PERS: Parameter-Efficient Multimodal Transfer Learning for Remote Sensing Visual Question Answering

Jinlong He, *Student Member, IEEE*, Gang Liu, *Member, IEEE*, Pengfei Li, *Student Member, IEEE*,
Xiaonan Su, Wenhua Jiang, Dongze Zhang, and Shenjun Zhong

***Abstract*—Remote sensing (RS) visual question answering (VQA) provides accurate answers through the analysis of RS images (RSIs) and associated questions. Recent research has increasingly adopted transformers for feature extraction. However, this trend leads to escalating training costs as a consequence of increased model sizes. Furthermore, existing studies predominantly employ transformers to extract features from a single modality, insufficiently integrating multimodal information and thereby undermining the potential advantages of transformers in feature extraction and fusion in these scenarios. To address these challenges, we propose parameter-efficient multimodal transfer learning for RSVQA. We introduce a lightweight, parameter-efficient adapter into the visual feature extraction module, initialized with weights pretrained on large-scale RSIs to reduce both training costs and parameters. A cross-attention mechanism is employed for multimodal interaction, enhancing the integration of information across modalities. Comprehensive experiments were conducted on three datasets: RSVQA-LR, RSVQA-HR, and RSVQAxBEN, achieving state-of-the-art performance. Moreover, exhaustive ablation studies demonstrate that our parameter-efficient adapter strategy achieves performance comparable to full-parameter training under partial parameter conditions, validating the efficacy of our approach.**

***Index Terms*—Multimodal representation learning, parameter-efficient transfer learning, remote sensing (RS) visual question answering (VQA).**

I. Introduction

REMOTE sensing (RS) visual question answering (VQA) entails obtaining information pertinent to posed questions from RS images (RSIs) and providing precise responses. The utilization of RS technology spans diverse domains including agriculture, forestry, natural resource management, and disaster mitigation. Research in multimodal RS increasingly relies on extensive RSI-text data. With advancements in RS technology, the proliferation of high-definition and detailed RSIs datasets has significantly enhanced the processing and analysis capabilities for RSIs. Applications in RSIs analysis encompass hyperspectral image classification [1], [2], scene classification [3], [4], [5], target detection [6], [7], [8], change detection [9], [10], and semantic segmentation [11], [12], [13]. Despite these advancements, a scarcity of high-quality RS multimodal image-caption datasets persists when compared to single-modality image datasets. As a precursor to RSVQA, Lobry et al. [14] introduced a novel RSI-text dataset along with the RSVQA model, which employs convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to extract visual and textual features, respectively. Subsequently, these features are fused through pointwise multiplication.

Subsequent research has continued to advance RSVQA. For instance, Yuan et al. [15] enabled the model to incrementally learn the problem by adjusting the prediction strategy. Building on this, Yuan et al. [16] introduced cross-modal global attention (CGA) and a spatial transformer to enhance visual information extraction. Subsequently, the gated recurrent unit was employed for semantic understanding and feature fusion [17]. Chappuis et al. [18] proposed Prompt-RSVQA, utilizing CNNs to extract image features, converts them into text, and integrates prompts into the language transformer, DistilBERT. Notably, Zhang et al. [19] designed a question-guided relation network (QRN) unit and introduced the SHRNet model, incorporating an attention mechanism to address the large-scale differences and location sensitivity of RSIs.

However, the aforementioned studies predominantly focus on enhancing single-modal feature extraction capabilities, adjusting learning strategies, or employing pointwise multiplication for fusion. This approach overlooks the significance of multimodal fusion and inadequately addresses cross-modal knowledge integration. Research by Chappuis et al. [20] demonstrated that pointwise multiplication, the simplest fusion method, yields the lowest performance compared to other existing fusion methods [14], [21], [22], thereby underscoring the substantial impact of cross-modal fusion on model effectiveness. Furthermore, although [17], [18], [19] adopt transformer technology to

Received 11 May 2024; revised 19 June 2024; accepted 14 August 2024. Date of publication 21 August 2024; date of current version 5 September 2024. This work was supported by the Ministry of Education Humanities and Social Science Research Planning Fund Project under Grant 23YJAZH079. *(Jinlong He and Pengfei Li are co-first authors.) (Corresponding author: Gang Liu.)*

Jinlong He, Gang Liu, Pengfei Li, Xiaonan Su, Wenhua Jiang, and Dongze Zhang are with the College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China (e-mail: hejinlong@hrbeu.edu.cn; liugang@hrbeu.edu.cn; pengfeili@hrbeu.edu.cn; suxiaonan@hrbeu.edu.cn; wenhuaj@hrbeu.edu.cn; zhangdongze@hrbeu.edu.cn).

Shenjun Zhong is with the Monash Biomedical Imaging, Monash University, Melbourne, VIC 3800, Australia, and also with the National Imaging Facility, University of Melbourne, Parkville, VIC 3052, Australia (e-mail: shenjun.zhong@monash.edu).

Digital Object Identifier 10.1109/JSTARS.2024.3447086

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

enhance feature extraction or fusion, they continue to rely on traditional CNNs or RNNs for modal feature extraction. This reliance hampers the full exploitation of transformer capabilities in feature extraction and escalates both model storage and training costs.

In multimodal RS research, handling diverse input types such as images and text necessitates models with substantial parameters to capture and learn the complex relationships and interactions among the data, thus requiring the introduction of additional vision encoding and text processing modules. Furthermore, to enhance the integration of information across modalities, additional fusion layers are introduced, which in turn increases the parameter count. Concurrently, with the prevalent use of transformer architectures, the multimodal transformer is designed to be deeper and wider to improve performance, consequently escalating both the parameter count and training costs. Under these circumstances, balancing the reduction of parameter updates, minimizing training costs, and maintaining high performance presents a significant challenge in current research on RS multimodality.

In response to the aforementioned challenges, we propose **P**arameter-**E**fficient multimodal transfer learning for **R**emote **S**ensing visual question answering (PERS). Within this framework, we develop a multimodal fusion module using cross-attention mechanisms to facilitate effective cross-modal representation learning. In addition, to reduce both the training costs and the number of parameters requiring updates, we introduce a parameter-efficient transfer learning technique utilizing lightweight adapters. Furthermore, we initialize the visual transformer (ViT) with weights pretrained on large-scale RSIs [23]. This adapter enables comparable performance to full-parameter fine-tuning by adjusting only a minimal subset of parameters. During training, we freeze all other parameters of the vision encoder except for those in the adapter. Experiments demonstrate that PERS achieves state-of-the-art (SOTA) results on existing three RSVQA datasets: RSVQA-LR, RSVQA-HR, and RSVQAxBEN. Simultaneously, we validate that the strategy of employing an adapter and freezing other vision encoder parameters can achieve performance comparable to training the entire model. Furthermore, we conduct ablation studies to illustrate the effectiveness of our technique.

In summary, this article makes the following contributions.

1) We introduce the PERS, incorporating a lightweight adapter in each layer of the vision encoder, initialized with weights pretrained on large-scale RSIs. This approach significantly reduces the number of parameters requiring updates and lowers overall training costs.
2) We develop a novel multimodal fusion module that effectively integrates self-attention and cross-attention mechanisms, facilitating a more comprehensive and effective learning process from both visual and textual inputs processed by their respective encoders.
3) We introduce and evaluate a new model for RSVQA, conducting rigorous experiments to assess its performance. Our model achieves SOTA results on the established benchmark RSVQA dataset, demonstrating its superiority and effectiveness.

II. Related Work

A. Visual Question Answering

In recent years, significant advancements have been made in research on single-modal tasks, including natural language processing (NLP) and image classification. However, effectively extracting and utilizing features from different modalities in multimodal tasks remain areas that require further investigation. VQA represents a precise intersection of visual and natural language modalities.

Recently, the availability of abundant image-caption datasets has significantly advanced research in VQA. Initially, Antol et al. [24] employed a pretrained VGGNet for image processing and LSTM for question processing. Subsequently, the field evolved to incorporate various modules within the VQA model for specialized processing tasks. For instance, Wang et al. [25] considered the information in the image, question, and answer, proposing a framework that utilizes trilinear attention and self-attention mechanisms, structured in a two-stage workflow. This approach, dubbed the MIRTT framework, achieved notable results. Farinhas et al. [23] argued that the flexibility of attention mechanisms could lead to distraction and lack of focus, prompting them to model the attention mechanism as a multimodal feature function in VQA to emulate human attention patterns. Furthermore, to address the issue of complex or discontinuous focus areas in images, Martins et al. [26] introduced a continuous unimodal attention mechanism.

However, most existing VQA methods utilize pipeline approaches for knowledge matching and extraction, which frequently result in cascading errors, poor performance, and biased answers. To mitigate these issues, Chen et al. [27] employed an external knowledge graph for zero-shot modeling on a novel dataset. In general, an effective VQA model must proficiently manage image inputs and questions, incorporating capabilities, such as fine-grained image recognition, spatial perception, and knowledge-based reasoning.

B. RS Visual Question Answering

In contrast to the abundance of image-caption datasets in the general domain, the RS domain features relatively few such datasets. This disparity has catalyzed rapid development in the general domain VQA, while progress in RSVQA has been comparatively slower.

Lobry et al. [14] introduced RS datasets RSVQA-LR and RSVQA-HR, which contain high-resolution and low-resolution RSI-caption pairs, respectively. The RSVQA model utilizes CNNs to extract image features and RNNs to extract textual semantic features. It performs elementwise multiplication of vector elements, each treated with the hyperbolic tangent function, for feature fusion. Results are obtained using an MLP, treating VQA as a classification task. This approach has yielded favorable results.

