + \tilde{F}_n, \tag{8}$$

where F_n represents the output of the N-th layer of QFormer. The above methods have been proven to be effective in natural scenes. However, in order to adapt the pre-trained VLP to remote sensing data while inheriting the prior knowledge structure of natural scenes, we insert adapters in parallel and remove skip connections. This allows the adapter to learn domain differences more efficiently without changing the pre-trained transformers. Firstly, we insert Adapter in parallel after $\mathrm{MCA}(\cdot)$. This added branch learns remote sensing information extracted by the $\mathrm{MCA}(\cdot)$. It helps bridge the domain gap between the natural scene pre-trained model and remote sensing images, enabling the model to acquire more representative image features. Furthermore, the cross-attention process in QFormer may destroy the integrity of the query information. Therefore, we further introduce residual learning of the query to maintain the information by introducing additional adapters. Given equation 4, 5, equation 8 can be modified as

$$F_l = \mathrm{FFN}(\mathrm{LN}(\tilde{F}_l) + Adapter(\tilde{F}_l) + Adapter(F_q)). \tag{9}$$

*2) **Gating Mechanism with Knowledge Sharing***: Introducing a query adapter may lead to additional parameter updates, which we do not expect from a lightweight perspective. To minimize additional parameter costs, we implement a weight-sharing mechanism specifically in the downsampling layer of the two adapter branches. Surprisingly, this weight-sharing mechanism even leads to better performance. This is attributed to weight sharing facilitating the fusion of diverse information. As shown in Fig. 3, the input feature vector is projected to a smaller dimension and then passed through a ReLU activation function. Inspired by the success of gated linear units in the Feed-Forward layers of Transformers [45], we introduce a gating mechanism in the up-projection layer. Specifically, we compute the element-wise product of two linear transformations, denoted as W_u and W_g, where W_u is activated by ReLU. This gating mechanism helps the adapter control the information flow, potentially emphasizing the most useful and relevant multimodal relationships. Specifically, we denote the output of the $\mathrm{MSA}(\cdot)$ at the N-th layer $F_q \in \mathbb{R}^{n\times d}$ as $X^{\mathcal{T}}$ and denote the output of the $\mathrm{MCA}(\cdot)$ at the L-th layer $\tilde{F}_n \in \mathbb{R}^{n\times d}$ as X^S. The adaptation process can be written as

$$X^{\mathcal{T}}_{\text{down}} = \mathrm{ReLU}\left(X^{\mathcal{T}} \cdot W_d^{\text{share}}\right), \tag{10}$$

$$X^{\mathcal{T}}_{\text{Adapter}} = \mathrm{ReLU}(X^{\mathcal{T}}_{\text{down}} \cdot W_u) \otimes X^{\mathcal{T}}_{\text{down}} \cdot W_g, \tag{11}$$

$$X^{S}_{\text{down}} = \mathrm{ReLU}\left(X^{S} \cdot W_d^{\text{share}}\right), \tag{12}$$

$$X^{S}_{\text{Adapter}} = \mathrm{ReLU}(X^{S}_{\text{down}} \cdot W_u^{\text{share}}) \otimes X^{S}_{\text{down}} \cdot W_u^{\text{share}}, \tag{13}$$

$$X_{\text{Adapter}} = X^{S}_{\text{Adapter}} + X^{\mathcal{T}}_{\text{Adapter}}, \tag{14}$$

where X_{Adapter} represents the output of the CRS-Adapter, which is added to the FFN output.

C. Class Prompt

Due to the token number constraint in QFormer, some features of RS entity targets (i.e., classes) may be lost during feature alignment. These classes define unique scene semantics, which are critical for image caption generation. Therefore, we propose a class-aware prompt generated by projecting the vision-encoded [CLS] token. This prompt helps guide the model in capturing geospatial targets and understanding visual representations, improving the ability of LLM to generate remote sensing captions. Specifically, we extract the [CLS] token V_{cls} from the output V of the vision encoder, this process can be expressed as

$$V_{cls} = V[\mathrm{cls}]. \tag{15}$$

Then, the [CLS] token V_{cls} is fed into a fully connected projection layer to align its dimensions with the input embeddings of the LLM decoder:

$$C_{prompt} = \mathrm{MLP}(\mathrm{V_{cls}}). \tag{16}$$

Finally, the projected Class Prompt C_{prompt} is integrated into the LLM.

