Semantic-Aware Remote Sensing Visual Question Answering via Segmentation-Guided Learning

Shuyi Wen, Aihua Mao, Ran Yi, and Yong-Jin Liu, *Senior Member, IEEE*

***Abstract*—Visual question answering (VQA) in remote sensing has emerged as a challenging yet promising research area due to the complexity of high-resolution satellite imagery, diverse land cover distributions, and the need for effective multimodal fusion. In this article, we propose a novel remote sensing VQA framework that leverages semantic segmentation results to enhance the extraction and focus of critical visual features. Our model employs a dual-branch feature extraction strategy: one branch captures pixel-level features using a convolutional neural network (CNN) backbone, while the other utilizes a graph neural network (GNN) to model the relationships among segmented regions, thereby extracting high-level, category-specific semantic information. In addition, an auxiliary contrastive learning module is integrated to align key terms in the question with corresponding regions in the segmentation map, effectively distinguishing foreground from background elements and reinforcing the model's attention on question-relevant areas. Extensive experiments on the EarthVQA, RescueNet-VQA, and FloodNet-VQA datasets demonstrate that our approach significantly outperforms state-of-the-art methods in terms of VQA accuracy. Our results highlight the benefits of incorporating semantic segmentation as a prior, offering a comprehensive and interpretable solution that lays the groundwork for future research in remote sensing VQA.**

***Index Terms*—Contrastive learning module, graph neural network (GNN), remote sensing visual question answering (VQA), semantic segmentation.**

I. Introduction

THE rapid development of remote sensing technology has significantly enhanced Earth observation capabilities, enabling the extraction of valuable information from high-resolution satellite imagery, unmanned aerial vehicles (UAVs), and multispectral data [1]. These advancements have led to widespread applications in various fields, including agriculture, forestry, environmental monitoring, and urban planning. With increasing availability and complexity of remote sensing data, traditional manual interpretation methods have been gradually replaced by automated and intelligent approaches, leveraging machine learning techniques for tasks such as scene classification [2], [3], object detection [4], [5], and semantic segmentation [6], [7].

Conventional remote sensing image analysis primarily focuses on visual interpretation, but often overlooks the interactive needs of users. To address this challenge, multimodal vision-language tasks, such as image captioning [8], [9], visual grounding [10], [11], and visual question answering (VQA) [12], [13], have recently gained significant attention. Different from image captioning, which generates general descriptions without fine-grained reasoning, and visual grounding, which focuses on locating specific objects but lacks comprehensive semantic understanding, VQA aims to answer natural language questions about a given image by comprehending both visual and textual information. The core components of VQA models include image and text feature extraction, multimodal feature fusion, and answer generation or classification. Compared to natural image VQA, remote sensing VQA presents unique challenges due to the distinct characteristics of remote sensing imagery, including complex land cover distributions, multiscale feature representations, and the need for effective cross-modal information fusion.

Most existing remote sensing VQA models are adapted from general VQA methods. These approaches typically employ convolutional neural networks (CNNs) [14] to extract global visual features or object detection models to obtain local features, while textual features are processed using recurrent neural networks (RNNs). Transformer-based architectures [15] and attention mechanisms are often used to enhance interactions between visual and textual modalities. However, when applying these methods to remote sensing VQA, two key challenges arise: 1) general VQA methods tend to overlook domain-specific semantic cues such as specialized land cover categories and environmental context, which are crucial for interpreting remote sensing imagery and 2) when addressing questions related to specific local objects, conventional attention mechanisms based solely on image feature values tend to mistakenly focus on regions that display similar features but are not pertinent to the question, especially in remote sensing scenarios where land cover differentiation is ambiguous.

To address these challenges, we propose a novel remote sensing VQA model designed based on semantic segmentation results to enhance visual reasoning. In remote sensing images, the content consists of diverse land cover categories with varying spatial distributions. Compared with methods that directly perform the VQA task, incorporating semantic segmentation labels into the visual feature vector can

Received 20 April 2025; revised 29 September 2025 and 9 January 2026; accepted 2 February 2026. Date of publication 10 February 2026; date of current version 3 March 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62576138, Grant 62461160309, and Grant 62302297; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012791; and in part by Guangzhou Municipal Science and Technology Program key projects under Grant 2023B01J1001. *(Corresponding author: Aihua Mao.)*

Shuyi Wen and Aihua Mao are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: 202321044412@mail.scut.edu.cn; ahmao@scut.edu.cn).

Ran Yi is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: ranyi@sjtu.edu.cn).

Yong-Jin Liu is with BNRist, Department of Computer Science and Technology, MOE Key Laboratory of Pervasive Computing, Tsinghua University, Beijing 100084, China (e-mail: liuyongjin@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TGRS.2026.3663435

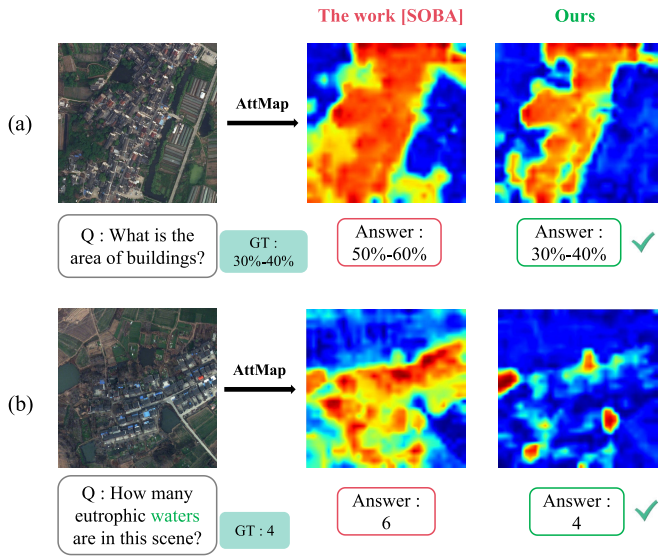

Fig. 1. Comparison of attention maps between our method and a representative method SOBA [16]. The attention maps are visualized as heatmaps, where redder regions indicate higher attention scores. (a) and (b) Correspond to the two challenges discussed earlier.

strengthen the learning of different land cover types. Unlike SOBA [16], which only integrates category labels into the learning process, our approach emphasizes utilizing segmentation results within the VQA task. As illustrated in Fig. 1, our model produces more precise attention maps than SOBA, demonstrating an improved ability to focus on critical image regions. In Fig. 1(a), our model captures high-level semantic information, enabling more accurate region boundaries. In Fig. 1(b), our method learns to attend to key regions relevant to the question, enhancing visual–text alignment. Specifically, we utilize remote sensing VQA datasets containing semantic segmentation masks. For a given image, we first extract visual features using either a CNN [14] or a Transformer-based [15] backbone. Simultaneously, a semantic segmentation model generates segmentation masks, which provide additional structural information. In the VQA pipeline, textual features are extracted using a long short-term memory (LSTM) [17] network, while multimodal feature fusion is performed using a Transformer. The fused representation is processed through a softmax classifier to select the most probable answer from a predefined answer set.

To incorporate high-level semantic information from remote sensing imagery, we introduce a dual-branch feature extraction strategy. One branch extracts pixel-level visual features, while the other builds the relationships between segmented regions using a graph neural network (GNN) [18], where each region serves as a graph node, and the feature relationships between regions form the graph edges. Furthermore, to improve the model's ability to focus on question-relevant image regions, we introduce an auxiliary contrastive learning module beyond the mainstream answer generation module. This module aligns key terms in the question with corresponding segmented regions in the image, distinguishing key (question-relevant) regions from nonkey regions. By separately fusing key region and nonkey region features with textual features and maximizing their differences, our model effectively enhances attention to key image regions while mitigating the influence of irrelevant areas. During training, different weights are assigned to various loss functions, and the model is optimized iteratively through backpropagation. We evaluate our method on the EarthVQA [16] and RescueNet-VQA [19] datasets, both of which contain semantic segmentation annotations. Our model achieves the highest VQA accuracy compared to all existing methods and ranks first on the EarthVQA test set evaluation hosted online on the Codabench platform.

Our contributions can be summarized as follows.

1) We introduce a novel remote sensing VQA framework designed based on semantic segmentation, enabling fine-grained attention to critical visual features at multiple levels.
2) Our model employs a parallel feature extraction strategy that combines pixel-level representations with a graph-based structure, ensuring a comprehensive understanding of high-level semantic information. Additionally, a contrastive learning module further enhances attention to question-relevant regions.
3) We validate our approach on remote sensing VQA datasets containing semantic segmentation masks. Our method surpasses existing state-of-the-art (SOTA) models in VQA accuracy and ranks first on the Codabench platform for EarthVQA evaluation.

II. Related Work

A. Visual Question Answering

VQA has gained significant attention in both computer vision and natural language processing in recent years. Its core objective is to enable models to infer reasonable answers based on an input image and a natural language question. The most widely used datasets for VQA include the manually annotated VQA v1.0 [20] and VQA v2.0 [21], as well as synthetic datasets such as GQA [22] and compositional language and elementary visual reasoning (CLEVR) [23], which obtain questions and answers from real-world or generated images.