However, Chappuis et al. [20] confirmed in their study that the elementwise fusion strategy in RSVQA was the least effective among the three existing feature fusion methods [14], [21], [22]. In subsequent research, Yuan et al. [15] sought to derive superior semantic information through alternative methods, introducing SPCL, which employs both hard and soft weighting strategies to facilitate learning that progresses from simple to complex problems. Concurrently, they developed a CGA module [16] to comprehend overall image features as guided by language. In addition, a cross-modal spatial transform module was developed to capture regional features pertinent to the query, thereby enhancing the fusion of image and text features and yielding positive outcomes. Zheng et al. [17] observed that existing RSVQA methods employed a nonspatial fusion strategy, which neglected the spatial information of images and the word-level details of questions. Consequently, they proposed a mutual attention initiation network for RSVQA, which utilizes an attention mechanism and bilinear techniques for effective fusion. This network incorporates word vectors into the CNN, enhancing the semantic interpretation of images.

Following the emergence and ongoing development of transformers, Bazi et al. [28] recognized their advantages and implemented them in both the visual and textual components of RSVQA. Concurrently, they utilized the attention mechanism to extract relevant modal information, achieving positive outcomes. Simultaneously, Chappuis et al. [18] noted that traditional visual language models, which obtain embeddings by fusing features from two deep models and process image and text separately, still faced challenges with efficient image information extraction. To address this, they introduced Prompt-RSVQA, which employs CNNs to extract image features. These image features are subsequently converted to text and integrated into the language transformer DistilBERT using prompts, yielding favorable results.

In recent work, Zhang et al. [19] observed that current methods infrequently account for geospatial objects with significant scale variations and location sensitivity, and seldom investigate the relationships between entities. They introduced SHRNet and incorporated a QRN unit to model and reason about high-order internal group-object relations within the hierarchy. Furthermore, they developed a spatial multiscale visual representation module based on hashing, coupled with a VQ interaction module, to derive more effective joint image-text embeddings for answer prediction, achieving promising results.

Although most current RSVQA research achieves commendable performance, it predominantly focuses on visual feature extraction and frequently overlooks the cross-modal interaction between visual and textual features. This oversight has prompted us to design a multimodal fusion module that effectively integrates information across different modalities. Unlike earlier, simpler multimodal fusion approaches, such as the dot product, our module employs self-attention and cross-attention mechanisms to comprehensively integrate information from both image and text modalities.

C. Parameter-Efficient Transfer Learning

Parameter-efficient transfer learning facilitates the transfer of parameters from a pretrained model to a new model, simplifying its training and reducing costs by freezing some of these parameters. Initially proposed and applied in the field of NLP, this technique has yielded notable results. Subsequently, researchers adapted this technology for the computer vision (CV) field, where it has also demonstrated strong performance [29], [30].

Currently, three mainstream methods exist for parameter-efficient transfer learning. The first method involves integrating adapters into the model, initially proposed for the CV field [31] and subsequently adapted for the NLP field [32]. Adapters are inserted between the model's layers, during training, only these adapters are updated while the remaining modules are frozen. This approach significantly reduces the number of training parameters required for the model. The second strategy, prompt tuning [33], [34], [35], adds training parameters to the model without altering its architecture, similar to the adapter strategy. Prompt tuning introduces a set of learnable tokens at either the input layer or the middle layers of the model. In models utilizing prompt tuning, the majority of components are frozen, updating only the added tokens or downstream classifier modules [33]. The final strategy involves updating only the low-rank matrix that approximates the weights during model training [36].

To date, in RSVQA research, the predominance of transformer-based methods in vision-text modeling has resulted in larger parameter sizes and increased training costs. To address this, we employ parameter-efficient transfer learning for RSVQA, initializing vision encoders with a ViT pretrained on extensive RSIs datasets and adopting a lightweight adapter strategy. This approach significantly reduces the number of parameters requiring updates, lowers training costs, and maintains robust model performance.

III. Method

The PERS framework, depicted in Fig. 1, employs a vision encoder and a language encoder to encode the visual and language modalities, respectively, and incorporates a parameter-efficient component within the vision encoder. Subsequently, after separately extracting information from each modality, the knowledge from both is thoroughly fused in a multimodal fusion module. Ultimately, this fully fused knowledge is fed into a classifier for the purpose of classification.

A. Problem Definition

We approach the VQA task as a multiclass classification challenge, utilizing a training set $D = \{I_i, Q_i^j, A_i^j\}_{i=1\ldots N}^{j=1\ldots M}$ that comprises N images, each image I_i associated with M question–answer pairs (q_i, a_i). For the VQA model, our objective is to correctly answer a question q associated with an image I, trained on the dataset, wherein the answer a corresponds to the label of the pair (q_i, a_i).

B. Vision and Language Encoders

1) Vision Encoder: In PERS, we utilize VIT-B with N_r layers as the primary architecture of the vision encoder, as depicted in Fig. 1. Concurrently, we initialize the encoder with weights from a ViT that has been pretrained on large-scale RSIs

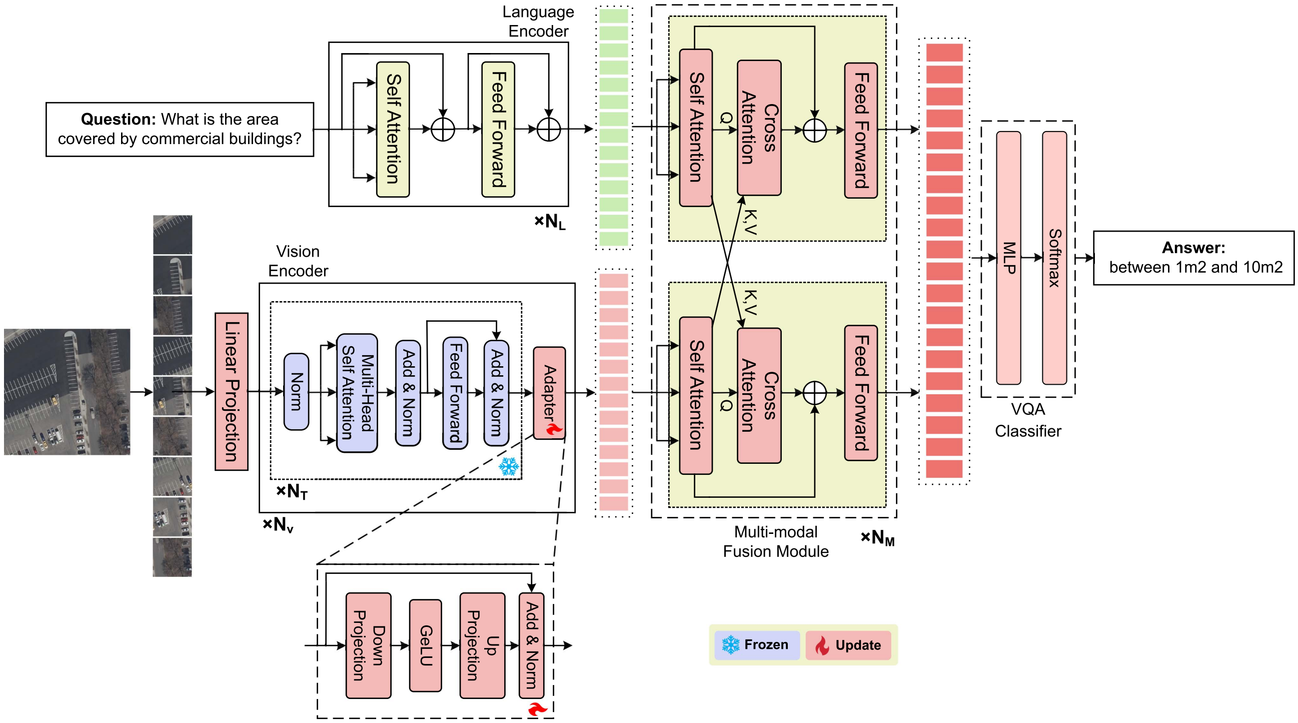


Fig. 1. Architecture of PERS.

datasets. This initialization provides the vision encoder with extensive RS visual knowledge from the outset. The parameter-efficient adapter module for vision encoder detailed in Section C. Parameter-efficient transfer learning.

For image I inputs to the vision encoder, we initially segment the image into patches I_p before inputting them into the transformer, where $v \in \mathbb{R}^{H\times W\times C}$, $v_p \in \mathbb{R}^{N\times(P^2\cdot C)}$. Here, $(H \times W)$ denotes the image resolution, C represents the number of channels, and (P, P) specifies the resolution of each patch. The total number of patches, N, is calculated as $N = HW/P^2$. These patches are subsequently flattened and mapped to D dimensions using a linear projection, as defined by the formula

$$z^I = \left[I_L; I_1E^I; \ldots; I_NE^I\right] + E^I_{\text{pos}} \tag{1}$$

where $E^I \in \mathbb{R}^{(P^2\cdot C)\times D}$, $I_L \in \mathbb{R}^D$ represents the learnable embedding, and E^I_{pos} denotes the positional embedding, added to the patch embeddings to preserve spatial information.

2) Language Encoder: As depicted in Fig. 1, we utilize N_L transformer layers to capture linguistic knowledge. For input processing, PERS employs the WordPiece tokenizer [37], similar to BERT [38], to segment the input text into subword tokens $\{\omega_1, \omega_2, \ldots, \omega_M\}$, where $\omega_M \in \mathbb{R}^V$ and V denotes the size of the vocabulary. These subword tokens are then embedded through linear projection, as described in the following:

$$z^l = \left[\omega_T; \omega_1E^l; \ldots; \omega_ME^l; \omega_{\text{SEP}}\right] + E^l_{\text{pos}} \tag{2}$$

where $E^l \in \mathbb{R}^{V\times D}$. Simultaneously, the tokens $\omega_T \in \mathbb{R}^D$ and $\omega_{\text{SEP}} \in \mathbb{R}^D$ are added to the text sequence. ω_T represents the initial sequence token, and ω_{SEP} denotes a special boundary token. Analogous to visual projection, it is necessary to append an embedding $E^l_{\text{pos}} \in \mathbb{R}^{(M+2)\times D}$ at the end of the text sequence to indicate positional information.