D. Training Optimization and Caption Generation

During training, images are input to the Vision Transformer to extract visual features, which interact with the CRS-Adapter-integrated QFormer to produce visual embeddings V_e. Subsequently, a fully connected layer linearly projects the vision-encoded [CLS] token to match the input dimensionality of the LLM, forming the Class Prompt C_{prompt}. The projected visual embeddings V_E and C_{prompt} are then concatenated as a prefix to the input text embeddings to generate the final image caption. Only the Class Prompt and CRS-Adapter are trainable during training, while all other components remain frozen. To optimize the model, we employ cross-entropy loss $\mathcal{L}_{lm}$, widely used for LLMs. This loss measures the difference between the predicted token probabilities and the ground-truth sequence Specifically, the loss function is defined as

$$\mathcal{L}_{lm} = -\sum_{t=1}^{T} \log(p_\theta(y_t^* \mid y_{1:t-1}^*)), \tag{17}$$

where $y_{1:t-1}^*$ denotes the ground-truth tokens preceding step t, T is the sequence length, and θ represents the model parameters. This loss function drives the model to generate accurate captions by reducing the gap between predicted and actual token distributions. During inference, we directly load the pre-trained model along with our trained adapter and projection layer parameters. Additionally, beam search is employed to improve the diversity of the generated captions to a certain extent.

IV. Experiments

This section begins by presenting the datasets and evaluation metrics utilized by the algorithm. Subsequently, the specific details of the implementation are provided. Meanwhile, we assess the algorithm's performance on three datasets and present several visualization examples. Finally, the findings from the ablation experiments are discussed.

TABLE I
EXPERIMENTS RESULTS ON UCM DATASET. THE BOLD AND UNDERLINED INDICATE THE BEST AND SECOND BEST RESULTS. THE ∗ DENOTES THE RESULTS OF OUR RE-IMPLEMENTED. THE GRAY SHADING REPRESENTS THE BASELINE.

Method	Params(M)	BLEU-l	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE
Traditional methods									
Word-Sentence [46]	13	79.31	72.37	66.71	62.02	43.95	71.32	278.71	-
GVFGA+LSGA [47]	-	83.19	76.67	71.03	65.96	44.36	78.45	332.70	48.53
MLCANet [48]	-	82.6	77.0	71.7	66.8	43.5	77.2	324.0	47.3
GLCM [49]	10	81.82	75.40	69.86	64.68	46.19	75.24	302.79	-
MLAT∗ [50]	161	82.26	75.39	69.59	64.24	42.68	77.29	308.26	-
HCNet [51]	258	88.26	83.35	**78.85**	**74.49**	48.65	83.91	351.83	-
Transformer-based methods									
Clipcap∗ [52]	32	82.13	74.88	68.72	63.20	42.32	77.42	323.91	-
PureT [53]	220	85.73	80.20	75.62	71.29	46.86	82.01	349.00	47.94
BLIP-2 [6]	188	88.04	82.23	76.89	71.84	47.32	83.36	380.31	54.78
BITA [20]	171	88.89	83.12	77.30	71.87	46.88	83.76	384.50	54.88
RSGPT [54]	-	86.12	79.14	72.31	65.74	42.21	78.34	333.23	-
Ours	9.6	**90.03**	**84.08**	78.17	74.33	**49.33**	**84.72**	**394.57**	**55.79**

TABLE II
EXPERIMENTS RESULTS ON NWPU DATASET. THE BOLD AND UNDERLINED INDICATE THE BEST AND SECOND BEST RESULTS. THE ∗ DENOTES THE RESULTS OF OUR RE-IMPLEMENTED. THE GRAY SHADING REPRESENTS THE BASELINE.

Method	Params(M)	BLEU-l	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE
Traditional methods									
MLCANet [48]	-	74.5	62.4	54.1	47.8	33.7	60.1	126.4	28.5
MLAT∗ [50]	161	85.27	76.74	70.07	64.79	43.27	74.96	185.56	-
Transformer-based methods									
Clipcap∗ [52]	32	83.94	74.21	66.19	59.54	41.42	73.85	172.65	-
PureT [53]	220	**88.82**	80.31	73.30	67.50	42.32	75.84	195.12	27.43
BLIP-2 [6]	188	87.78	80.24	73.51	67.62	45.01	78.44	193.86	33.28
BITA [20]	171	88.54	**80.70**	73.76	67.60	45.27	**78.53**	197.04	33.65
Ours	9.6	88.47	79.69	**74.15**	**67.66**	**45.35**	78.14	**199.74**	**33.98**

TABLE III
EXPERIMENTS RESULTS ON RSICD DATASET. THE BOLD AND UNDERLINED INDICATE THE BEST AND SECOND BEST RESULTS. THE ∗ DENOTES THE RESULTS OF OUR RE-IMPLEMENTED. THE GRAY SHADING REPRESENTS THE BASELINE.