Early VQA methods used CNNs for visual feature extraction and RNNs for text processing, followed by feature fusion for answer prediction [24]. With deep learning advancements, attention mechanisms improved visual–text interaction [25]. For instance, stacked attention networks (SANs) [26] employs a multilayer attention mechanism to iteratively refine the focus on key regions. Additionally, bottom-up and top-down (BUTD) [9] was the first to introduce an object-based attention mechanism, leveraging Faster R-CNN [27] to detect salient objects and extract their features. Given the effectiveness of the Transformer [15] architecture, deep modular co-attention networks for visual question answering (MCAN) [12] was developed to improve the alignment of multimodal features. To enhance the robustness of VQA models, D-VQA [28] introduced a unimodal bias detection module, improving its generalization ability. Recent multimodal pretrained models, including ViLBERT [29] and learning cross-modality encoder representations from transformers (LXMERT) [30], adopt a dual-stream BERT structure with co-attention, significantly improving VQA performance. With the widespread adoption

of large language models (LLMs), conditionally generative models such as BLIP-2 [31] and Instruct-BLIP [32] have demonstrated strong adaptability to the VQA domain through fine-tuning. Our model leverages attention mechanisms and Transformer architectures to better integrate visual and textual features, improving VQA reasoning capabilities.

B. Remote Sensing Semantic Segmentation

In recent years, research on semantic segmentation has expanded from general images to remote sensing imagery, aiming to assign semantic labels to each pixel in high-resolution satellite, UAV, or aerial images. Early approaches primarily relied on machine learning-based pixel classification methods [33], such as support vector machines (SVMs), random forest (RF), and extreme gradient boosting (XGBoost). However, these methods, limited by low-dimensional feature representations, struggle with complex land cover classification.

With the rise of deep learning, remote sensing semantic segmentation methods can be broadly categorized into CNN and Transformer-based approaches. CNN-based methods extract local features through convolution operations. For instance, FarSeg [34] enhances foreground-background contrast by redistributing feature maps, while ResUNet-a [35] integrates multibranch ResNet [36] into a UNet encoder–decoder structure, combining pyramid scene parsing pooling for multilevel feature fusion. AFNet [6] employs attention mechanisms to integrate multichannel and multiscale features, improving segmentation accuracy. Given the high computational cost of processing large-scale feature maps, lightweight networks such as ABCNet [37] and ShelfNet [38] have been developed to balance efficiency and accuracy. Transformer-based methods leverage self-attention to capture global dependencies. Zhang et al. [7] proposed a hybrid model combining CNN and Transformer, using Swin Transformer [39] as the encoder and CNN as the decoder to fuse global and local features. Similarly, Cui et al. [40] improved segmentation by integrating Swin Transformer with UPerNet, utilizing convolutional block attention to highlight post-disaster features. Our model leverages semantic segmentation results from remote sensing images, incorporating specialized modules to enhance VQA accuracy in this domain.

C. Remote Sensing VQA

Compared to the well-established research on natural image VQA, remote sensing VQA remains a developing field with significant potential for advancement. Lobry et al. [13] introduced VQA into the remote sensing domain by creating the remote sensing visual question answering (RSVQA) dataset, which includes both low-resolution and high-resolution subsets. Building on this, the RSVQAxBEN [41] dataset was constructed using BigEarthNet [42], providing a large-scale benchmark. The MAIN model [43] improves image–text alignment through attention mechanisms and bilinear fusion while introducing the RSIVQA dataset. Inspired by human learning processes, Yuan et al. [44] proposed a progressive learning approach, enabling models to train on questions of increasing difficulty. Moving beyond traditional CNN and LSTM-based feature extraction, Bazi et al. [45] introduced a Transformer-based framework, employing CLIP [46] to separately extract visual and textual features and leveraging a decoder to model cross-modal relationships. Furthermore, Prompt-RSVQA [47] introduced a novel paradigm that converts visual information into textual prompts for answer generation. Recent research [48] has also explored spatial reasoning, incorporating hash-based multiscale visual learning to better capture the spatial relationships within remote sensing images. Beyond general remote sensing VQA tasks, domain-specific applications have emerged, such as post-disaster damage assessment [19], [49] and change detection [50], demonstrating the potential in diverse spatiotemporal contexts.

More recently, the EarthVQA dataset [16] was introduced, significantly expanding the number of reasoning-based QA pairs while maintaining a balanced distribution of judging and counting questions. The SOBA model, developed for this dataset, enhances local semantic learning and improves the reasoning capability of the model by incorporating category labels obtained from semantic segmentation into the visual feature vectors, thereby enabling more effective learning of regional features. A similar work [51] concatenates segmentation features and textual features to generate an attention map, which is then applied to the visual features to learn category-guided image representations. Differently, our approach leverages semantic segmentation results to inform the design of the VQA framework itself. By constructing graph-structured representations and incorporating question-guided contrastive learning, our method explicitly exploits semantic regions and their relationships to focus on question-relevant areas. This task-oriented use of semantic segmentation enables more effective visual–textual alignment and leads to superior VQA performance across different question types.

III. Method

In this article, we propose a novel model framework for remote sensing VQA, which is designed based on semantic segmentation, as illustrated in Figs. 2 and 3. Our model proposes a dual-branch visual feature extraction strategy that combines basic image features with region-level graph-structured features. To enhance cross-modal understanding, we design a bidirectional cross-attention mechanism for multimodal feature fusion. Additionally, a contrastive learning module is incorporated to help the model focus more effectively on question-relevant image regions.

First, in Section III-A, we introduce semantic segmentation as a prerequisite for our approach. Then, in Section III-B, we present the overall architecture of our model. Section III-C focuses on the multimodal feature fusion strategy and the incorporation of semantic graph-structured features. In Section III-D, we describe an auxiliary contrastive learning strategy that processes images into positive and negative samples. Finally, in Section III-E, we analyze the total loss function used during training.

A. Prior Knowledge: Semantic Segmentation of Remote Sensing Images

Semantic segmentation of remote sensing images aims to assign a semantic category label to each pixel. As shown in

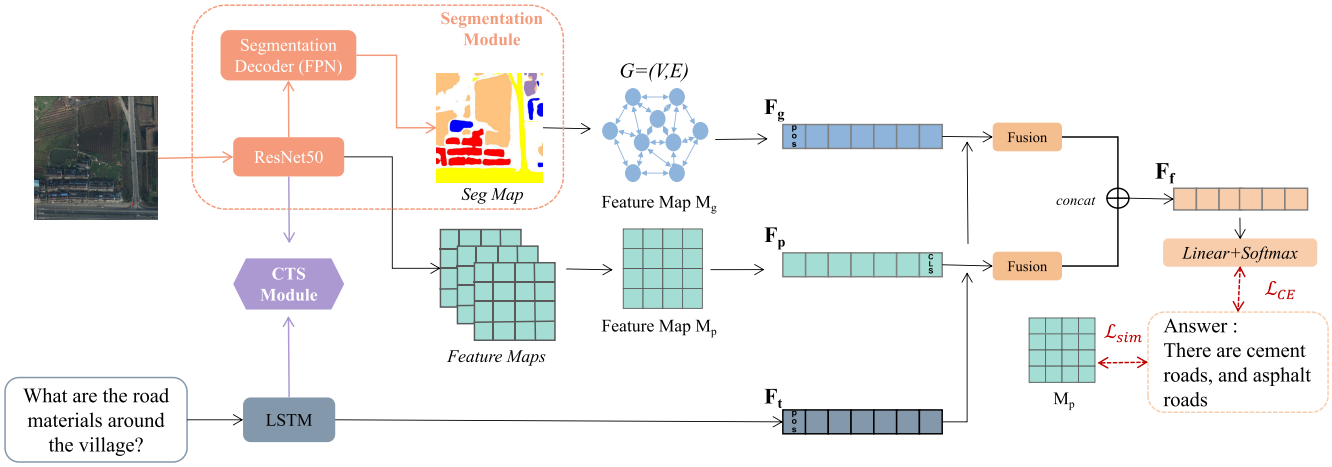

Fig. 2. Overall pipeline of our proposed method. It highlights the specific feature vectors involved in the experiment, with the segmentation module presented solely as a prior step for the VQA task. We design a dual-branch feature extraction strategy: one branch extracts pixel-level visual features, while the other builds the relationships between segmented regions using a GNN. The CTS module, serving as an auxiliary contrastive learning component, is detailed in Fig. 3. Additionally, [POS] denotes positional encoding, while [CLS] represents the category label.

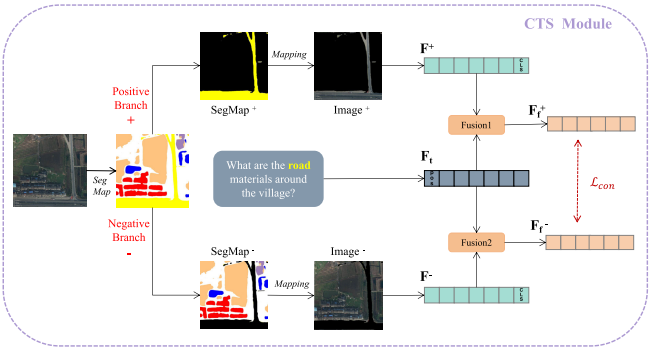

Fig. 3. Auxiliary contrastive learning module. We identify key terms in the question and map them to critical image regions, dividing the image into positive and negative samples for contrastive training.