C. Parameter-Efficient Transfer Learning

As shown in Fig. 1, PERS incorporates a lightweight, parameter-efficient adapter module within the vision encoder. The vision encoder comprises N_r transformer layers, with each layer including a self-attention layer followed by a feedforward layer. This configuration precedes the inclusion of the parameter-efficient adapter module. To limit the parameter count, the adapter receives an embedding, z_d, of dimension d from the preceding layer, processes it through a down feedforward layer $W_{\text{down}} \in \mathbb{R}^{(d\times m)}$, and outputs an embedding, z_m, of size m. Subsequently, the embedding passes through a GeLU activation function without a change in dimension and is then reprojected to the embedding $z_{d'}$ of dimension d via an up projection layer $W_{\text{up}} \in \mathbb{R}^{(m\times d)}$.

In the adapter, accounting for biases, we add $2md + d + m$ parameters per layer. Thus, we can control the number of parameters by judiciously setting the sizes of m and d, ensuring $m \ll d$. After dimension reduction, the dimension size m ensures a balance between performance and parameter count. In addition, to prevent the parameter initialization in the projection layer from being too small or nearing zero, thus averting large-scale knowledge forgetting, a skip-connection is introduced at the outset of the adapter, rendering the module effectively an approximate identity function. The computation within the adapter is formalized as follows:

$$\text{Adapter}(x) = x + s \cdot \text{GeLU}\left(xW_{\text{down}}\right)W_{\text{up}} \tag{3}$$

where $x \in \mathbb{R}^d$ represents the input embeddings, and s denotes the scaling factor.

Within the vision encoder, we freeze all modules except for the adapter, which remains the only component updated throughout the training process. The self-attention layers and the feedforward layers within the transformer span the successive N_r layers and, when combined with an adapter, constitute the basic components of the vision encoder. As depicted in Fig. 1, we concatenate the N_v layer with this component, which not only preserves the initialized knowledge but also enables the acquisition of new visual modality knowledge through updates to the parameters of the lightweight adapter module.

D. Multimodal Fusion Module

After obtaining the embeddings of the image and language modalities, we utilize both self-attention and cross-attention mechanisms within the multimodal fusion module to effectively fuse the contextualized representations of these modalities. Within this module, as depicted in Fig. 1, we employ N_M transformer layers. Each transformer layer receives an embedding from the previous layer, which it then feeds into the self-attention layer. This process enables the model to handle long-distance dependencies more effectively and capture global information enriched with new knowledge. Following the self-attention layer, the information from multiple modalities is integrated through the cross-attention mechanism, defined as follows:

$$z^{Ic} = \text{ATTN}\left(z^{Is}, z^{ls}, z^{ls}\right)$$
$$z^{lc} = \text{ATTN}\left(z^{ls}, z^{Is}, z^{Is}\right) \tag{4}$$

where z^{Is} and z^{ls} denote the visual and linguistic representations, respectively. The cross-attention layer establishes a correlation between the inputs and outputs of the two modalities, enabling the model to assimilate multimodal knowledge and enhance its performance. Once the inputs for the cross-attention layer are obtained, z^{Ic} and z^{lc} are passed to the feedforward layer, which consists of an MLP fully connected layer. Finally, after multimodal feature fusion, the integrated information from both modalities is conveyed to the classifier to perform the classification task.

IV. Experiments and Results

We conducted comprehensive experiments on three benchmark datasets: RSVQA-LR, RSVQA-HR [14], and RSVQAxBEN [39], evaluating accuracy across various question types as well as average and overall accuracy. Simultaneously, we performed detailed ablation studies on various modules within the model to assess the effectiveness of our proposed method. Furthermore, we conducted a visual analysis to enhance the model's interpretability, displaying attention maps of the model applied to RSIs under various scenarios.

A. Datasets

1) RSVQA-LR: The RSVQA-LR dataset comprises 77 232 question–answer pairs, divided into training, validation, and test sets at proportions of 77.8%, 11.1%, and 11.1%, respectively. Images in the dataset derive from Sentinel-2 imagery acquired in the Netherlands, focusing on nine tiles with low cloud coverage. These image data are publicly accessible through the ESA's Copernicus Open Access Center. Each block is segmented into 772 images, each image featuring a resolution of 256 × 256 pixels and captured in the RGB color band. Collectively, these images encompass a total area of 6.55 square kilometers. Given that the Sentinel-2 satellite captures images at a resolution of 10 m, smaller objects such as houses, roads, and trees are not discernible, indicating that the images span extensive spatial and temporal scales. The question types for the datasets include "count," "comparison," "presence," and "rural/urban" classification.

2) RSVQA-HR: The RSVQA-HR dataset contains 1 066 316 question–answer pairs and was divided into a training set, a validation set, Test Set 1, and Test Set 2, in proportions of 61.5%, 11.2%, 20.5%, and 6.8%, respectively. Test Set 1 covers an area similar to the training and validation sets, while Test Set 2 focuses on the city of Philadelphia. The images in the dataset derive from the USGS High Resolution Orthophoto dataset, which comprises 15 cm aerial RGB images covering most cities in the United States, and are publicly accessible via the USGS EarthExplorer tool. The dataset includes 161 map blocks from the Northeast Coast of the United States, segmented into 10 659 images, each with a resolution of 512 × 512 pixels and covering 5898 square meters. Given the high resolution of USGS imagery, RSVQA-HR contains more useful information. The question types include "count," "comparison," "presence," and "area," where questions about object areas are only feasible in such high-resolution datasets.

3) RSVQAxBEN: The RSVQAxBEN dataset comprises 590 326 Sentinel-2 L2A image blocks from the BigEarthNet (BEN) archive, each equipped with 12 spectral bands. Questions in this dataset are categorized into two types: "yes/no" and "land cover." Each image block generates 25 unique questions, yielding a total of 14 758 150 question–answer pairs with 26 875 unique answers. To streamline the dataset, we limited the answers to the 1000 most frequent ones, which constitute 98.1% of the total answer set. This constraint ensures alignment with previous benchmarks [39]. We allocated 66%, 11%, and 23% of the image samples and corresponding question–answer pairs to the training, validation, and testing sets, respectively.

B. Experimental Details

Our model was implemented in Python 3.8 and PyTorch 2.0, and trained on two NVIDIA RTX 4090 GPUs, each with 24 GB of RAM. The vision encoder was initialized with pretrained weights from the ViT-B model [37], which had been trained across a diverse array of RSIs datasets. Both N_T and N_v were set to 3. For data augmentation, we employed RandAugment [40], randomly cropping the input images to a resolution of 256 × 256 pixels.

For the language encoder, we initialized with pretrained BERT weights [41], setting N_L to 12. The model was trained for 20 epochs with a batch size of 32, using the AdamW optimizer [42] at an initial learning rate of $2e^{-5}$. A cosine schedule was applied

TABLE I
COMPARISON OF THE PERS WITH EXISTING METHODS ON THE RSVQA-LR DATASET

Method	Count	Presence	Comparison	Rural/urban	Average accuracy	Overall accuracy
RSVQA [14]	67.01%	87.46%	81.50%	90.00%	81.49%	79.08%
EasyToHard [16]	69.22%	90.66%	87.49%	91.67%	84.76%	83.09%
Bimodal [28]	72.22%	**91.06%**	91.16%	92.66%	86.78%	85.56%
SHRNet [19]	73.87%	91.03%	90.48%	94.00%	87.34%	85.85%
MADNet [43]	72.85%	90.96%	**91.68%**	95.00%	87.62%	85.97%
PERS(ours)	**75.79%**	90.86%	91.49%	**98.33%**	**89.12%**	**86.89%**

Best performances in different categories are indicated in bold.

TABLE II
COMPARISON OF THE PERS WITH EXISTING METHODS ON THE RSVQA-HR TEST SET 1

Method	Count	Presence	Comparison	Area	Average accuracy	Overall accuracy
RSVQA [14]	68.63%	90.43%	88.19%	85.24%	83.12%	83.23%
EasyToHard [16]	69.06%	91.39%	89.75%	85.92%	83.97%	84.16%
Bimodal [28]	69.80%	92.03%	91.83%	86.27%	84.98%	85.30%
SHRNet [19]	**70.04%**	**92.45%**	91.68%	86.35%	85.13%	85.39%
MADNet [43]	70.02%	92.36%	**91.87%**	86.58%	85.21%	85.51%
PERS(ours)	69.79%	92.09%	91.86%	**91.50%**	**86.31%**	**86.15%**

Best performances in different categories are indicated in bold.

TABLE III
COMPARISON OF THE PERS WITH EXISTING METHODS ON THE RSVQA-HR TEST SET 2

Method	Count	Presence	Comparison	Area	Average accuracy	Overall accuracy
RSVQA [14]	61.47%	86.26%	85.94%	76.33%	77.50%	78.23%
EasyToHard [16]	61.95%	87.97%	87.68%	78.62%	79.06%	79.29%
Bimodal [28]	63.06%	89.37%	89.62%	80.12%	80.54%	81.23%
SHRNet [19]	**63.42%**	**89.81%**	89.44%	80.37%	80.76%	81.37%
MADNet [43]	63.38%	89.69%	89.82%	80.58%	80.87%	81.51%
PERS(ours)	62.23%	88.91%	**90.04%**	**86.09%**	**81.82%**	**81.85%**

Best performances in different categories are indicated in bold.

TABLE IV
COMPARISON OF THE PERS WITH EXISTING METHODS ON THE RSVQAXBEN DATASET

Method	Yes/no	Land cover	Average accuracy	Overall accuracy
RSVQA [14]	79.92%	20.57%	50.25%	69.83%
Prompt-RSVQA [18]	86.07%	26.56%	56.32%	75.40%
VBFusion [44]	86.56%	26.26%	55.80%	76.10%
DBBT [45]	87.83%	36.26%	62.04%	79.15%
LiT-4-RSVQA [46]	89.72%	39.50%	64.61%	81.27%
PERS(Ours)	**92.49%**	**42.91%**	**67.70%**	**83.75%**

Best performances in different categories are indicated in bold.