Method	Params(M)	BLEU-l	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE
Traditional methods									
Word-Sentence [46]	13	72.40	58.61	49.33	42.50	31.97	62.60	206.29	-
GVFGA+LSGA [47]	-	67.79	56.00	47.81	41.65	32.85	59.29	260.12	46.83
MLCANet [48]	-	75.70	63.40	53.90	46.10	35.10	64.60	235.60	44.40
GLCM [49]	10	77.67	64.92	56.42	49.37	36.27	67.79	254.91	-
Transformer-based methods									
PureT [53]	220	77.05	65.75	56.63	49.19	37.66	67.41	275.56	48.87
BLIP-2 [6]	188	76.91	66.04	57.02	49.60	40.28	69.86	294.66	53.11
BITA [20]	171	77.38	66.54	57.65	50.36	**41.99**	71.74	304.53	**54.79**
Ours	9.6	**78.20**	**67.85**	**58.74**	**51.07**	41.02	**72.14**	**304.62**	54.16

A. Datasets

To evaluate the performance of the proposed method, experiments are performed on the UCM, NWPU, and RSICD datasets. A detailed explanation is provided below:

The UCM-Caption Dataset [55]: This dataset includes 2,100 images categorized into 21 land-use classes, including agricultural, aerial, beach, and dense residential. Each image has a resolution of 256 × 256 pixels and is provided in RGB format. In total, the dataset contains 10,500 unique annotations. The UCM-Caption Dataset is widely used for aerial image captioning research.

The NWPU-Caption Dataset [48]: This dataset contains 31,500 aerial images across 45 land-use categories, including airport, bridge, desert, and farmland. Each image is 256 × 256 pixels in size. The dataset includes five unique captions for each image, resulting in 157,500 human-written descriptions.

The RSICD Dataset [56]: This dataset comprises 10,921 remote sensing images sourced from different platforms. These images cover 30 land-use categories, such as airport, desert, farmland, and playground. Each image has a resolution of 224 × 224 pixels with some variations. To maintain annotation consistency, the dataset provides up to five captions per image, resulting in 54,605 sentences.

B. Evaluation Metrics

It is essential to evaluate the quality of generated captions for RSIC, and a variety of metrics are commonly employed in image captioning. In this study, we utilize BLEU1-4,

METEOR, ROUGE_L, CIDEr, and SPICE to comprehensively assess caption performance.

1) BLEU measures the n-gram overlap between generated and reference texts, focusing on precision. It evaluates n-grams for $n = 1, 2, 3, 4$ (BLEU-1 to BLEU-4), making it effective for tasks like machine translation and captioning.

2) METEOR combines exact word matching with stemming and synonym recognition. It calculates a weighted harmonic mean of precision and recall, capturing semantic relationships better than n-gram-based metrics.

3) ROUGE_L evaluates similarity using the longest common subsequence (LCS). It emphasizes recall and is suitable for assessing content overlap in summarization and captioning.

4) CIDEr uses TF-IDF weighting to highlight n-grams that are both significant and frequently shared across multiple reference captions, ensuring that the generated captions closely align with the consensus of human annotations.

5) SPICE evaluates semantic content by comparing objects, attributes, and relationships. It uses scene graphs to assess deeper meaning beyond surface-level text similarity.

C. Experimental Settings

1) *Data Setting*: For fair comparison with existing methods, each dataset is split into three subsets: 80% of the image-caption pairs are used for training, 10% for validation, and the remaining 10% for testing. During preprocessing, image sizes are resized to 224×224.

2) *Frozen Pre-trained Model*: We utilize the ViT-g/14 from EVA-CLIP [57] as the visual encoder. For the language model, we adopt OPT-2.7B[1] as the decoder-only LLM, directly loading its pre-trained weights. The QFormer[2] is initialized with fully pre-trained BERT parameters from BLIP-2.

3) *Fine-Tuning Setting*: Fine-tuning is performed using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a weight decay of 0.05. The training process spans 40 epochs, starting with a linear learning rate warm-up phase where the learning rate gradually increases from 1×10^{-6} to 1×10^{-4}. After this warm-up phase, a cosine learning rate decay strategy is applied, with the minimum learning rate set to 0. Mixed precision training is employed to enhance computational efficiency, and the training and validation batch sizes are configured as 64 and 32, respectively. During inference, we apply a beam search algorithm to achieve a balance between computational efficiency and prediction accuracy. All experiments are implemented on an NVIDIA GeForce RTX 4090 GPU.

D. Comparison with Existing Methods

To evaluate the performance of the proposed method, we conduct comparisons with several mainstream approaches, all of which are based on encoder-decoder architectures. These include both traditional methods and pre-trained transformer-based approaches.