Fig. 2, if different category labels are visualized in different colors, the segmented image can be divided into distinct regions, where each region represents a cluster of pixels belonging to the same category within a localized area. This segmentation result serves as prior knowledge to guide the VQA task, helping the model focus on key regions in the image. To obtain these segmentation results, we utilize a pretrained semantic feature pyramid network (FPN) model based on CNNs. Unlike general-purpose segmentation models such as SAM, which are designed for open-domain scenarios and may lack fine-grained understanding of domain-specific semantics, the Semantic FPN is well-suited for remote sensing imagery due to its ability to capture structured spatial hierarchies and land cover semantics. Specifically, this model adopts ResNet-50 as the encoder and an FPN as the decoder, enabling the generation of region-level segmentation maps that capture high-level semantic information from remote sensing images. These segmentation results are subsequently leveraged to enhance the reasoning capabilities of our VQA model by providing explicit structural and categorical information about different land cover types.

B. Overall Architecture

Following the prior step of performing semantic segmentation on the image based on category labels, we obtain a corresponding semantic mask that encodes category-level information. In the remote sensing VQA task addressed in this work, this mask serves as a bridge to guide the model's focus toward more critical visual information, ultimately improving the accuracy of the predicted answers. Overall, our proposed framework for remote sensing VQA consists of the mainstream answer generation module and the auxiliary contrastive learning module:

Mainstream Answer Generation Module based on Dual-branch Feature Extraction: To effectively capture both fine-grained pixel-level and high-level semantic representations in remote sensing images, we propose a dual-branch visual feature extraction strategy. One branch employs a standard ResNet-50 backbone to extract basic visual features I, resulting in pixel-level feature maps, denoted as M_p for each image. The other branch builds upon these features and incorporates a graph-structured learning module G, which is constructed based on the connectivity of segmented semantic regions. This branch generates region-level feature maps, denoted as M_g for each image, capturing structured relationships between land cover categories. The introduction of graph-based learning enables the model to enhance both intra-region coherence and inter-region interactions, overcoming the limitations of traditional pixel-wise feature learning. To further improve generalization and robustness, we apply common data augmentation techniques such as random cropping, rotation, and scaling during training. In parallel, we use LSTM to extract textual features T. The extracted visual features from both visual feature extraction branches, denoted as F_p and F_g, are then fused with the textual feature vector F_t, followed by element-wise addition to obtain the final fused feature representation F_f. This representation is subsequently mapped to the answer space, where the answer with the highest probability is selected as the output.

Auxiliary Contrastive Learning Module: Beyond the mainstream answer prediction module, we incorporate an auxiliary contrastive learning module (CTS module). Based on the semantic segmentation results, we map key terms in the question to corresponding key regions in the image. Using these key regions, we generate visual positive samples (I^+) and negative samples (I^-) by selectively masking certain regions. The feature fusion process in this module follows a similar approach to pixel-level feature extraction, resulting in fused features for positive and negative samples, denoted as F_f^+ and F_f^-, respectively. These features are only used as part of the training loss during gradient descent, effectively enhancing the model's ability to focus on question-relevant features.

The mainstream answer generation module and the auxiliary contrastive learning module work together to improve the model's reasoning and answering capabilities.

C. Multimodal Feature Fusion

1) Basic Fusion Architecture—Transformer: In our model, the fundamental feature fusion mechanism is based on the multihead attention mechanism in the Transformer framework. To make the attention formulation easier to understand, we use I to denote visual features and T to denote textual features in this section. In our method, I corresponds to different types of visual features, including pixel-level features F_p, graph features F_g, and positive/negative sample features F^+ and F^-, while T corresponds to the textual features F_t. This unified notation is used to describe the basic mechanism of visual–textual feature fusion.

For any given input feature X(which can be I or T), we first project it into query (Q), key (K), and value (V) representations through three independent learnable linear layers

$$Q = W_q X, \quad K = W_k X, \quad V = W_v X \tag{1}$$

where W_q, W_k, and W_v are weight matrices that are automatically updated and optimized through backpropagation during training. The attention mechanism then follows the standard formulation

$$\text{Attention}\,(Q, K, V) = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V \tag{2}$$

where d_K represents the dimension of K. To establish stronger relationships between image features, the model applies self-attention on the visual features I, updating them as follows:

$$I' = \text{Self-Attention}\,(I) = \text{Attention}\left(W_q I, W_k I, W_v I\right). \tag{3}$$

Since the VQA task is inherently multimodal, effective interaction between question text features T and image features I is crucial. To achieve this, our model employs bidirectional cross-attention, considering the mutual influence between vision and language. The updated representations are computed as follows:

$$I' = \text{Cross-Attention}\,(I, T) = \text{Attention}\left(W_q I, W_k T, W_v T\right) \tag{4}$$

$$T' = \text{Cross-Attention}\,(T, I) = \text{Attention}\left(W_q T, W_k I, W_v I\right). \tag{5}$$

Equation (4) represents the attention mechanism where textual information influences the image representation. By computing attention scores over the textual features, the model weights the textual value vectors accordingly, generating a text-aware visual representation I'. Conversely, (5) generates a visual-aware textual representation T' based on the image features.

Notably, visual feature representation in our model includes pixel-level features F_p, graph-structured features F_g, positive sample features F_f^+, and negative sample features F_f^-, and the fusion of each of them with textual features is consistently performed using multihead attention. The only exception is the fusion of graph-structured features, where the graph attention network (GAT) additionally considers edge features. For clarity, I and T are used as general representations of image and text features in this section. I corresponds to different types of visual features, including F_p, F_g, F^+, and F^-, while T corresponds to the textual feature F_t. This notation is used to describe the basic mechanism of visual–textual feature fusion.

2) Graph Structure Incorporation: Although pixel-level features F_p capture most of the detailed information from the entire image, they tend to be overly generalized and redundant for remote sensing images. Furthermore, these features often contain a significant amount of irrelevant noise, which can negatively impact the final answer prediction. Given that VQA questions primarily focus on specific regions of an image or its overall content, we introduce a graph structure $G = (V, E)$ based on the semantic segmentation results. This allows the model to emphasize key regions from a macroscopic perspective.

The nodes V in the graph are derived from the connected component regions obtained through semantic segmentation. In the segmentation map, each pixel is assigned a class label l, where $l \in [1, \text{num-class}]$, and num-class represents the total number of distinct categories. Pixels with the same class label that are spatially adjacent are grouped into clusters. The number of clusters corresponds to the number of connected components in the pixel-level segmentation mask. Each cluster is then converted into a graph node, with its feature representation computed as the average feature of all pixels within the cluster

$$V_{f_k} = \text{Average-Pool}\,(\text{Cluster}_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} f_{k_i} \tag{6}$$

where n_k denotes the number of pixels in the kth cluster, and $\sum_{k=1}^{\text{num-clusters}} \sum_{i=1}^{n_k} = M_p$. Since all pixels within a cluster share the same semantic category, we append a class label identifier $[l]$ to the last dimension of V_{f_k}, computed as

$$[l] = \frac{l}{\text{num-class}} + 10^{-8}. \tag{7}$$

The edges E of the graph should effectively capture the relationships between different nodes. The edge feature $E_{f_{ij}}$ between two nodes is computed as the dot product of their feature vectors

$$E_{f_{ij}} = V_{f_i} \cdot V_{f_j}^T. \tag{8}$$

This formulation enables the construction of a fully connected graph structure G, producing the region-level feature representation M_g. To model relationships within the graph, we employ a GAT that extends the multihead self-attention mechanism in Transformers by incorporating edge features. Specifically, the edge feature $E_{f_{ij}}$ serves as the basis for

attention between nodes. The updated node feature V'_{f_i} is obtained by a weighted sum of neighboring node features, where the weights are normalized using a softmax over $E_{f_{ij}}$

$$V'_{f_i} = \sum_{j \in N(i)} \text{softmax}_j E_{f_{ij}} V_{f_j} \tag{9}$$

where $N(i)$ denotes the neighboring nodes of node i. This formulation allows the GAT to effectively capture both intra-region and inter-region dependencies while leveraging the semantic relations encoded in the edge features, enhancing the visual representation M_g. During cross-modal interaction, we apply the same bidirectional cross-attention mechanism used previously, enabling the exchange of information between the graph-structured visual features and textual features. This results in updated feature representations F'_g and F''_t.

For the pixel-level feature branch, F_p is directly fused with textual features F_t to obtain F'_p and F'_t. To unify the updated feature representations, we perform layer normalization on the aggregated features

$$F_f = \text{LayerNorm}\left(F'_g + F''_t + F'_p + F'_t\right). \tag{10}$$

Finally, the fused feature representation F_f is projected into the answer space, and the answer with the highest predicted probability is selected as the final response.