TABLE V
PERFORMANCE OF PERS IN RSVQA-LR AND RSVQA-HR DATASETS AFTER TRAINING ON DIFFERENT DATASET SIZES

Dataset	Accuracy	Training set size				
		10%	20%	30%	40%	100%
RSVQA-LR	Average	85.28%	86.42%	87.09%	88.12%	**89.12%**
	Overall	83.63%	84.63%	85.49%	86.04%	**86.89%**
RSVQA-HR Test Set 1	Average	85.70%	85.83%	85.93%	86.02%	**86.31%**
	Overall	85.51%	85.64%	85.76%	85.98%	**86.15%**
RSVQA-HR Test Set 2	Average	79.93%	81.02%	81.58%	81.64%	**81.82%**
	Overall	79.88%	80.94%	81.31%	81.57%	**81.85%**

Best performances under different training data sizes are indicated in bold.

to gradually reduce the learning rate to $1e^{-8}$, with a decay weight of 0.05 also applied.

C. Comparison Results and Analysis

In this section, we compare the performance of PERS with several top-performing models on the RSVQA-LR, RSVQA-HR, and RSVQAxBEN datasets. First, as shown in Table I, we evaluate several models across four question types within the RSVQA-LR dataset: "count," "comparison," "presence," and "rural/urban" region classifications. We also assess the overall and average accuracies of these models. PERS achieves SOTA results in the RSVQA-LR dataset for "count" and "rural/urban" questions. The performance on "presence" and "comparison" questions is lower by only 0.2% and 0.19% compared to bimodal [28] and MADNet [43], respectively. For "rural/urban" questions, our model surpasses MADNet [43] by 3.33%, and for "count" questions, it exceeds SHRNet [19] by 1.92%. In addition, the average and overall accuracies of our model are 1.5% and 0.92% higher, respectively, than those of MADNet.

Tables II and III demonstrate the SOTA performance of PERS across two test sets in the RSVQA-HR dataset, with comparably strong performance observed in both the RSVQA-LR and RSVQA-HR datasets. We evaluate the performance across four question types in this dataset: "count," "comparison," "presence," and "area," assessing both the average and overall accuracy of various models. As indicated in Tables II and III, although PERS did not surpass the SHRNet model on "Count" and "Presence" questions, the differences were marginal, with gaps of only 0.25% and 0.36% in Test Set 1, and 1.19% and 0.9% in Test Set 2, respectively. Notably, PERS achieves SOTA results for the "area" question type in both test sets, and for the "comparison" question type in Test Set 2. Specifically, for the "area" question type, PERS outperforms the previous best by 4.92% in Test Set 1 and by 5.51% in Test Set 2. Across both test sets, PERS achieves SOTA results in terms of both average and overall accuracy.

As shown in Table IV, we compared the performance of various models on the "yes/no" and "land cover" question types within the RSVQAxBEN dataset, including both the average and overall accuracy of all models. Our model achieves SOTA performance in both question types, as well as in terms of average and overall accuracy. It surpasses the previously best-performing model, LiT-4-RSVQA, by 2.77% and 3.41% on the "yes/no" and "land cover" questions, respectively. The average and overall accuracy of PERS are 3.09% and 2.48% higher, respectively, than those of the previously proposed model.

D. Ablation Study

To further validate the efficacy of the proposed method, we conducted comprehensive ablation studies across all three RSVQA datasets. As indicated in Table VI, omitting the pretrained ViT weights from the large-scale RSIs datasets resulted in a performance decline, with the most notable decrease being a 1.19% drop in average accuracy in the RSVQA-LR dataset. Removing the multimodal fusion module had a more substantial impact than omitting pretrained weights. Eliminating the multimodal fusion module resulted in decreases of 2.47% and 2.54% in overall and average accuracy, respectively, in the RSVQA-LR dataset, and more than a 1% reduction in RSVQA-HR Test Set 1. In RSVQA-HR Test Set 2, overall and average accuracies decreased by 1.6% and 1.51%, respectively. Similarly, in the

TABLE VI
Ablation Study Using PERS Models With Different Variants on RSVQA-LR, RSVQA-HR, and RSVQAxBEN Datasets

Variant	RSVQA-LR		RSVQA-HR Test Set 1		RSVQA-HR Test Set 2		RSVQAxBEN	
	Overall accuracy	Average accuracy	Overall accuracy	Average accuracy	Overall accuracy	Average accuracy	Overall accuracy	Average accuracy
w/o pre-train	86.01%	87.93%	85.25%	85.31%	81.16%	81.08%	83.14%	66.77%
w/o multimodal fusion	84.42%	86.58%	84.88%	85.01%	80.25%	80.31%	82.44%	65.90%
Full parameters training	86.84%	**89.17%**	**86.26%**	**86.43%**	**81.93%**	81.78%	**84.02%**	**67.91%**
LoRA	85.55%	86.32%	85.50%	85.68%	81.78%	81.73%	82.21%	65.70%
PERS	**86.89%**	89.12%	86.15%	86.31%	81.85%	**81.82%**	83.75%	67.70%

Best results shown in bold.

TABLE VII
Comparison of Different Number of Layers in the Multimodal Fusion Module on RSVQA-LR, RSVQA-HR and RSVQAxBEN Datasets

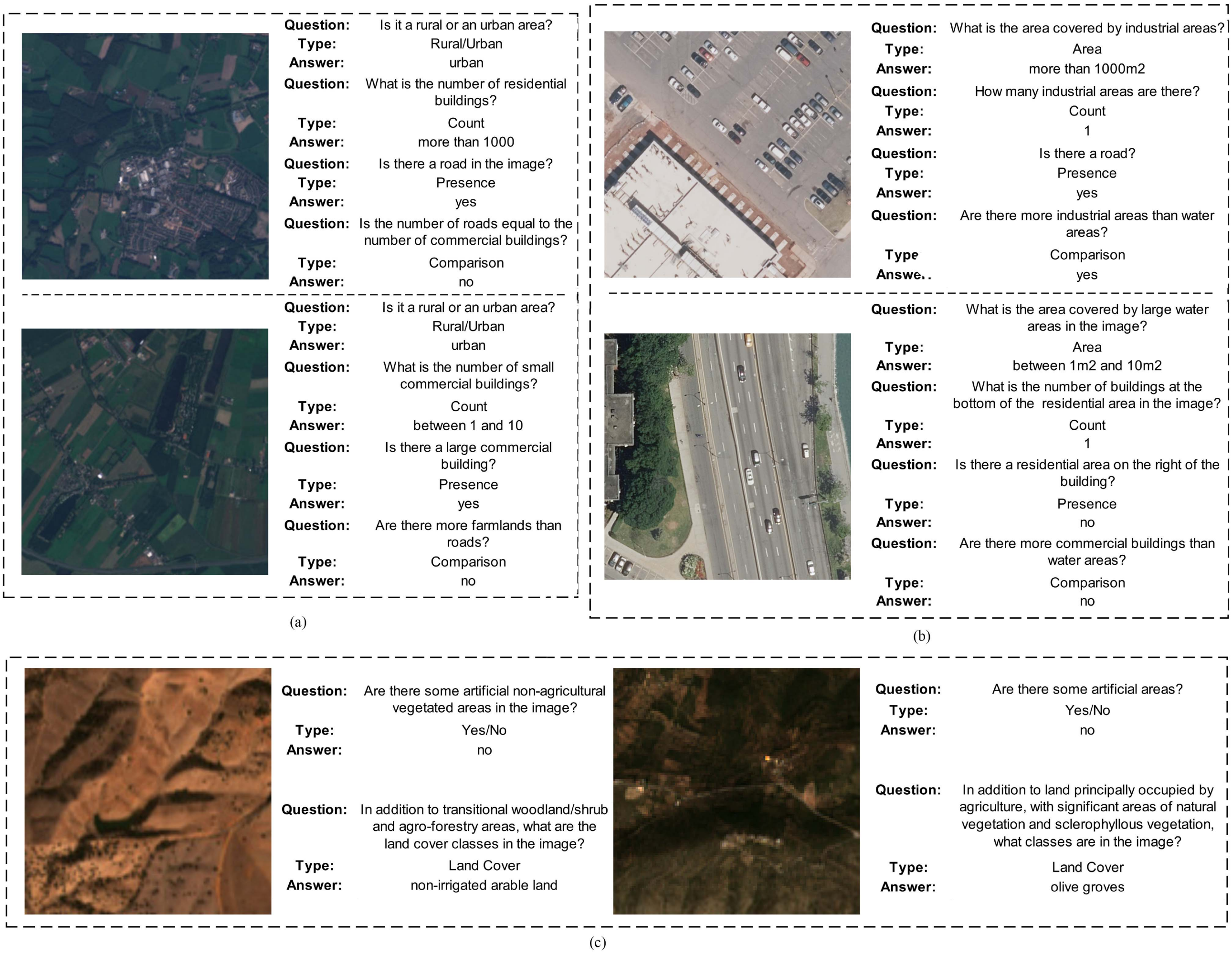
Number of layers	RSVQA-LR		RSVQA-HR Test Set 1		RSVQA-HR Test Set 2		RSVQAxBEN	
	Overall accuracy	Average accuracy	Overall accuracy	Average accuracy	Overall accuracy	Average accuracy	Overall accuracy	Average accuracy
1	86.45%	87.90%	85.99%	86.18%	81.62%	81.39%	83.23%	67.42%
2	86.65%	87.51%	86.03%	86.19%	81.54%	81.28%	83.59%	67.54%
3	86.79%	88.62%	**86.17%**	**86.34%**	81.75%	81.78%	83.68%	67.48%
4	**86.89%**	**89.12%**	86.15%	86.31%	**81.85%**	**81.82%**	**83.75%**	**67.70%**
5	86.79%	88.23%	86.03%	86.15%	81.67%	81.59%	83.69%	67.54%
6	86.58%	88.12%	86.08%	86.12%	81.38%	81.47%	83.64%	67.59%

Best results shown in bold.