1) Traditional methods: We first compare our proposed method with several advanced traditional RSIC methods, including Word Sentence [46], GVFGA+LSGA [47], GLCM [49], MLCA-Net [48], MLAT [50], and HCNet [51]. These methods employ various attention mechanisms and feature integration strategies. A brief overview is as follows. Word Sentence [46] extracts key terms from images and constructs sentences through sequencing tasks. GVFGA+LSGA [47] employs two attention mechanisms to reduce redundant visuals and enhance VL feature integration. GLCM [49] combines global and local features, enhancing generation and interpretability by linking words to relevant visual features. MLCA-Net [48] resolves scale variation and category ambiguity through multi-level and contextual attention. MLAT [50] integrates multi-scale features from convolutional layers and uses LSTM to aggregate information, refining features before passing to a Transformer decoder for sentence generation. HCNet [51] employs a hierarchical feature aggregation module and designs a cross-module feature interaction module to reduce the modality gap between text and image features.

2) Pre-trained transformer-based methods: 2) Pre-trained transformer-based methods: To fully evaluate our performance, we select several RSIC methods based on pre-trained transformers, as described below. ClipCap [52] uses a mapping network to convert CLIP embeddings into prefixes, which are fed into a fine-tuned GPT-2, allowing efficient image captioning without extra pre-training. PureT [53] utilizes Swin-Transformer for grid-level feature extraction and integrates a refining encoder and decoder in a Transformer-based framework. RSGPT [54] fine-tunes InstructBLIP on high-quality human-annotated remote sensing image captioning datasets for enhancing performance. Additionally, we compare with the baseline BLIP-2 and its extension BITA to validate our effective use of the VLP paradigm. BLIP-2 [6] serves as our baseline, introducing a QFormer to bridge the modal gap. After two stages of pre-training, the QFormer generates visual representations interpretable by the LLM. BITA [20] builds upon BLIP-2 by introducing interactive Fourier Transformer to replace the traditional transformer, enabling more effective alignment of remote sensing image-text features.

E. Experimental Results and Analysis

1) ***Quantitative Comparison***: Tables I, II, and III show the quantitative results of different methods on the UCM-caption, NWPU-caption, and RSICD datasets. It is evident that our method outperforms other comparison methods, benefiting from the VLP paradigm, which combines the strong representation and reasoning capabilities of visual pretraining models and LLMs. Furthermore, our proposed CRS-Adapter effectively learns domain-specific knowledge in remote sensing and bridges the multimodal gap between the visual pretraining model and the LLM more efficiently.

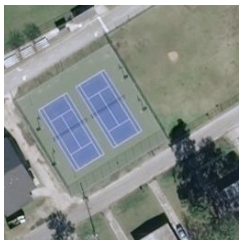
The results from the UCM-caption dataset show that traditional methods (e.g., GVFGA+LSGA, MLCA-Net, and GLCM) consistently underperform across multiple metrics compared to methods utilizing Transformers as both the encoder and decoder (e.g., PureT and BLIP-2). In addition, the HCNet demonstrates significant performance improvements over traditional methods due to the integration of an additional feature alignment module. The non-remote sensing method

[1] Pre-trained weights for OPT 2.7B
[2] Pre-trained weights for BLIP-2 QFormer

UCM

GT: Two straight freeways in the desert with no car on them.
HCNet: Two straight freeways in the desert.
Base: There are two straight freeways in the desert with no cars.
Ours: There are two straight freeways in the desert with no cars on them.

GT: Lots of boats docked at the harbor and the water is deep blue.
HCNet: Many boats docked at the harbor.
Base: Lots of boats docked at the harbor and the water is blue.
Ours: Lots of boats docked **neatly** at the harbor and the water is deep blue.



GT: There are two tennis courts on the lawn and surrounded by some plants
HCNet: Two tennis courts arranged neatly.
Base: There are two tennis courts arranged neatly on the lawn
Ours: There are two tennis courts arranged **neatly** and surrounded by some plants.

GT: A part of a golf course with green turfs and some trees.
HCNet: A golf course with grass and some trees.
Base: This is part of a golf course with turfs and trees..
Ours: This is a part of a golf course with green turfs and **some bunkers** and trees.

RSICD

GT: Five planes are near a large building in an airport
HCNet: Some planes are close to a big building in the airport.
Base: There are five planes near a large building at the airport.
Ours: Five planes are parked near **a large terminal** in an airport.

GT: White waves in green ocean is near yellow beach
HCNet: A green lawn is near a piece of yellow beach.
Base: White waves in a green ocean near a yellow beach
Ours: White waves are **between** green ocean and yellow beach.

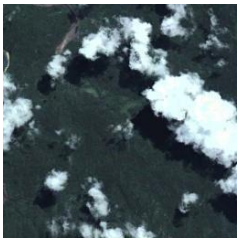
GT: Many buildings are in a commercial area
HCNet: There are many buildings in a commercial area.
Base: Many buildings are some trees located in a business area
Ours: Many **tall** buildings and **some green trees** are in a commercial area.