D. Contrastive Learning Module

In the VQA task, the reasoning process involves establishing a balanced relationship between the image, question, and answer. Specifically, in remote sensing VQA datasets, questions often pertain to specific regions within an image. Therefore, selectively focusing on these relevant regions can significantly enhance the accuracy of the answers. Based on semantic segmentation results, the key terms in a question can be mapped to corresponding category regions within the image. For each training sample, the entire image can be divided into key regions and nonkey regions based on the alignment between the regional categories and the question keywords. These regions are then designated as the visual positive sample I^+, and the visual negative sample I^-, respectively, and their relationship with the full image can be expressed as $I = I^+ + I^-$.

Beyond the standard answer-generation module, we introduce an additional contrastive learning module to ensure a more effective learning process for question-relevant visual features. In this module, we compute the fused representations of the positive and negative samples with textual features, separately denoted as $F_f^+ = \text{Fusion}(I^+, T)$ and $F_f^- = \text{Fusion}(I^-, T)$. The fusion process follows the same approach as that used for pixel-level features F_p.

To enhance learning from the positive samples, i.e., emphasizing the positive influence of question-relevant regions while reducing the negative impact of irrelevant regions, we incorporate a contrastive loss function to guide gradient updates. Before defining this loss, we introduce a scoring function s to measure the similarity between positive/negative samples and the full image representation. A higher score indicates a greater contribution to the overall representation, distinguishing positive from negative samples. To ensure proper normalization, we use the cosine similarity between the multimodal fusion representations as the scoring function

$$\begin{aligned} s_p &= \text{Score}\left(F_f, F_f^+\right) = \frac{F_f \cdot F_f^{+T}}{\|F_f\|\|F_f^+\|} \\ s_n &= \text{Score}\left(F_f, F_f^-\right) = \frac{F_f \cdot F_f^{-T}}{\|F_f\|\|F_f^-\|}. \end{aligned} \tag{11}$$

Following contrastive learning principles from unsupervised learning, we define the contrastive loss function as

$$\mathcal{L}_{\text{con}} = E_{F_f, F_f^+, F_f^-}\left[-\log\left(\frac{e^{s_p}}{e^{s_p} + e^{s_n}}\right)\right]. \tag{12}$$

This loss encourages the model to maximize the similarity between F_f^+ and the overall fused representation F_f, while simultaneously minimizing the similarity between F_f^- and F_f. By doing so, the model learns to prioritize question-relevant visual regions, improving the quality of the generated answers.

E. Loss Function

1) Cross-Entropy (CE) Loss: Since the VQA task is a multiclass classification problem in which the model selects the most probable answer from a predefined answer set, CE loss is employed as the primary objective function for optimization. Given the multimodal fused representation F_f, the predicted answer y' is obtained through a linear projection followed by a softmax activation. The CE loss between the predicted answer y' and the ground-truth answer y is defined as

$$\mathcal{L}_{\text{CE}} = \text{CE}\left(y, y'\right) = -y \odot \log\left(y'\right) = -\sum_{i=1}^{N} y_i \log\left(y'_i\right) \tag{13}$$

where N denotes the number of answers in the answer set.

2) Contrastive Loss: As described in Section III-D, the contrastive learning process separates the image into positive and negative samples based on their relevance to the question. By minimizing the contrastive loss $\mathcal{L}_{\text{con}}$, the model is encouraged to focus more on question-relevant visual regions while reducing the influence of irrelevant regions during training.

3) Text-Visual Similarity Loss: The VQA dataset contains a mix of yes/no (Y/N), counting, and reasoning questions. Given the complexity of textual answers in reasoning-type questions, enhancing the similarity between the selected answer's textual features and the corresponding image features can significantly improve the accuracy for these questions. To achieve this, we first align the model-predicted answer y' with its corresponding textual content and extract answer text features T_a using an LSTM. To measure the similarity between the answer text features and image features, we adopt cosine similarity as the measurement method. The objective is to maximize the cosine similarity between T_a and the image representation I for reasoning-based questions. Thus, the text-visual similarity loss is defined as

$$\mathcal{L}_{\text{sim}} = \text{Cosine-Similarity}\left(T_a, I\right) = \frac{T_a \cdot I^T}{\|T_a\|\|I\|}. \tag{14}$$

By combining all sublosses with assigned weights, the total training loss is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{con}} + \omega\mathcal{L}_{\text{sim}} \tag{15}$$

where α and ω are hyperparameters that control the relative contributions of contrastive loss and similarity loss. It is important to note that since the primary objective of our model is answer prediction, $\mathcal{L}_{\text{CE}}$ serves as the main loss, while $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{sim}}$ act as auxiliary losses to enhance model learning.

IV. Experiment

A. Dataset Descriptions

To evaluate the effectiveness of the proposed model, we conduct experiments on two remote sensing VQA datasets that provide corresponding semantic segmentation masks.

1) EarthVQA: The EarthVQA dataset [16] is an extension of the LoveDA dataset [52], which was originally designed for land cover segmentation. It consists of 6000 high-spatial-resolution (HSR) images collected from 18 urban and rural administrative regions, including Wuhan, Nanjing, and Changzhou. Each image is paired with a semantic segmentation mask that annotates eight land cover categories. In addition to the visual data, the dataset contains 208 593 question-answer (QA) pairs, which are categorized into six VQA task types: basic judging (Basic-J), relational-based judging (Relation-J), basic counting (Basic-C), relational-based counting (Relation-C), object situation analysis (Object-A), and comprehensive analysis (Comp-A). Specifically, 2522 images and their corresponding QA pairs are used for training, 1669 images for validation, and 1809 images for testing.

Since the ground-truth answers for the test set are not publicly available, model performance on the test set is evaluated by generating answer files in ".json" format and submitting them to the Codabench platform for standardized assessment.

2) RescueNet-VQA: The RescueNet-VQA dataset [19] is derived from RescueNet [35], a dataset designed for post-disaster damage assessment. It consists of 4375 high-resolution images captured following Hurricane Michael between October 11 and October 14, 2018, using UAV-based aerial imaging over affected regions such as Mexico Beach, FL, USA. Each image is accompanied by a semantic segmentation mask that delineates 11 land cover categories. For the VQA task, the dataset includes 103 192 QA pairs, which have been refined in recent updates. The QA pairs are categorized into eight distinct VQA types: simple counting (SC), complex counting (CC), building condition recognition (BCR), road condition recognition (RCR), level of damage (LOD), risk assessment (RA), density estimation (DE), and area-based (AB).

Model training is conducted on the training set, while performance evaluation is performed on the test set to measure the effectiveness of the proposed approach.

3) FloodNet-VQA: The FloodNet-VQA dataset [53] is derived from the FloodNet dataset, which was collected using UAV-based aerial imagery after Hurricane Harvey in Texas. It consists of 3200 high-resolution images (4000×3000) captured at low altitude, each annotated with semantic segmentation masks covering nine land cover categories. For the VQA task, the dataset provides approximately 11 000 manually generated QA pairs, averaging 3.5 questions per image. The QA pairs are grouped into four categories: SC, CC, condition recognition (CR), and Y/N.

Since the ground-truth answers of the validation and test sets are not publicly available, we split the original training set by using 80% of the samples for training and the remaining 20% for evaluating the effectiveness of our model.

B. Implementation Details

Our model is implemented using Python 3.8 and deployed on the PyTorch 1.13.0 framework. For image features, input images undergo mathematical feature augmentation, and feature extraction is performed using a two-layer ResNet-50 backbone. For text features, we employ a two-layer LSTM to encode the input question representations. In the feature fusion stage, we set the number of self-attention layers and bidirectional cross-attention layers to 3. The number of attention heads is set to 8, and the hidden size for both visual and textual features is fixed at $d = 384$.

We utilize the AdamW optimizer with the default momentum coefficients $\beta = (0.9, 0.999)$ and a weight decay factor of 0.05. The learning rate is adjusted using a "poly" learning rate decay strategy, where the initial learning rate is set to 5e−5, and it follows an exponential decay with power = 0.9. Performance is evaluated on the validation set at the end of each epoch to adjust and optimize hyperparameters. For loss weight hyperparameters, we set $\alpha = 0.1$ and $\gamma = 0.1$ in our experiments. Training is conducted with a batch size of 16 for a total of 48 000 iterations. All experiments are performed on a single RTX 3090 GPU with 24 GB of VRAM.

C. Performance Comparison and Analysis

In this section, we present a quantitative analysis of our proposed model's performance on EarthVQA [16], RescueNet-VQA [19], and FloodNet-VQA [53] datasets, as shown in Tables I–III. In the visualization analysis section, we will present qualitative comparison results, as shown in Fig. 4. We compare our model with several mainstream natural-domain VQA methods and remote sensing VQA methods, as detailed below.

1) *General VQA Methods:* We compare our model with several classic VQA architectures that have evolved over time, including SAN [26], BUTD [9], MCAN [12], and D-VQA [28]. Additionally, we include LXMERT [30], which leverages pretrained multimodal representations, as well as BLIP-2 [31] and Instruct-BLIP [32], which are fine-tuned LLMs for vision-language tasks.
2) *RSVQA [13]:* A pioneering foundational model that simply concatenates visual and textual features to predict answers.
3) *MAIN [43]:* A model that enhances image–text alignment through attention mechanisms and bilinear techniques.
4) *SOBA [16]:* The baseline model for EarthVQA, which incorporates additional learning of semantic information from images.