RSVQAxBEN dataset, overall and average accuracies decreased by 1.31% and 1.8%, respectively. In additton, training with full parameters yielded results comparable to adapter-based fine-tuning, with gaps ranging from 0.04% to 0.12% in the RSVQA-LR and RSVQA-HR datasets, respectively, with the largest gap being 0.27% in the RSVQAxBEN dataset. Notably, the overall accuracy in the RSVQA-LR and the average accuracy in the RSVQA-HR Test Set 2 were slightly higher than those achieved with full parameter training. Meanwhile, we replaced the adapter with a fully connected layer enhanced by low-rank adaptation (LoRA). As a parameter-efficient transfer learning technique, the model's performance, after the replacement with LoRA, was inferior to our method across the three datasets. Specifically, on the RSVQA-HR Test Set 2, the overall accuracy was the closest, differing by 0.07%. The largest performance gap was in the average accuracy on the RSVQA-LR dataset, with a difference of 2.8%. On average, across the three datasets, the performance was 1.14% lower than our method.

To determine the optimal number of layers in the multimodal fusion module, we conducted a comprehensive ablation study. As shown in Table VII, we evaluated the performance of the model across the RSVQA-LR, RSVQA-HR, and RSVQAxBEN datasets with varying numbers of layers, ranging from 1 to 6. When the multimodal fusion module consisted of a single layer, the model achieved an overall accuracy of 86.45% and an average accuracy of 87.9% on the RSVQA-LR dataset. In RSVQA-HR Test Set 1, the model's overall accuracy was 85.99%, and the average accuracy was 86.18%. In RSVQA-HR Test Set 2, the model's overall accuracy was 81.62%, and the average accuracy was 81.63%. In the RSVQAxBEN dataset, the overall accuracy was 83.23% and the average accuracy was 67.42%. Interestingly, we observed that the model's performance improved with an increase in the number of module layers, achieving optimal performance when the multimodal fusion module contained four layers. Moreover, increasing the number of layers to five or more resulted in performance leveling off or slightly degrading, suggesting that adding additional layers to the multimodal fusion module does not necessarily improve performance and may introduce unnecessary complexity.

To further examine our method's sensitivity to the size of training datasets, we conducted detailed ablation studies on the RSVQA-LR and RSVQA-HR datasets. As indicated in Table V, our experiments encompassed various training set sizes, including 10%, 20%, 30%, 40%, and the entire training set. We was observed that PERS achieved an overall accuracy of 83.63% and an average accuracy of 85.28% on the RSVQA-LR dataset with only 10% of the training data, demonstrating strong adaptability to small-scale datasets. As the size of the training set gradually increased, the model's accuracy steadily improved. When the training dataset size reached 30%, the overall accuracy of PERS was nearly equivalent to that of the current best-performing model, MADNet [43]. Specifically, when the training set size increased to 40%, PERS outperformed MADNet [43] in both average and overall accuracy. Similarly, with only 10% of the training data for the RSVQA-HR datasets, the overall accuracies for Test Set 1 and Test Set 2 were 85.51% and 79.88%, respectively. When the training set size was increased to 20%, PERS had already surpassed MADNet, and at 40%, its performance was nearly comparable to that achieved with the full datasets. Fig. 3 further illustrates the performance of PERS and bimodal [28] after training on datasets of varying sizes within the RSVQA-LR and RSVQA-HR datasets. In these line charts,



Question: Is it a rural or an urban area?
Type: Rural/Urban
Answer: urban
Question: What is the number of residential buildings?
Type: Count
Answer: more than 1000
Question: Is there a road in the image?
Type: Presence
Answer: yes
Question: Is the number of roads equal to the number of commercial buildings?
Type: Comparison
Answer: no

Question: Is it a rural or an urban area?
Type: Rural/Urban
Answer: urban
Question: What is the number of small commercial buildings?
Type: Count
Answer: between 1 and 10
Question: Is there a large commercial building?
Type: Presence
Answer: yes
Question: Are there more farmlands than roads?
Type: Comparison
Answer: no

(a)

Question: What is the area covered by industrial areas?
Type: Area
Answer: more than 1000m2
Question: How many industrial areas are there?
Type: Count
Answer: 1
Question: Is there a road?
Type: Presence
Answer: yes
Question: Are there more industrial areas than water areas?
Type Comparison
Answe.. yes

Question: What is the area covered by large water areas in the image?
Type: Area
Answer: between 1m2 and 10m2
Question: What is the number of buildings at the bottom of the residential area in the image?
Type: Count
Answer: 1
Question: Is there a residential area on the right of the building?
Type: Presence
Answer: no
Question: Are there more commercial buildings than water areas?
Type: Comparison
Answer: no

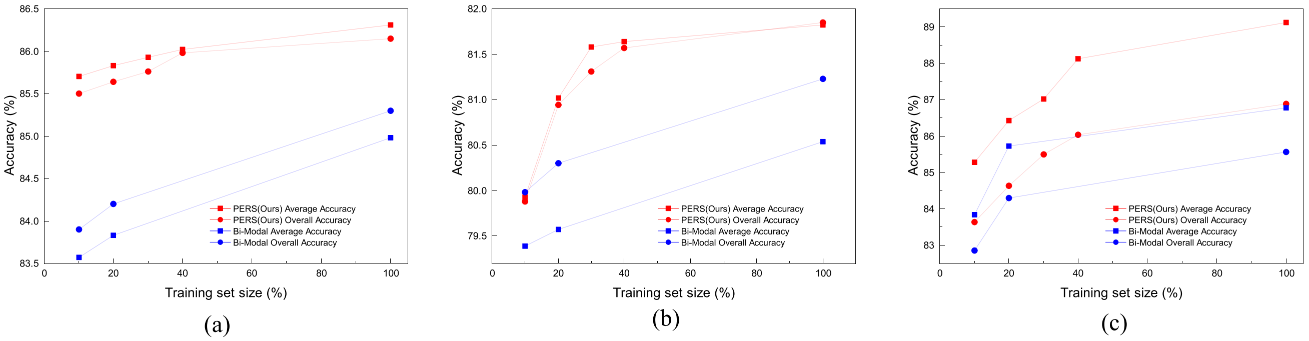
(b)

Question: Are there some artificial non-agricultural vegetated areas in the image?
Type: Yes/No
Answer: no
Question: In addition to transitional woodland/shrub and agro-forestry areas, what are the land cover classes in the image?
Type: Land Cover
Answer: non-irrigated arable land

Question: Are there some artificial areas?
Type: Yes/No
Answer: no
Question: In addition to land principally occupied by agriculture, with significant areas of natural vegetation and sclerophyllous vegetation, what classes are in the image?
Type: Land Cover
Answer: olive groves

(c)

Fig. 2. Examples of three RSVQA datasets, including different question types.



(a) (b) (c)

Fig. 3. Comparison of performance of different models on RSVQA-LR and RSVQA-HR datasets after training on different training set size. (a) RSVQA-LR. (b) RSVQA-HR Test Set 1. (c) RSVQA-HR Test Set 2.

PERS's performance is represented by the red line, while the blue line depicts the performance of bimodal [28]. It is evident that PERS consistently outperforms bimodal [28] across the same sizes of training datasets. This demonstrates that PERS can achieve competitive performance even with a relatively limited amount of training data.

To explore the impact of the dimension m in the adapter, we conducted comprehensive ablation experiments for different sizes of m, as shown in Table VIII. The other dimension size, d, represents the input and output dimension of the adapter and must be consistent with the hidden layer dimension size of ViT, thus, it is fixed at 768. In Table VIII, the smallest m is 8 and the largest is 512, with intermediate values of 16, 32, 64, 128, and 256, which are common reduced dimensions in parameter-efficient transfer learning techniques. In addition, 384, which is half of the ViT hidden layer size of 768, is the

TABLE VIII
PERFORMANCE OF PERS IN RSVQA-LR AND RSVQA-HR DATASETS AFTER TRAINING ON DIFFERENT M SIZES IN ADAPTER

m	RSVQA-LR		RSVQA-HR Test Set 1		RSVQA-HR Test Set 2	
	Overall accuracy	Average accuracy	Overall accuracy	Average accuracy	Overall accuracy	Average accuracy
8	86.67%	86.83%	86.47%	86.62%	81.53%	81.60%
16	86.55%	87.19%	86.40%	86.54%	81.66%	81.68%
32	86.57%	87.23%	86.49%	86.63%	81.49%	81.63%
64	86.41%	87.48%	86.59%	86.75%	81.73%	**81.83%**
128	86.81%	87.82%	86.42%	86.59%	81.02%	81.02%
256	86.73%	88.17%	**86.67%**	**86.79%**	81.55%	81.68%
384	**86.89%**	**89.12%**	86.15%	86.31%	**81.85%**	81.82%
512	86.53%	88.39%	86.51%	86.64%	80.47%	80.59%

Best results shown in bold.

TABLE IX
IMPACT OF *M* VARIATIONS WITHIN THE ADAPTER ON TRAINABLE PARAMETERS

m	Trainable parameters	Increase in number of parameters (versus m=8)	Percentage increase (versus m=8)
8	188 700 705	0	0.00%
16	188 737 593	36888	0.02%
32	188 811 369	110664	0.06%
64	188 958 921	258 216	0.14%
128	189 254 025	553 320	0.29%
256	189 844 233	1 143 528	0.61%
384	190 434 441	1 733 736	0.92%
512	191 024 649	2 323 944	1.23%

final size of *m* chosen for our method. As m gradually increases from 8 to 256 and 384, the model's performance on different datasets reaches its peak. However, when further increased to 512, the model's performance decreases compared to 384 on the RSVQA-LR and RSVQA-HR Test Set 2 datasets, with average decreases of 0.55% and 1.31%, respectively. On the RSVQA-HR Test Set 1, there is a slight increase of approximately 0.35%. Therefore, it is evident that the size of *m* in the adapter should be neither too large nor too small, and approximately half the size of *d* is optimal. Meanwhile, we also studied the change in the number of trainable parameters, as shown in Table IX. Compared to when m is 8, as *m* increases, the growth rate of the number of trainable parameters also increases. However, even when *m* is raised to 512, the proportion of additional trainable parameters is only 1.23%. When the model performs best at 384, the number of trainable parameters increases by only 0.92% compared to the smallest *m*. As shown in Table VIII, the model's performance is optimal at *m* values of 256 and 384, and the increase in parameters is within an acceptable range. Therefore, these two *m* values are ideal choices.