GT: some cars are parked in a parking lot with several green trees.
HCNet: Some cars are parked in a parking area with a few trees.
Base: Several cars are parked in a parking lot with green trees
Ours: Some cars are parked in a parking lot with **several empty parking spaces** and green trees

NWPU

GT: A slender, Straight bridge connecting the land on both sides of the blue waters.
HCNet: A bridge connects the land on both sides of the blue water.
Base: A slender, straight bridge connects the land across the blue water
Ours: A slender, Straight bridge connects the land on both sides of the turquoise water

GT: Several dark green circular farmlands are neatly arranged on the ground.
HCNet: Several dark green circular farmlands are neatly placed on the ground.
Base: There are some dark green circular farmlands arranged neatly.
Ours: Some dark green circular farmlands of **the same size** are neatly arranged on the ground.



GT: The cumulus clouds are located above the land covered by dense vegetation
HCNet: The clouds are above the land with a lot of vegetation.
Base: The stratus clouds are located above the land covered by a lot of vegetation.
Ours: The stratus clouds are located above the land covered by dense vegetation.

GT: The mobile home park has some dense mobile homes and a swimming pool and some trees are among these mobile homes
HCNet: There are many buildings and trees in the mobile home park
Base: The mobile home park has some dense mobile homes and some trees.
Ours: The mobile home park has some dense mobile homes arranged in rows and some trees and a swimming pool.

Fig. 4. Examples of sentences generated by PE-RSIC from UCM dataset (first row), RSICD dataset (second row), and NWPU dataset (third row). The red words and blue words indicate the disadvantages of other methods and the advantages of our method.

BLIP-2 leverages the powerful representation and reasoning capabilities of a pretrained visual encoder and LLM. By introducing interactions between visual and textual features through QFormer, it achieves significantly better results than most advanced comparison methods across various metrics. This highlights the superiority of the VLP paradigm. Our approach further enhances QFormer by incorporating a trainable adapter that can bridge the domain gap while retaining prior knowledge from the natural domain, leading to more robust remote sensing visual representations. A related approach, BITA, similarly builds on BLIP-2 by proposing the IFT module. BITA outperforms PureT and BLIP-2 on a range of metrics. Compared to BITA, our method achieves better results, with our CIDEr score improving by approximately 10 points over theirs, and demonstrating significant advantages in parameter efficiency. The results on the NWPU and RSICD datasets are similar methods based on Transformers outperform traditional methods. Among the Transformer-based approaches, those adopting the VLP paradigm demonstrate superior performance. Our method also achieves highly competitive results.

2) ***Qualitative Comparison***: Fig. 4 shows some typical remote sensing scenes from three experimental datasets, along with the corresponding captions generated by three different methods: HCNet, BLIP-2 (base), and PE-RSIC. Clearly, compared to HCNet and BLIP-2, the caption generation results of our method are superior in terms of word accuracy, sentence comprehensiveness, and semantic consistency. Specifically, in Fig. 4(B) and 4(C), our model not only provides a complete description of the images but also generates detailed descriptions of the spatial arrangement of objects, enriching sentence detail and enhancing the sense of the scene. Likewise, in Fig. 4(F), our method not only accurately describes the colors of the beach and the waves but also their positional relationship, whereas some methods even produce incorrect descriptions. In Fig. 4(E), while other methods fail to capture the term "terminal" accurately, our approach not only correctly identifies the number and size of the aircraft but also demonstrates strong scene perception capabilities, offering a more comprehensive depiction of the airport environment. Moreover, compared to other methods, our model delivers more detailed object descriptions, such as "tall buildings" and "green trees" in Fig. 4(G) and "farmlands of the same size" in Fig. 4(J). Notably, our method continues to perform well in complex scenes. For example, in Fig. 4(D), our approach identifies"bunkers," which are missed by other methods and even the ground-truth labels. Similarly, in Fig. 4(H), PE-RSIC successfully highlights the feature "empty parking spaces." Finally, in Fig. 4(L), our method is the only one to describe all objects in the scene.

3) ***Trained Parameter Efficiency***: Fig. 6(a) illustrates the relationship between parameter count and CIDEr score during training across different methods. Evidently, our approach achieves the best performance while requiring the minimal trainable parameters compared to other methods. Specifically, our model only requires 9.6M parameters, representing a 95% reduction in trainable parameters compared to full fine-tuning. Notably, CRS-Adapter accounts for 5.9M parameters (with a bottleneck dimension r = 128), and Class Prompt accounts for 3.6M parameters. The proposed PETL framework achieves remarkable efficiency and effectiveness in RSIC tasks, showcasing the viability of adapting VL pre-trained models from the natural domain to address challenges in the RS domain.