The performance comparison on the EarthVQA dataset is presented in Table I. Our model significantly outperforms all general VQA and remote sensing VQA methods, achieving

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE EARTHVQA TEST SET. (*) DENOTES THE LIGHTWEIGHT VERSION. WE HIGHLIGHT THE **BEST** RESULT IN EACH COLUMN

Method	Category(Accuracy)(%)						AA(%)	OA(%)	Params(M)
	Basic-J	Relation-J	Basic-C	Relation-C	Object-A	Comp-A			
General VQA Methods:									
SAN [26]	88.73	80.61	77.16	64.98	52.73	42.70	67.82	75.52	32.30
BUTD [9]	90.01	82.02	77.16	60.95	56.29	42.29	68.12	76.49	34.95
MCAN [12]	89.65	81.65	79.83	63.16	57.28	43.71	69.21	77.07	55.17
D-VQA [28]	89.73	82.12	77.38	63.99	55.14	43.20	68.59	76.59	37.79
LXMERT [30]	90.00	81.93	**80.54**	68.26	54.87	45.07	70.11	77.46	224.12
BLIP-2 [31]	88.13	81.92	70.26	58.58	42.72	28.34	61.66	71.07	$\approx 4B$
Instruct-BLIP [32]	89.67	79.69	76.96	63.34	59.72	45.68	69.18	75.25	$\approx 4B$
Remote Sensing VQA Methods:									
RSVQA [13]	82.43	79.34	70.68	55.53	42.45	35.46	60.98	70.70	30.21
MAIN [43]	85.32	80.44	75.01	56.63	51.55	39.25	64.70	73.71	41.41
SOBA [16]	89.63	82.64	80.17	67.86	61.40	49.30	71.83	78.14	40.46
ACR [54]	89.80	83.60	79.70	68.00	61.60	49.20	71.98	78.50	47.90
Ours*	90.03	85.68	77.60	68.94	60.97	53.23	72.74	79.38	55.69
Ours	**90.37**	**87.30**	77.87	**68.92**	**61.89**	**55.30**	**73.61**	**80.45**	63.68

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE RESCUENET-VQA TEST SET. WE HIGHLIGHT THE **BEST** IN EACH COLUMN

Method	Category(Accuracy)(%)								AA(%)	OA(%)
	SC	CC	BCR	RCR	LOD	RA	DE	AB		
MCAN [12]	43.55	43.21	79.24	86.74	75.00	84.81	86.08	40.00	67.33	69.43
RSVQA [13]	44.35	41.79	76.96	83.15	77.33	79.75	81.01	40.00	65.54	68.01
MAIN [43]	48.39	42.14	73.92	84.59	79.33	79.75	84.81	37.50	66.30	68.32
SAM-VQA [55]	37.10	56.43	95.69	92.83	76.33	74.68	81.01	66.25	72.54	77.10
SOBA [16]	51.61	64.29	**99.49**	91.39	77.00	82.28	83.54	53.75	75.42	80.26
Ours	**55.65**	**71.43**	99.24	**94.62**	**78.67**	**87.34**	**89.87**	**66.25**	**80.38**	**83.79**

the highest average accuracy (AA) of 73.61% and overall accuracy (OA) of 80.45%. Compared to the baseline model SOBA, our model demonstrates a 1.78% improvement in AA and a 2.31% improvement in OA. Compared with the latest adaptive conditional reasoning (ACR) [54], which adaptively selects different reasoning paths based on the semantic type of the question, our model still holds significant advantages. Furthermore, on the Codabench platform, which serves as the official benchmark for EarthVQA test set evaluations, our model outperforms all existing methods, ranking first and achieving SOTA performance. When analyzing performance across different question categories, we observe that our model achieves the highest accuracy in all question types except Basic Counting. The lower accuracy in this category may be attributed to the nature of these questions, which focus on the distribution of specific land cover categories across an entire image. In such cases, methods that rely solely on global pixel-level feature learning may be more suitable. Beyond accuracy metrics, we also compare model parameter sizes. As shown in the last column of Table I, our model achieves competitive accuracy while maintaining a parameter size comparable to other methods. Notably, when compared to approaches that rely on pretrained models or LLMs, our model demonstrates a significant advantage in parameter efficiency. In addition, we added comparison experiments using a lightweight version of our method, as indicated by the row marked with (*) in Table I. Since the graph structure introduces considerable computation overhead, it is removed in the lightweight version, leading to a similar parameter size to MCAN [12]. However, our lightweight model still exhibits a clear advantage in VQA accuracy.

TABLE III

COMPARISON WITH STATE-OF-THE-ART REMOTE SENSING METHODS ON THE FLOODNET-VQA TEST SET. WE HIGHLIGHT THE **BEST** IN EACH COLUMN

Method	Category(Accuracy)(%)				AA(%)	OA(%)
	SC	CC	CR	Y/N		
MCAN [12]	18.18	14.63	92.17	89.36	53.59	67.92
RSVQA [13]	17.69	12.20	91.96	85.11	51.74	66.41
MAIN [43]	15.91	21.95	90.98	94.85	55.92	68.68
SOBA [16]	27.03	**29.27**	90.23	93.62	60.04	70.94
Ours	**27.27**	26.83	**92.48**	**95.74**	**60.58**	**72.08**

To further validate the effectiveness of our model, we conduct additional experiments on the RescueNet-VQA dataset, with results summarized in Table II. Although both EarthVQA and RescueNet-VQA belong to the domain of remote sensing VQA, disaster assessment tasks often incorporate specialized knowledge related to damage evaluation. Therefore, we introduce SAM-VQA [55], a model specifically designed for disaster damage assessment, as an additional comparison method. As shown in Table II, our model achieves the highest VQA accuracy, with AA reaching 80.38% and OA achieving 83.79%, maintaining a substantial margin of 4.96% and 3.53% over the second-best method, respectively. When broken down by question category, our model consistently achieves the highest accuracy in all question types, except for a minor 0.25% gap in BCR questions. In addition, we further

TABLE IV
ABLATION STUDY OF DIFFERENT MODEL COMPONENTS ON THE EARTHVQA TEST SET

Method	Category(Accuracy)(%)						OA(%)
	Basic-J	Relation-J	Basic-C	Relation-C	Object-A	Comp-A	
w/o GNN	90.03	85.68	77.60	68.94	60.97	53.23	79.38
w/o CON	89.40	84.37	76.08	65.27	**62.15**	51.58	78.20
w/o MHATT	79.71	71.68	65.02	62.67	44.49	45.07	67.16
w/o SIM	90.06	85.98	77.73	68.14	63.02	54.68	79.79
All (Ours)	**90.37**	**87.30**	**77.87**	**68.92**	61.89	**55.30**	**80.45**

include comparative experiments on another disaster dataset, FloodNet-VQA, as presented in Table III. It can be observed that our model achieves the highest accuracy across nearly all question types compared with mainstream remote sensing VQA methods, thereby demonstrating the effectiveness of the proposed approach.

Overall, the experimental results on both datasets strongly confirm the effectiveness of our proposed model and further demonstrate the robustness and generalization capability of the model in remote sensing VQA tasks.

D. Ablation Studies

1) Effectiveness of Model Components: Our model follows the conventional framework of VQA, where visual and textual features are extracted separately before being fused to predict the final answer. To enhance performance, we introduce four key components: 1) a graph-based visual feature extraction module (GNN); 2) a contrastive learning module (CON); 3) a multihead attention mechanism within the feature fusion stage (MHATT); and 4) a similarity-based mechanism for answer generation in reasoning-type questions (SIM). The GNN enables the model to capture high-level semantic information about land cover categories, the CON enhances the model's ability to focus on question-relevant image regions, the MHATT facilitates sufficient interaction between visual and textual information during feature fusion, and the SIM improves the alignment between reasoning-based answers and image content.

To assess the effectiveness of these components, we conduct ablation experiments by removing each module separately and comparing the model's performance under these conditions. Table IV presents the results, where the first four rows represent models without GNN, CON, MHATT, and SIM, respectively, while the last row corresponds to the complete model. The OA results indicate that the removal of each component leads to a performance drop: GNN (−1.07%), CON (−2.25%), MHATT (−13.29%), and SIM (−0.66%). These results confirm that each module plays a crucial role in the model's overall effectiveness, and all components are indispensable.

2) Analysis of Encoders in Semantic Segmentation: Since our model is designed based on semantic segmentation results to enhance the learning of more effective visual features and ultimately improve remote sensing VQA performance, we further analyze the segmentation encoder's impact on the model. A semantic segmentation model typically consists of an encoder–decoder structure, where the encoder extracts semantic features from the input image, while the decoder maps these features back to pixel-level spatial resolution. Given that our remote sensing VQA framework includes a visual feature extraction process, the role of the segmentation encoder aligns with that of the feature extraction module in the VQA pipeline. Therefore, evaluating different encoder architectures provides insights into their effectiveness for VQA.

In our experiments, we consider two categories of encoders: CNN-based and Transformer-based architectures. For CNN-based encoders, we evaluate ResNet [36] and EfficientNet [56], while for Transformer-based encoders, we assess Vision Transformer [57] and Swin Transformer [39]. Each segmentation model is trained end-to-end along with the VQA model, and the VQA accuracy is compared across different architectures. As shown in Table V, the first four rows present the OA results under different encoder settings. The results indicate that CNN-based encoders are more effective in capturing critical semantic features from remote sensing images, with ResNet demonstrating particularly strong feature extraction capabilities.