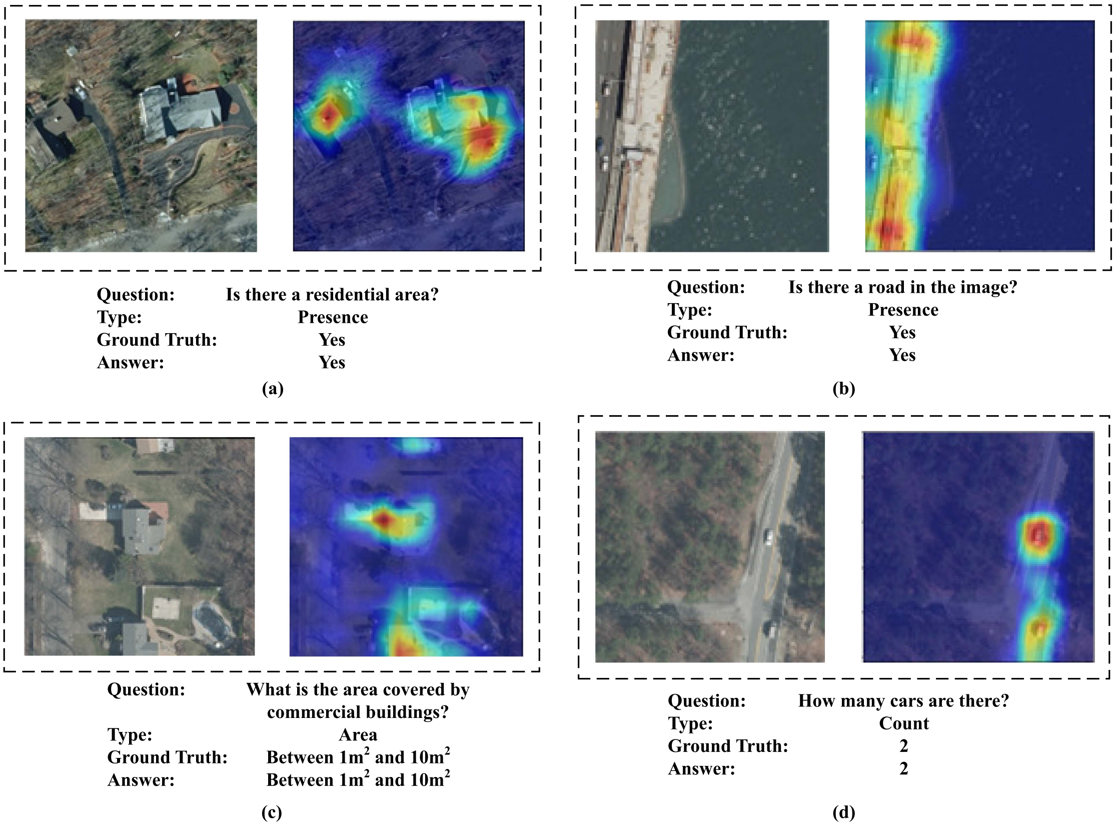
E. Visualization

To further enhance our model's interpretability, we employed Grad-CAM [47] to visualize the cross-attention maps that our model predicts for the RSVQA task. This approach highlights the image regions most influential in the model's decision-making process for answering specific questions. As illustrated in Fig. 4, we present four examples of attention maps overlaid on the original images from the RSVQA-HR dataset. In Fig. 4(a), the image corresponds to the question "Is there a residential area?" Consequently, the model focuses on the residential areas depicted. Similarly, Fig. 4(d) corresponds to the question "How many cars are there?" The model focuses on the cars, enabling an accurate count and a prediction of "2." For the "area" question type, as shown in Fig. 4(c), the question asked is "What is the area covered by commercial buildings?" The model concentrates on the commercial buildings depicted. The highlighted section of the image represents the area of the commercial buildings, enabling the model to accurately estimate this area as "between 1 and 10m^2," consistent with the ground truth. These visualizations further elucidate how the model interprets image and text information to generate corresponding answers, underscoring its capability to comprehensively understand the input images and language data.

V. DISCUSSION

The experimental results demonstrate that PERS outperforms existing top models in both overall and average accuracy across three benchmark RS datasets: RSVQA-LR, RSVQA-HR, and RSVQAxBEN. PERS achieves optimal performance in "count" and "rural/urban" question types within the RSVQA-LR dataset. In the RSVQA-HR dataset, PERS shows superior performance in the "area" question type in Test Set 1 and both "area" and "comparison" types in Test Set 2, with similarly outstanding results in RSVQAxBEN. PERS is specifically designed for the RSVQA task and incorporates a lightweight, parameter-efficient adapter utilizing pretrained weights. The modular design of the adapter enables independent training and adjustment, reducing interference with the main model. Training adapter keeps the majority of the vision encoder parameters unchanged, preserving the extensive knowledge embedded in the pretrained model. Specifically, in the vision encoder, the trainable parameters account for only about 1.98% of the total parameters, with the total reaching approximately 90 M. Although full-parameter training can bring slight improvements, as shown in Table VI, the largest performance improvement appears in the RSVQAxBEN dataset, which is only 0.27%. Moreover, in some datasets, such as RSVQA-LR, the overall accuracy decreases. Meanwhile, the increased training cost is significant, demonstrating the effectiveness and efficiency of this method. This approach enhances performance and generalization on specific tasks, providing a significant advantage in learning new tasks. In addition, the adapter serves as a regularizer to some extent. By limiting changes to the model parameters, the adapter technique mitigates overfitting, especially in data-constrained scenarios. This regularization effect enhances the model's generalization on new tasks. This method reduces training costs and resource consumption while maintaining competitive performance, thus achieving unparalleled success in the RS VQA task. This underscores the feasibility of parameter-efficient approaches for effectively managing large-scale RSIs and highlights the potential of such architectures in RS applications.

Central to PERS, the multimodal fusion module employs self-attention and cross-attention mechanisms, representing a



Question: Is there a residential area?
Type: Presence
Ground Truth: Yes
Answer: Yes

(a)

Question: Is there a road in the image?
Type: Presence
Ground Truth: Yes
Answer: Yes

(b)

Question: What is the area covered by commercial buildings?
Type: Area
Ground Truth: Between 1m^2 and 10m^2
Answer: Between 1m^2 and 10m^2

(c)

Question: How many cars are there?
Type: Count
Ground Truth: 2
Answer: 2

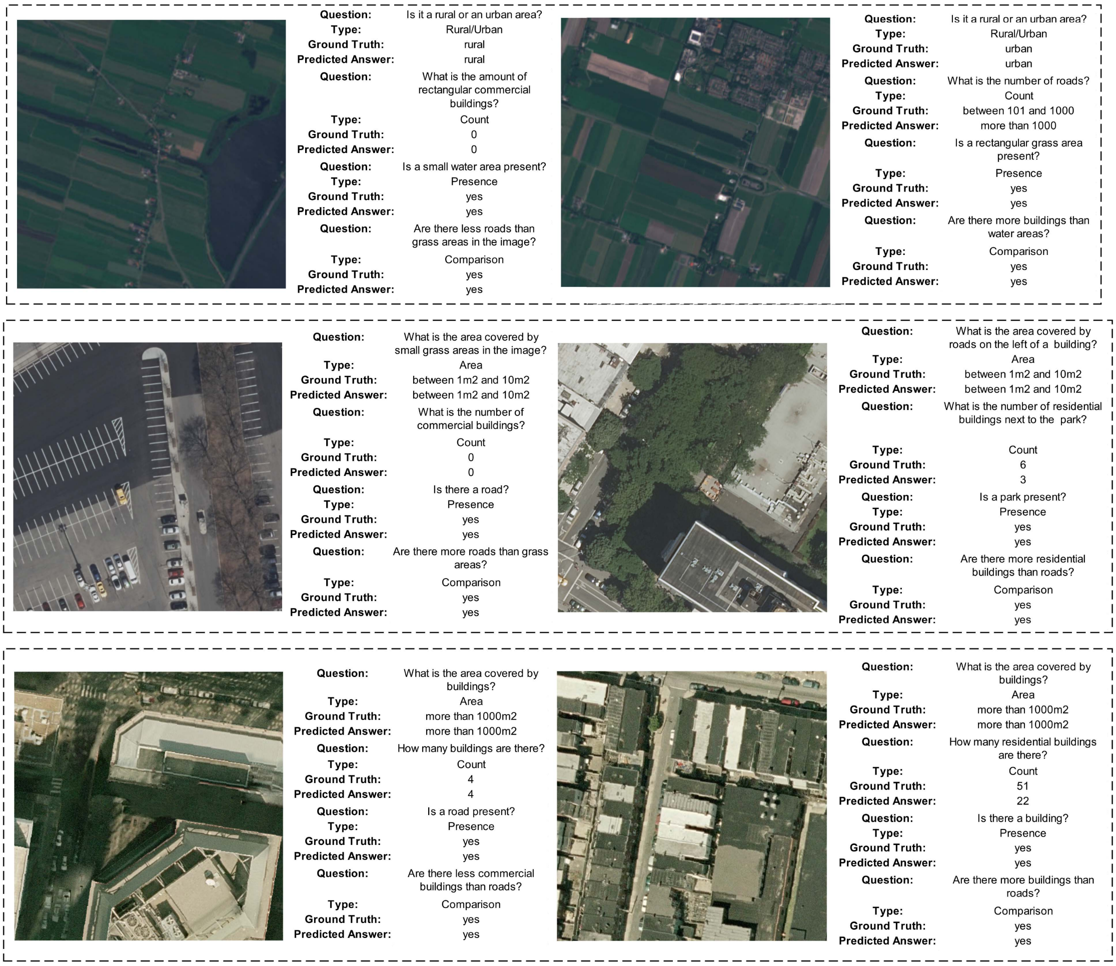
(d)

Fig. 4. Attention maps of RSIs for different question types.

significant advancement in integrating visual and textual information. This architecture facilitates a nuanced and context-sensitive interpretation of input data, crucial for accurately addressing complex questions about RSIs and surpassing existing models on baseline datasets. Table VI demonstrates that the multimodal fusion module significantly enhances model performance. The effective fusion of multimodal information not only achieves higher accuracy but also underscores the importance of intricate intermodal interactions in boosting RSVQA system performance.