F. limitation

Despite the promising results achieved by PE-RSIC across multiple benchmark datasets, the proposed framework has certain limitations in handling remote sensing images containing small objects. Specifically, the model sometimes fails to accurately describe fine-grained objects that occupy only a small portion of the image. As shown in Fig. 5, a vehicle is small and embedded within complex urban backgrounds, which make it difficult for our model to accurately describe them. Our current method utilizes a global [CLS] token extracted from the vision encoder to generate a Class Prompt. It lacks the spatial granularity to focus on fine or localized targets. To address this, future work may consider incorporating finer-grained object cues while avoiding excessive computational overhead to enhance the sensitivity of model to small object.

GT: A tan cross-shaped roof, The main house has a light blue round roof, A church with a tower, And buildings, Roads and cars around.

Ours: A tan cross-shaped roof, the main house with a light blue round roof, a church with a tall tower, and buildings and roads nearby.

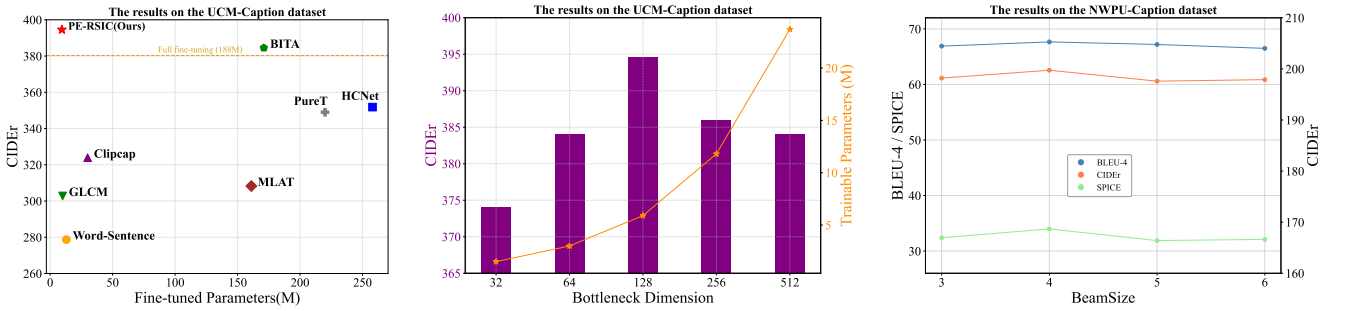
Fig. 5. An example of sentence generated by PE-RSIC compared to GT.

G. Ablation Study

In this section, we design ablation in the UCM-Caption dataset and NWPU-Caption dataset to demonstrate the effectiveness of the design module.

Comparison to Traditional PEFT Methods. As shown in Table IV, our method significantly outperforms traditional PEFT techniques with only a slight increase in parameter count. Specifically, in terms of CIDEr, it +18.36 over Bitfit, +19.34 over Adapter, and +12.26 over Low-Rank Adaptation (LoRA), with notable advantages in other metrics as well. This demonstrates the superiority of our framework over conventional PEFT methods. Unlike PEFT techniques designed for natural domains, our framework is specifically tailored to address the unique challenges of complex visual feature analysis and modality alignment in the remote sensing domain, specifically designed for RSIC.

Bottleneck Dimension and BeamSize. We experiment with different bottleneck dimensions of CRS-Adapter and beam sizes during inference. From Fig. 6(b), we observe that as the bottleneck dimension increases, both the parameter size and model performance improve. However, when the embedding dimension reaches 128, the model performance begins to decline. This may be because larger embedding dimensions lead to overfitting. Therefore, we choose the bottleneck dimension of our adapter to be 128. As shown in Fig. 6(c), the model achieves its highest BLEU-4, CIDEr, and SPICE scores when the beam size is equal to 4, with performance decreasing as the beam size deviates from this value.



(a) (b) (c)

Fig. 6. (a) Comparison of PE-RSIC with different models in terms of CIDEr performance versus the number of trainable parameters. (b) Bottleneck dimension: The impact of different bottleneck dimensions on the model performance CIDEr and trainable Parameters(M). (c) Beam size: The impact of different beam sizes on the model performance BELU-4, SPICE and CIDEr.

TABLE IV
ABALTION RESULTS ON UCM DATASET. THE BOLD INDICATE THE BEST RESULTS.

Method	Params(M)	BLEU-l	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE
Full Fine-Tuning	188.0	88.04	82.23	76.89	71.84	47.32	83.36	380.31	54.78
Adapter	9.6	82.45	78.32	72.89	68.57	45.12	79.86	375.23	53.14
LoRA	7.1	87.78	81.93	76.12	72.56	47.89	82.84	382.31	54.35
BitFit	3.8	87.21	80.85	76.48	71.24	45.04	81.93	376.21	53.24
Ours	9.6	**90.03**	**84.08**	**78.17**	**74.33**	**49.33**	**84.72**	**394.57**	**55.79**

TABLE V
ABALTION RESULTS ON UCM DATASET. THE BOLD INDICATES THE BEST RESULTS.