To further investigate the impact of the decoder, we conduct experiments using ResNet as the encoder and compare two decoder variants: FPN [58] and UNet [59]. Since our primary objective is to improve VQA performance rather than segmentation accuracy, we analyze whether the choice of decoder affects the final QA results. The results in Table V show that different decoders have little effect on VQA accuracy. This suggests that semantic feature extraction is the key factor influencing VQA performance, but the segmentation mask primarily affects attention on different visual regions rather than fundamentally altering the extracted features. Based on the same encoder–decoder configurations as Table V for VQA accuracy, we further report the corresponding semantic segmentation results in Table VI. From the mIoU metric, segmentation performance shows a generally positive correlation with VQA accuracy across different encoders and decoders. However, the performance gaps among different configurations remain relatively small. This observation is consistent with the core motivation of our work: rather than relying on marginal improvements in segmentation precision, we focus on leveraging semantic segmentation results to guide the design of the VQA architecture, enabling more targeted visual feature learning. Therefore, we focus primarily on analyzing the visual feature extraction network rather than the decoder structure.

3) Analysis of Edge Construction in GNN: Our model incorporates a graph-based visual feature extraction module to capture more precise land cover semantic information. In constructing the graph, as formulated in (8), we define nodes based on regional average feature vectors and establish edges using the dot product of feature vectors. To assess the impact of different edge construction methods, we compare our approach, denoted as dot product-based edge construction (MUL), with two alternative strategies. The first alternative, denoted as distance-based edge construction (DIS), constructs edges based solely on the Euclidean distance between region center points. In this setting, the influence between nodes in the graph attention mechanism is inversely proportional to their distance. The second approach, denoted as Gaussian kernel similarity-based edge construction (EUC), constructs edges

TABLE V
ABLATION STUDY OF DIFFERENT ENCODERS IN SEMANTIC SEGMENTATION ON THE EARTHVQA TEST SET

Encoder		Decoder	Category(Accuracy)(%)						OA(%)
CNN-based	Transformer-based		Basic-J	Relation-J	Basic-C	Relation-C	Object-A	Comp-A	
✓(ResNet)	-	FPN	88.39	84.08	76.53	59.62	55.14	48.90	77.12
✓(EfficientNet)	-	FPN	87.25	84.05	76.81	61.95	55.17	49.05	77.02
-	✓(Vision Transformer)	FPN	87.70	83.24	76.27	60.07	52.68	**51.07**	76.72
-	✓(Swin Transformer)	FPN	87.53	**84.62**	74.35	62.11	**57.74**	48.76	76.93
✓(ResNet)	-	UNet	**89.82**	83.76	**76.85**	**66.10**	55.80	45.30	**77.15**

TABLE VI
COMPARISON OF SEMANTIC SEGMENTATION RESULTS BROUGHT BY DIFFERENT ENCODERS AND DECODERS (THE SAME AS TABLE V) ON THE EARTHVQA VALIDATION SET

Encoder		Decoder	OA(%)	mIOU(%)	Precision(%)	Recall(%)	F1-score(%)
CNN-based	Transformer-based						
✓(ResNet)	-	FPN	77.12	47.92	63.61	67.21	62.76
✓(EfficientNet)	-	FPN	77.02	47.69	59.23	64.10	64.58
-	✓(Vision Transformer)	FPN	76.72	47.29	62.10	65.66	63.32
-	✓(Swin Transformer)	FPN	76.93	46.81	61.29	64.63	61.72
✓(ResNet)	-	UNet	77.15	47.98	64.44	67.44	63.28

TABLE VII
ABLATION STUDY OF DIFFERENT EDGE CONSTRUCTION METHODS ON THE EARTHVQA TEST SET

Method	Category(Accuracy)(%)						OA(%)
	Basic-J	Relation-J	Basic-C	Relation-C	Object-A	Comp-A	
DIS	89.46	86.78	77.21	68.42	**63.60**	53.95	79.85
EUC	90.26	86.76	77.44	69.47	60.68	54.18	79.94
MUL (Ours)	**90.37**	**87.30**	**77.87**	**68.92**	61.89	**55.30**	**80.45**

TABLE VIII
COMPARISON OF FOURFOLD CROSS-VALIDATION ON OA METRIC ON THE EARTHVQA TEST SET. THE TRAINING AND VALIDATION SETS ARE DIVIDED INTO FOUR SUBSETS (P1, P2, P3, P4). M1, M2, M3, AND M4 MEANS CHOOSING P1, P2, P3, AND P4, RESPECTIVELY, AS THE VALIDATION SET AND THE REMAINING SUBSETS AS THE TRAINING SETS

Method	M1	M2	M3	M4	Overall
SOBA [16]	77.30	77.14	77.06	77.09	77.15
Ours	77.40	77.25	77.28	77.33	77.32

using a Gaussian kernel similarity function, formulated as

$$E_{f_{ij}} = \exp\left(-\frac{\left\|V_{f_i} - V_{f_j}\right\|^2}{2\delta^2}\right) \tag{16}$$

where smaller distances yield higher similarity values, indicating stronger interactions between nodes.

Table VII presents the OA results for the VQA task under different edge definitions. The results demonstrate that using the vector dot product as the edge feature yields the highest accuracy, suggesting that this method better captures semantic relationships between image regions compared to distance-based and Gaussian kernel-based approaches.

V. HYPERPARAMETER ANALYSIS

During the training process of our model, the total loss function consists of three sublosses: the classification loss for the primary VQA task, the contrastive loss maximizing the distinction between positive and negative samples in contrastive learning, and the similarity loss that enforces alignment between textual answers and image features in reasoning-type questions. Since the classification loss serves as the primary objective in VQA, it dominates the overall training process, while the other two losses act as auxiliary losses. To control their contributions, we introduce weighting hyperparameters α and ω ($\alpha, \omega \in (0, 1)$).

Experimental results indicate that assigning relatively small values to the auxiliary loss weights leads to better performance. To further investigate their impact, we conduct experiments by fixing one weight at 0.1 while varying the other among 0.1, 0.3, 0.5, and we evaluate the resulting VQA accuracy. The obtained results, illustrated in the line chart in Fig. 5, show that the highest OA is achieved when both α and ω are set to 0.1. This finding confirms that the auxiliary losses primarily serve as supplementary objectives to facilitate model optimization, rather than acting as dominant loss components during training.

VI. DISCUSSION

A. Cross Validation on Different Dataset Splits

To further demonstrate the reliability and robustness of the performance gains achieved by our approach, we conduct K-fold cross-validation (with $K = 4$ in this experiment) on different dataset split settings. Specifically, the EarthVQA training and validation sets, comprising a total of 4191 images and their corresponding question–answer pairs, are divided into four subsets (1000, 1000, 1000, and 1191 samples). In each run, one subset is used for validation while the remaining are used for training. As reported in Table VIII, the cross-validation results on the EarthVQA test set show that our segmentation-guided model consistently outperforms the baseline SOBA model [16], confirming the effectiveness of our strategy.