Furthermore, as shown in Table VII, we analyzed the model's sensitivity to varying numbers of multimodal fusion layers. As the number of layers increased from 1 to 6, the model's performance on three datasets progressively improved, peaking at three to four layers before gradually declining. Similarly, we analyzed the model's sensitivity to the dimension size m in the adapter, as shown in Table VIII. The performance trend was similar to that observed with the Table VII: as m increased from 8 to 512, the model's performance gradually improved, peaking at 256 and 384, before declining. From these two tables, it is evident that there is a consistent trend of performance improvement followed by a decline as the parameters increase. Notably, we set the maximum number of fusion layers to 6, which is commonly used in practical applications. The theoretical maximum value of m is consistent with d at 768, a frequently used hidden layer size in non-PEFT research. When these two parameters are approximately halved to 3 and 384, the model performs optimally. This observation may serve as a reference for future studies in similar contexts.

Meanwhile, we conducted a detailed sample analysis for PERS, with several typical examples shown in Fig. 5. These samples were selected from the RSVQA-LR, RSVQA-HR Test Set 1, and RSVQA-HR Test Set 2 datasets. PERS exhibits varying performance across different types of questions. Compared to other question types, PERS is most prone to errors in "count" questions. As indicated in Tables I–III, this is also the type where PERS performs the worst overall. For other types of questions, PERS can achieve accuracy rates of 89.12% or even 98.33%. However, in the RSVQA-LR dataset, the accuracy for "count" questions is only 75.79%, and in the RSVQA-HR Test Set 1 and Test Set 2, the accuracy is 69.79% and 62.23%, respectively. From the examples in Fig. 5, it can be observed that PERS performs well when the count result is zero. However, the accuracy decreases as the number of targets, such as roads or houses, increases. For instance, in the bottom right example, the actual number of houses is 51, but the model counts 22. In scenarios with fewer targets, such as the middle right example, where the actual number is 6 and the model counts 3. It can be seen that not all buildings are fully visible in the image, if only buildings with larger segmented areas are considered; the model might interpret there to be only three buildings, specifically in the upper left, upper right, and lower right parts of the image. This phenomenon is not unique to our study. The comparative studies listed in Tables I–III also show poor performance specifically



Question: Is it a rural or an urban area?
Type: Rural/Urban
Ground Truth: rural
Predicted Answer: rural
Question: What is the amount of rectangular commercial buildings?
Type: Count
Ground Truth: 0
Predicted Answer: 0
Question: Is a small water area present?
Type: Presence
Ground Truth: yes
Predicted Answer: yes
Question: Are there less roads than grass areas in the image?
Type: Comparison
Ground Truth: yes
Predicted Answer: yes

Question: Is it a rural or an urban area?
Type: Rural/Urban
Ground Truth: urban
Predicted Answer: urban
Question: What is the number of roads?
Type: Count
Ground Truth: between 101 and 1000
Predicted Answer: more than 1000
Question: Is a rectangular grass area present?
Type: Presence
Ground Truth: yes
Predicted Answer: yes
Question: Are there more buildings than water areas?
Type: Comparison
Ground Truth: yes
Predicted Answer: yes

Question: What is the area covered by small grass areas in the image?
Type: Area
Ground Truth: between 1m2 and 10m2
Predicted Answer: between 1m2 and 10m2
Question: What is the number of commercial buildings?
Type: Count
Ground Truth: 0
Predicted Answer: 0
Question: Is there a road?
Type: Presence
Ground Truth: yes
Predicted Answer: yes
Question: Are there more roads than grass areas?
Type: Comparison
Ground Truth: yes
Predicted Answer: yes

Question: What is the area covered by roads on the left of a building?
Type: Area
Ground Truth: between 1m2 and 10m2
Predicted Answer: between 1m2 and 10m2
Question: What is the number of residential buildings next to the park?
Type: Count
Ground Truth: 6
Predicted Answer: 3
Question: Is a park present?
Type: Presence
Ground Truth: yes
Predicted Answer: yes
Question: Are there more residential buildings than roads?
Type: Comparison
Ground Truth: yes
Predicted Answer: yes

Question: What is the area covered by buildings?
Type: Area
Ground Truth: more than 1000m2
Predicted Answer: more than 1000m2
Question: How many buildings are there?
Type: Count
Ground Truth: 4
Predicted Answer: 4
Question: Is a road present?
Type: Presence
Ground Truth: yes
Predicted Answer: yes
Question: Are there less commercial buildings than roads?
Type: Comparison
Ground Truth: yes
Predicted Answer: yes

Question: What is the area covered by buildings?
Type: Area
Ground Truth: more than 1000m2
Predicted Answer: more than 1000m2
Question: How many residential buildings are there?
Type: Count
Ground Truth: 51
Predicted Answer: 22
Question: Is there a building?
Type: Presence
Ground Truth: yes
Predicted Answer: yes
Question: Are there more buildings than roads?
Type: Comparison
Ground Truth: yes
Predicted Answer: yes

Fig. 5. Performance of PERS on typical samples in RSVQA-LR and RSVQA-HR datasets, from top to bottom, are RSVQA-LR,RSVQA-HR Test Set 1,RSVQA-HR Test Set 2, respectively.

on "count" questions. This insight may guide future research in RS.

Despite PERS's advantages, its limitations offer significant opportunities for further research. Although effective, the inclusion of multimodal fusion modules in PERS may introduce inefficiencies in scenarios that demand real-time processing. Future research on PERS could explore more dynamic and scalable multimodal integration methods and more efficient parameter training strategies, potentially employing advanced transfer learning or adaptive learning techniques, to better balance performance with computational costs. Furthermore, current research on multimodal large language models is expanding, demonstrating significant achievements in both general and specialized domains, such as healthcare and autonomous driving. This trend suggests a promising research trajectory for RS applications.

VI. Conclusion

In this article, we propose PERS, a novel parameter-efficient multimodal transfer learning model for RSVQA. We initialize the vision encoder with a ViT pretrained on large-scale RSIs datasets and introduce a lightweight, parameter-efficient adapter module within this encoder. By training on only a minimal subset of parameters, we achieve performance nearly equivalent to that of fine-tuning with full parameters. Concurrently, we develop a multimodal fusion module that utilizes both self-attention and cross-attention mechanisms, enabling the model to fully leverage multimodal information. Our approach achieved SOTA performance on three benchmark RS VQA datasets, and we further validated our model's exceptional performance on extremely limited training datasets. Moreover, we demonstrate our model's interpretability by showcasing attention maps.

References

[1] M. Zhang, H. Luo, W. Song, H. Mei, and C. Su, "Spectral-spatial offset graph convolutional networks for hyperspectral image classification," *Remote. Sens.*, vol. 13, no. 21, 2021, Art. no. 4342, doi: 10.3390/rs13214342.

[2] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019, doi: 10.1109/TGRS.2019.2899129.

[3] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019, doi: 10.1109/TGRS.2018.2864987.

[4] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021, doi: 10.1109/TGRS.2020.3016820.

[5] W. Cai and Z. Wei, “Remote sensing image classification based on a cross-attention mechanism and graph convolution,” *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, 2022, Art. no. 8002005, doi: 10.1109/LGRS.2020.3026587.

[6] G. Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016, doi: 10.1109/TGRS.2016.2601622.

[7] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, “Attentional local contrast networks for infrared small target detection,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 11, pp. 9813–9824, Nov. 2021, doi: 10.1109/TGRS.2020.3044958.

[8] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, “Cross-scale feature fusion for object detection in optical remote sensing images,” *IEEE Geosci. Remote. Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2021, doi: 10.1109/LGRS.2020.2975541.

[9] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, “Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model,” *IEEE Geosci. Remote. Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021, doi: 10.1109/LGRS.2020.2988032.

[10] H. Chen, Z. Qi, and Z. Shi, “Remote sensing image change detection with transformers,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 5607514, doi: 10.1109/TGRS.2021.3095166.

[11] L. Ding, H. Tang, and L. Bruzzone, “LANet: Local attention embedding to improve the semantic segmentation of remote sensing images,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 1, pp. 426–435, Jul. 2021, doi: 10.1109/TGRS.2020.2994150.

[12] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, “SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images,” *IEEE Geosci. Remote. Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021, doi: 10.1109/LGRS.2020.2988294.

[13] W. Liu, F. Su, X. Jin, H. Li, and R. Qin, “Bispace domain adaptation network for remotely sensed semantic segmentation,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 5600211, doi: 10.1109/TGRS.2020.3035561.

[14] S. Lobry, D. Marcos, J. Murray, and D. Tuia, “RSVQA: Visual question answering for remote sensing data,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 58, no. 12, pp. 8555–8566, Dec. 2020, doi: 10.1109/TGRS.2020.2988782.

[15] Z. Yuan, L. Mou, and X. X. Zhu, “Self-paced curriculum learning for visual question answering on remote sensing data,” in *2021 IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2999–3002.

[16] Z. Yuan, L. Mou, Q. Wang, and X. X. Zhu, “From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 5623111, doi: 10.1109/TGRS.2022.3173811.

[17] X. Zheng, B. Wang, X. Du, and X. Lu, “Mutual attention inception network for remote sensing visual question answering,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, 2022, Art. no. 5606514, doi: 10.1109/TGRS.2021.3079918.