Method	Baseline	CRS-Adapter	Class Prompt	Full Feature	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE
Model1	✓				88.04	82.23	76.89	71.84	47.32	83.36	380.31	54.78
Model2	✓	✓			89.07	83.62	78.50	73.57	47.91	84.53	389.26	54.84
Model3	✓	✓		✓	89.25	**84.12**	77.89	74.05	48.97	**84.84**	391.68	55.35
Model4(Ours)	✓	✓	✓		**90.03**	84.08	**78.17**	**74.33**	**49.33**	84.72	**394.57**	**55.79**

TABLE VI
ABALTION RESULTS ON UCM DATASET. THE BOLD INDICATES THE BEST RESULTS.

Method	Resdual	Gating	Weight-Sharing	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE
Model5				83.71	77.42	71.38	66.84	42.13	79.29	360.71	50.64
Model6	✓			88.64	82.02	77.53	72.16	46.23	82.31	377.62	53.41
Model7	✓		✓	89.32	83.61	78.14	73.07	47.11	83.54	388.46	54.06
Model4	✓	✓	✓	**90.03**	**84.08**	**78.17**	**74.33**	**49.33**	**84.72**	**394.57**	**55.79**

TABLE VII
ABALTION RESULTS ON UCM DATASET. THE BOLD INDICATES THE BEST RESULTS.

Method	Positions	Params(M)	Times(S)	BLEU-4	CIDEr	SPICE
Model8	Top	6.6	2.07	68.97	320.84	51.56
Model9	Bottom	6.6	2.07	72.43	378.67	53.47
Model10	0-8	8.1	2.10	73.01	385.34	54.61
Model4	All	9.6	2.16	**74.33**	**394.57**	**55.79**

Effectiveness of Main Components. As shown in Table V, to verify the effectiveness of each component, we conduct ablation experiments by progressively incorporating components into the baseline BLIP-2 until the complete PE-RSIC framework is established. Compared to full fine-tuning, introducing the CRS-Adapter component alone already leads to a certain degree of performance improvement. When both components are utilized together, the CIDEr score increases by 14.26, and the BLEU-4 score increases by 2.49. We compared the performance when projecting the [CLS] token versus the full feature into the LLM. Model 3 represents the projection of the full feature from the penultimate layer to the LLM with a fully connected layer to match the dimensionality of LLMs. Model 4 only projects the [CLS] token into the LLM, which significantly reduces the computational cost of the LLM, achieving better results. These results highlight the effectiveness of the individual components designed within the PE-RSIC framework.

Components of CRS-Adapter. As shown in table VI, we conduct an ablation study by incrementally adding components of the CRS-Adapter. Model 5 represents adapter residual learning branch after removing the MSA module. Model 6 contains two full branches without gating mechanisms and weight shar-

ing. However, adding a residual learning branch still improves performance. Model 7 introduces weight sharing only in the downsampling layer. Finally, Model 4 represents the complete CRS-Adapter architecture. Our experiments demonstrate the effectiveness of each component in CRS-Adapter.

Positions of CRS-Adapter and Speed Performance. As shown in Table VII, we evaluate the impact of inserting CRS-Adapters at different QFormer layers on model performance and inference efficiency. Adapters are inserted into: top 6 layers (Model8), bottom 6 layers (Model9), first 9 layers (Model10), and all 12 layers (Model4). Performance improves with more inserted layers. Model4 achieves the best result, indicating that deeper integration helps better model the semantic alignment between remote sensing images and text. In addition, Model9 clearly outperforms Model8 despite using the same number of adapters, suggesting that inserting adapters into deeper layers is more effective than in earlier ones. Regarding inference time, deeper insertion introduces only a minor overhead (from 2.07s to 2.16s).

V. Conclusion

In this paper, we propose PE-RSIC, a Parameter-Efficient Transfer Learning (PETL) framework tailored for RSIC. Our framework leverages the pre-trained BLIP-2 model as the baseline and introduces two key components: the CRS-Adapter and the Class Prompt. The CRS-Adapter is specifically designed to adapt pre-trained natural-domain vision-language models to the remote sensing domain by learning cross-modal relationships in a lightweight manner. It is integrated in parallel within the QFormer blocks, utilizing a dual-branch structure to retain query integrity and a gating mechanism to regulate the flow of information. Meanwhile, the Class Prompt, encoded from the [CLS] token of the vision encoder output, effectively captures geospatial class-specific features and guides the decoder in generating more accurate and context-aware captions for remote sensing images. By freezing the parameters of the pre-trained components and fine-tuning only the adapter and prompt modules, PE-RSIC achieves efficient domain adaptation with minimal computational and memory overhead, outperforming full fine-tuning methods with only 5% of the trainable parameters. Moreover, numerous advanced VLPs have emerged to better handle diverse tasks. Future work will focus on effectively integrating CRS-Adapter with other QFormer-based VLPs and further optimizing its memory footprint during both training and inference. We hope our work will inspire research toward a unified PETL framework for multimodal foundation models, addressing the diverse task and scenario requirements in remote sensing.