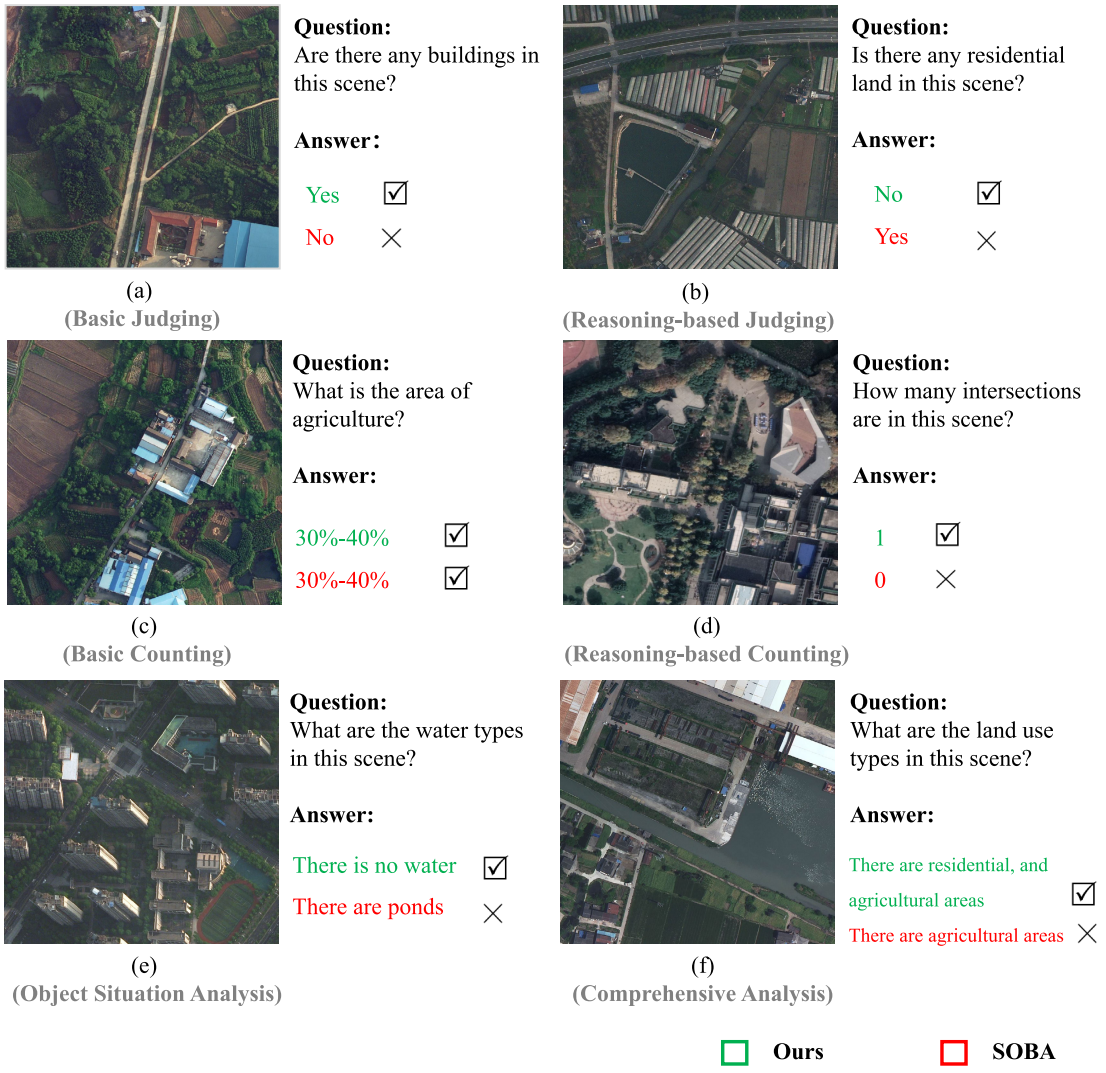

Fig. 4. Qualitative comparison between our model and the baseline method SOBA on the EarthVQA dataset. (a)–(f) Performance on six different question categories across various scenarios.

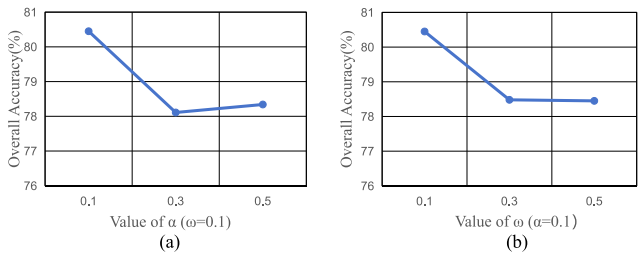

Fig. 5. Line charts illustrating the OA of the task under different hyperparameter settings for α and ω in the loss function. (a) Effect of adjusting α with ω fixed at 0.1. (b) Effect of adjusting ω with α fixed at 0.1.

B. Evaluation of the Text Encoder in the Model

To evaluate the effect of the text encoder in our model, we attempt to replace it with Transformer-based language models and adjust the parameters accordingly to ensure fair comparison. We adopt a Transformer trained end-to-end from scratch as a text feature extractor. Furthermore, we compare pretrained BERT [60], specifically using the *bert-base-uncased* model. As shown in Table IX, our approach achieves higher VQA accuracy than the Transformer trained from scratch and achieves slightly lower OA than the pretrained BERT, while still demonstrating superior performance on almost all question types except the counting category (Basic-C, Relation-C). Nevertheless, the performance gain of *bert-base-uncased* model remains relatively modest, as reflected in Table IX. This is mainly because Transformer-based language models are pretrained on general-domain corpora, whose linguistic knowledge may not fully align with the remote sensing field. Moreover, remote sensing VQA questions are more constrained with a limited vocabulary, compared to open-domain VQA tasks (e.g., VQA v2.0 [21]).

C. Evaluation of the Model at Different Resolution Ratios

Considering that remote sensing images are often acquired at different resolution levels, we conducted experiments to evaluate the adaptability of our model to varying image resolutions. Specifically, we rescaled the test images of EarthVQA

TABLE IX

COMPARISON OF ACCURACY ON THE EARTHVQA TEST SET WITH DIFFERENT TEXT ENCODERS

Text Encoder	Category(Accuracy)(%)						OA(%)
	Basic-J	Relation-J	Basic-C	Relation-C	Object-A	Comp-A	
Transformer	89.95	86.30	**79.69**	68.64	**62.56**	54.51	80.26
bert-base-uncased [60]	90.26	87.25	78.44	**69.08**	61.84	54.90	**80.48**
LSTM(Ours)	**90.37**	**87.30**	77.87	68.92	61.89	**55.30**	80.45

TABLE X

COMPARISON OF ACCURACY ON THE EARTHVQA TEST SET WITH DIFFERENT RESOLUTIONS

Test Resolution	Category(Accuracy)(%)						OA(%)
	Basic-J	Relation-J	Basic-C	Relation-C	Object-A	Comp-A	
512×512	90.07	85.27	77.64	62.94	59.76	50.91	78.74
2048×2048	89.54	86.80	76.51	69.08	58.57	52.43	79.30
1024×1024(Ours)	90.37	87.30	77.87	68.92	61.89	55.30	80.45

TABLE XI

COMPARISON OF ACCURACY ON THE EARTHVQA VALIDATION SET RELATED ON SEGMENTATION MASKS

Mask Settings	Category(Accuracy)(%)						OA(%)
	Basic-J	Relation-J	Basic-C	Relation-C	Object-A	Comp-A	
w/o masks	80.05	80.27	76.45	79.63	67.05	43.69	74.46
Ours	79.20	82.59	77.46	73.46	65.23	44.50	75.23
w GT masks	88.27	78.70	77.41	78.97	67.09	46.32	76.00

while keeping the aspect ratio unchanged and measured the VQA accuracy at different resolutions, as shown in Table X. Our model was trained on images with a resolution of 1024×1024, while the test images were resized to 512×512 and 2048×2048. The results demonstrate that the highest accuracy is achieved when the training and testing resolutions are consistent. Although the performance slightly decreases at other resolutions, the model still maintains relatively high accuracy, indicating good adaptability across different resolutions.

D. Analysis of Segmentation's Contribution to VQA

To evaluate the performance gain brought by semantic segmentation and to explore the upper bound of such improvements, we designed a set of experiments. Since the segmentation masks of the EarthVQA test set are not available, we conducted comparisons on the EarthVQA validation set under three settings: 1) our model without considering semantic segmentation and its related module design; 2) our model; and 3) our model using ground-truth masks for module computation on the validation set. As shown in Table XI, incorporating segmentation-based mask designs into the VQA module leads to performance improvements, while employing ground-truth masks further demonstrates the potential upper bound of such enhancements.

E. Comparison With Large Vision-Language Models

Considering that LLM-based VQA has become a mainstream trend, we included comparisons between our method and LLaVA-Next [61] and GeoChat [62] for a fair comparison, as reported in Table XII. The models of LLaVA-Next and GeoChat exhibit relatively poor classification performance on the test set. The reason could be that they are fundamentally generative models that produce open-vocabulary answers,

TABLE XII

COMPARISON WITH LLM-BASED METHODS ON THE EARTHVQA TEST SET

Method	Category(Accuracy)(%)						OA(%)
	Basic-J	Relation-J	Basic-C	Relation-C	Object-A	Comp-A	
LLaVA-Next [61]	52.76	61.49	42.07	31.64	9.64	2.89	45.45
GeoChat [62]	62.05	45.40	69.87	35.01	14.09	2.59	46.72
Ours	90.37	87.30	77.87	68.92	61.89	55.30	80.45

which differ from the closed-set classification-based VQA task studied in this article. This observation further highlights that, in specialized and high-certainty VQA scenarios, task-specific lightweight closed-set models still possess irreplaceable advantages over general-purpose generative LLMs.

VII. VISUALIZATION ANALYSIS

A. Visualization of Model Results

To provide a qualitative comparison, we evaluate our proposed model against SOBA, the current SOTA baseline model on the EarthVQA dataset. As shown in Fig. 4(a)–(f), we present six representative remote sensing scenes, each corresponding to a different question type. Since the ground-truth answers for the test set of EarthVQA are not publicly available, we employ a manual evaluation approach to assess answer correctness. Specifically, we recruited five volunteers to independently judge the two models' answers. The answer is marked as correct only if all five evaluators agree on its correctness. From the sample cases, we observe that our model consistently provides correct answers across all six scenarios, whereas SOBA exhibits errors in most of the cases. This demonstrates the superior adaptability of our model to the remote sensing VQA task, highlighting its effectiveness in understanding and reasoning over aerial imagery.

B. Validation of the Consistency Between the Answer and the Attention Region

In our model, the final predicted answer probability distribution is derived from the aligned and fused visual–textual features, where the answer with the highest probability corresponds to the strongest semantic association between certain image regions and the question words. As illustrated in Fig. 6, for each test sample we output the top-5 answers along with their corresponding confidence scores. Different candidate answers, obtained from the fused image–question features, often attend to distinct regions in the image. We highlight the top-3 answers with their associated attention regions, where the thickness of the bounding boxes decreases with the corresponding scores. Besides, we also present attention heatmaps corresponding to image–question pairs. From scenes Fig. 6(a)–(c), it can be observed that the most attended region, corresponding to the reddest area in the attention heatmaps, consistently matches the highest-probability answer, which aligns with the ground truth in the case of correct predictions.