[18] C. Chappuis, V. Zermatten, S. Lobry, B. L. Saux, and D. Tuia, “Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, New Orleans, LA, USA, Jun. 19–20, 2022, pp. 1371–1380, doi: 10.1109/CVPRW56347.2022.00143.

[19] Z. Zhang et al., “A spatial hierarchical reasoning network for remote sensing visual question answering,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 61, 2023, Art. no. 4400815, doi: 10.1109/TGRS.2023.3237606.

[20] C. Chappuis, S. Lobry, B. Kellenberger, B. L. Saux, and D. Tuia, “How to find a good image-text embedding for remote sensing visual question answering?,” in *Proc. MACLEAN: Mach. Learn. Earth Observ. Workshop Co-Located Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases*, ser. CEUR Workshop Proceedings, T. Corpetti, D. Ienco, R. Interdonato, M. Pham, and S. Lefèvre, Eds., vol. 3088, CEUR-WS.org, 2021, pp. 1–10. [Online]. Available: https://ceur-ws.org/Vol-3088/paper1.pdf

[21] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, Nov. 1–4, 2016, J. Su, X. Carreras, and K. Duh, Eds. Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2016, pp. 457–468, doi: 10.18653/v1/d16-1044.

[22] H. Ben-Younes, R. Cadène, M. Cord, and N. Thome, “MUTAN: Multimodal tucker fusion for visual question answering,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 22–29, 2017, IEEE Computer Society, 2017, pp. 2631–2639, doi: 10.1109/ICCV.2017.285.

[23] A. Farinhas, A. F. T. Martins, and P. M. Q. Aguiar, “Multimodal continuous visual attention mechanisms,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, ICCVW 2021*, Montreal, BC, Canada, Oct. 11–17, 2021, IEEE, 2021, pp. 1047–1056, doi: 10.1109/ICCVW54120.2021.00122.

[24] S. Antol et al., “VQA: Visual question answering,” in *2015 IEEE Int. Conf. Comput. Vis.*, ICCV 2015, Santiago, Chile, Dec. 7–13, 2015, IEEE Computer Society, 2015, pp. 2425–2433, doi: 10.1109/ICCV.2015.279.

[25] J. Wang, Y. Ji, J. Sun, Y. Yang, and T. Sakai, “MIRTT: Learning multimodal interaction representations from trilinear transformers for visual question answering,” in *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic*, 16–20 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2021, pp. 2280–2292, doi: 10.18653/v1/2021.findings-emnlp.196.

[26] A. F. T. Martins, A. Farinhas, M. V. Treviso, V. Niculae, P. M. Q. Aguiar, and M. A. T. Figueiredo, “Sparse and continuous attention mechanisms,” in *Proc. Adv. Neural Inf. Process. Syst. 33: Annu. Conf. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020, pp. 20989–21001. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/f0b76267fbe12b936bd65e203dc675c1-Abstract.html

[27] Z. Chen, J. Chen, Y. J. Z. Geng, Z. P. Yuan, and H. Chen, “Zero-shot visual question answering using knowledge graph,” in *Proc. 20th Int. Semantic Web Conf. Semantic Web*, ser. Lecture Notes in Computer Science, A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, P. M. Barnaghi, A. Haller, M. Dragoni, and H. Alani, Eds., vol. 12922, Springer, 2021, pp. 146–162, doi: 10.1007/978-3-030-88361-4_9.

[28] Y. Bazi, M. M. A. Rahhal, M. L. Mekhalfi, M. A. A. Zuair, and F. Melgani, “Bi-modal transformer-based approach for visual question answering in remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4708011, doi: 10.1109/TGRS.2022.3192460.

[29] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, “AdaptFormer: Adapting vision transformers for scalable visual recognition,” in *Proc. Adv. Neural Inf. Process. Syst. 35: Annu. Conf. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. New Orleans, LA, USA, 2022, pp. 16664–16678. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/69e2f49ab0837b71b0e0cb7c555990f8-Abstract-Conference.html

[30] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li, “St-adapter: Parameter-efficient image-to-video transfer learning,” in *Proc. Adv. Neural Inf. Process. Syst. 35: Annu. Conf. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., New Orleans, LA, USA, 2022, pp. 1–16. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/a92e9165b22d4456fc6d87236e04c266-Abstract-Conference.html

[31] S. Rebuffi, H. Bilen, and A. Vedaldi, “Learning multiple visual domains with residual adapters,” in *Proc. Adv. Neural Inf. Process. Syst. 30: Annu. Conf. Neural Inf. Process. Syst.*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds. Long Beach, CA, USA, 2017, pp. 506–516. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/e7b24b112a44fdd9ee93bdf998c6ca0e-Abstract.html

[32] N. Houlsby et al., “Parameter-efficient transfer learning for NLP,” in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, 2019, pp. 2790–2799. [Online]. Available: http://proceedings.mlr.press/v97/houlsby19a.html

[33] Y. Lee, Y. Tsai, W. Chiu, and C. Lee, “Multimodal prompting with missing modalities for visual recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 14943–14952, doi: 10.1109/CVPR52729.2023.01435.

[34] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, “Prompting visual-language models for efficient video understanding,” in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Proceedings, Part XXXV, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13695, Springer, 2022, pp. 105–124, doi: 10.1007/978-3-031-19833-5_7.

[35] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” in *Proc. 59th Ann. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Association for Computational Linguistics, 2021, pp. 4582–4597, doi: 10.18653/v1/2021.acl-long.353.

[36] E. J. Hu et al., “LORA: Low-rank adaptation of large language models,” in *Proc. 10th Int. Conf. Learn. Representations*, Virtual Event, Apr. 25–29, 2022, OpenReview.net, 2022, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[37] D. Wang et al., “Advancing plain vision transformer towards remote sensing foundation model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251402536

[38] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proc. 54th Annu. Meeting Association Comput. Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds., Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[39] S. Lobry, B. Demir, and D. Tuia, “RSVQA meets BigEarthNet: A new, large-scale, visual question answering dataset for remote sensing,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Brussels, Belgium, 2021, pp. 1218–1221, doi: 10.1109/IGARSS47720.2021.9553307.

[40] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “RandAugment: Practical automated data augmentation with a reduced search space,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 3008–3017.

[41] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Minneapolis, MN, USA, Jun. 2–7, 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.

[42] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. 7th Int. Conf. Learn. Representations*, New Orleans, LA, USA, May 2019, OpenReview.net, pp. 1–19. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[43] Y. Li et al., “Enhancing remote sensing visual question answering: A mask-based dual-stream feature mutual attention network,” *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024. [Online]. Available: https://doi.org/10.1109/LGRS.2024.3389042

[44] T. Siebert, K. N. Clasen, M. Ravanbakhsh, and B. Demir, “Multi-modal fusion transformer for visual question answering in remote sensing,” *Remote Sens.*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252781182

[45] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *Proc. 38th Int. Conf. Mach. Learn.*, Virtual Event, 2021, pp. 10347–10357, [Online]. Available: http://proceedings.mlr.press/v139/touvron21a.html

[46] L. W. Hackel, K. N. Clasen, M. Ravanbakhsh, and B. Demir, “LIT-4-RSVQA: Lightweight transformer-based visual question answering in remote sensing,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Pasadena, CA, USA, Jul. 16–21, 2023, pp. 2231–2234, doi: 10.1109/IGARSS52108.2023.10281674.

[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.

Jinlong He (Student Member, IEEE) received the B.S. degree in software engineering from Harbin Engineering University, Harbin, China, in 2022, where he is currently working toward the M.S. degree in software engineering with the Harbin Engineering University.

His research interests include medical visual question answering, medical multimodality, remote sensing visual question answering, and medical multimodal large language models.

Gang Liu (Member, IEEE) received the B.S. degree in computer and application from the Department of Computer and Information Science, Harbin Engineering University, Harbin, China, in 1999, the M.S. degree in computer application technology from Harbin Engineering University, in 2004, and the Ph.D. degree in computer application technology from the College of Computer Science and Technology, Harbin Engineering University, in 2008.

He was a Visiting Scholar with the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2005, and with Monash University, Melbourne, VIC, Australia, in 2014. His research interests include artificial intelligence, large language models, vision-language pretraining models, remote sensing visual question answering, and multimodal knowledge graphs.

Pengfei Li (Student Member, IEEE) received the B.S. degree from the School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou, China, in 2021, and the M.S. degree in electronic and information engineering from the College of Computer Science and Technology, Harbin Engineering University, Harbin, China, in 2024.

His research interests include computer vision, natural language processing, remote sensing visual question answering, vision-language pretraining, medical visual question answering, and medical image processing.

Xiaonan Su is currently working toward the M.S. degree in computer science and technology with Harbin Engineering University, Harbin, China.

His research interests include remote sensing image and text retrieval.

Wenhua Jiang received the B.S. degree in computer science and technology from Changzhou University, Changzhou, China, in 2022. He is currently working toward the M.S. degree in electronic and information engineering with the School of Computer Science and Technology, Harbin Engineering University, Harbin, China.

His research interests include graph-based information retrieval, visual question answering, and natural language processing.

Dongze Zhang received the B.S. degree in computer science and technology from Harbin Engineering University, Harbin, China, in 2023. He is currently working toward the master’s degree with the School of Computer Science and Technology, Harbin Engineering University.

His research interests include medical VQA, multimodality, and natural language processing.

Shenjun Zhong received the B.S. degree in engineering from the University of Shanghai for Science and Technology, Shanghai, China, in 2009, and the M.S. degree in information technology and the Ph.D. degree in biomedical imaging and information technology from Monash University, Melbourne, VIC, Australia, in 2012 and 2016, respectively.

He is currently a Research Scientist with Monash Biomedical Imaging, Monash University, and an Informatics Fellow with National Imaging Facility, Saint Lucia, QLD, Australia. His research interests include biomedical and neuroimaging, medical image analysis, and the application of deep learning in multimodality large language models and bioinformatics.