References

[1] Q. Li, W. Zhang, W. Lu, and Q. Wang, "Multibranch mutual-guiding learning for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–10, 2025.

[2] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Rgb-induced feature modulation network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[3] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "Spectralgpt: Spectral remote sensing foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5227–5244, 2024.

[4] Q. Li, M. Zhang, Z. Yang, Y. Yuan, and Q. Wang, "Edge-guided perceptual network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–10, 2024.

[5] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to text: Survey of image and video captioning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 4, pp. 297–312, 2019.

[6] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[7] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[8] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[9] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[10] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5227–5237.

[11] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International conference on computer, information and telecommunication systems (Cits)*. IEEE, 2016, pp. 1–5.

[12] Y. Li, X. Zhang, J. Gu, C. Li, X. Wang, X. Tang, and L. Jiao, "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[13] H. Kandala, S. Saha, B. Banerjee, and X. X. Zhu, "Exploring transformer and multilabel classification for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, p. 1–5, Jan 2022.

[14] X. Ma, R. Zhao, and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 2001–2005, 2020.

[15] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 3, pp. 436–440, 2020.

[16] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, vol. 8, pp. 2608–2620, 2019.

[17] Z. Zhang, W. Diao, W. Zhang, M. Yan, X. Gao, and X. Sun, "Lam: Remote sensing image captioning with label-attention mechanism," *Remote Sensing*, vol. 11, no. 20, p. 2349, 2019.

[18] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, "A multi-level attention model for remote sensing image captions," *Remote Sensing*, vol. 12, no. 6, p. 939, 2020.

[19] Q. Wang, Z. Yang, W. Ni, J. Wu, and Q. Li, "Semantic-spatial collaborative perception network for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[20] C. Yang, Z. Li, and L. Zhang, "Bootstrapping interactive image-text alignment for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[21] H. Lin, D. Hong, S. Ge, C. Luo, K. Jiang, H. Jin, and C. Wen, "Rs-moe: A vision-language model with mixture of experts for remote sensing image captioning and visual question answering," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2025.

[22] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu, "Vlp: A survey on vision-language pre-training," *Machine Intelligence Research*, vol. 20, no. 1, pp. 38–56, 2023.

[23] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," *arXiv preprint arXiv:2202.10936*, 2022.

[24] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[25] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[26] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.
[27] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proceedings of ECCV*. Springer, 2020, pp. 121–137.
[28] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.
[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
[30] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[32] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.
[33] W. Dai, J. Li, D. Li, A. Meng Huat Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," *arXiv preprint arXiv:2305.06500*, 2023.
[34] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.
[35] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," *AI Open*, 2023.
[36] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
[37] ——, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.
[38] J. Wang, C. Wang, F. Luo, C. Tan, M. Qiu, F. Yang, Q. Shi, S. Huang, and M. Gao, "Towards unified prompt tuning for few-shot text classification," *arXiv preprint arXiv:2205.05313*, 2022.
[39] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 664–16 678, 2022.
[40] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.
[41] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv preprint arXiv:2111.03930*, 2021.
[42] H. Lu, Y. Huo, G. Yang, Z. Lu, W. Zhan, M. Tomizuka, and M. Ding, "Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling," *arXiv preprint arXiv:2302.06605*, 2023.
[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
[44] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," *arXiv preprint arXiv:2106.10199*, 2021.
[45] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.
[46] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 532–10 543, 2020.
[47] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.
[48] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "NWPU-captions dataset and mlca-net for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.
[49] Q. Wang, W. Huang, X. Zhang, and X. Li, "Glcm: Global–local captioning model for remote sensing image captioning," *IEEE Transactions on Cybernetics*, vol. 53, no. 11, pp. 6910–6922, 2022.
[50] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
[51] Z. Yang, Q. Li, Y. Yuan, and Q. Wang, "Hcnet: Hierarchical feature aggregation and cross-modal feature alignment for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
[52] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.
[53] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2585–2594.
[54] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *arXiv preprint arXiv:2307.15266*, 2023.
[55] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *International conference on computer, information and telecommunication systems (Cits)*. IEEE, 2016, pp. 1–5.
[56] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.
[57] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 358–19 369.

Xuezhi Zhao is currently working toward the M.S. degree in computer technology with the School of Computer Science, the School of Artifcial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.

Zhigang Yang is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing and computer vision.

Qiang Li (Member, IEEE) is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/)