C. Validation of Correlation With Semantic Segmentation

Our model is designed to leverage semantic segmentation results to extract more relevant and accurate high-level

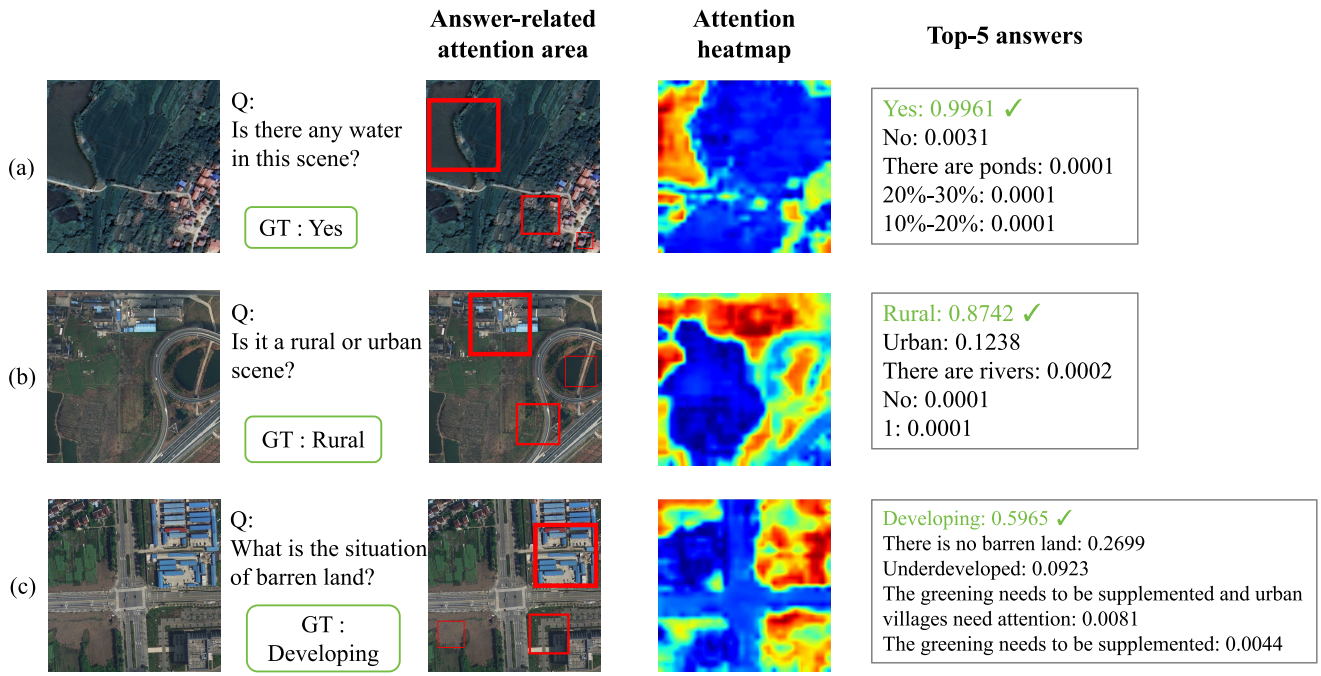

Fig. 6. Visualization of the answer-related attention area and the attention heatmaps in the remote sensing image. We output top-5 answers along with their confidence scores and highlight top-3 answers with associated attention regions, where the thickness of the bounding boxes decreases with the corresponding scores. (a)–(c) Illustrate the performance on three different scenarios.

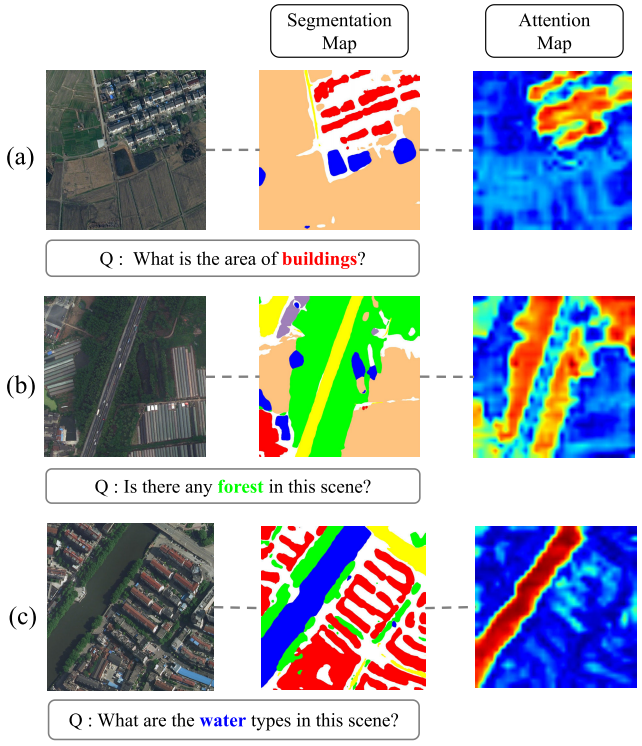

Fig. 7. Visualization of the original image, segmentation map, and attention map during the experiment. The results validate that the high-attention regions in the image correspond closely to the segmented category regions and align with the category references in the question keywords. (a)–(c) Illustrate three representative remote sensing scenes along with corresponding questions.

semantic information that captures category-specific features. Consequently, the visual feature extraction process in the VQA task is inherently correlated with the prior semantic segmentation process. As illustrated in Fig. 7, we present three representative remote sensing scenes along with corresponding questions. For each scene, we provide its semantic segmentation map and the attention heatmap generated by our trained model. In the heatmaps, redder regions indicate areas receiving closer attention, meaning they contribute more significantly to feature learning.

From the comparisons of scenes (a)–(c), we observe that the regions receiving the closest attention closely correspond to the locations of relevant categories in the semantic segmentation maps. Additionally, the scopes of these highlighted areas align well with the segmented regions. The keywords in the questions are emphasized using the category colors from the segmentation maps, which consistently match the most attended regions in the attention maps. This validates that our model, which is designed to incorporate semantic segmentation results, effectively focuses on critical image regions. As a result, it achieves improved learning of high-level category-specific semantic information, enhancing its ability to reason over remote sensing imagery.

VIII. Conclusion

In this article, we propose a novel remote sensing VQA model that leverages semantic segmentation results to enhance the focus on critical visual features. Specifically, we introduce a graph-based structure and adopt a dual-branch feature extraction strategy to capture high-level category-specific semantic information. Furthermore, beyond the conventional VQA framework, we incorporate a contrastive learning module to ensure more effective learning of positive samples while mitigating the influence of irrelevant regions.

Extensive experimental evaluations demonstrate that our proposed model significantly outperforms existing SOTA approaches in remote sensing VQA. The qualitative and quantitative results validate the effectiveness of our method in enhancing both accuracy and interpretability, highlighting the importance of semantic segmentation in guiding visual attention for VQA tasks. Additionally, our model provides a reference for integrating semantic segmentation into VQA tasks, although its effectiveness benefits from the inherent clarity of land-cover categories in remote sensing images and the consistent resolution between the training and validation sets, as discussed earlier. Therefore, investigating the generalizability of our model to ordinary images without clearly defined land-cover categories, as well as to images with varying levels of resolution, is a promising direction for future research.

REFERENCES

[1] T. Arnold, M. De Biasio, A. Fritz, and R. Leitner, "Uav-based multispectral environmental monitoring," in *Proc. IEEE SENSORS*, 2010, pp. 995–998.

[2] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review," *Int. J. Remote Sens.*, vol. 39, no. 9, pp. 2784–2817, May 2018.

[3] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[4] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 3–22, Nov. 2018.

[5] X. Li, J. Deng, and Y. Fang, "Few-shot object detection on remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[6] X. Yang et al., "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 238–262, Jul. 2021.

[7] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[8] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[9] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[10] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "ReferItGame: Referring to objects in photographs of natural scenes," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 787–798.

[11] Y. Zhan, Z. Xiong, and Y. Yuan, "RSVG: Exploring data and models for visual grounding on remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5604513.

[12] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6274–6283.

[13] 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.

[14] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 39–48.

[15] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[16] J. Wang, Z. Zheng, Z. Chen, A. Ma, and Y. Zhong, "EarthVQA: Towards queryable Earth via relational reasoning-based remote sensing visual question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 6, pp. 5481–5489.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[18] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[19] A. Sarkar and M. Rahnemoonfar, "RESCUENet-VQA: A large-scale visual question answering benchmark for damage assessment," in *Proc. IGARSS - IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 1150–1153.

[20] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[21] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6904–6913.

[22] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6700–6709.

[23] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 2901–2910.

[24] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 30–38.

[25] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1–9.

[26] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 39, 2016, pp. 1137–1149.

[28] Z. Wen, G. Xu, M. Tan, Q. Wu, and Q. Wu, "Debiased visual question answering from feature and sample perspectives," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 3784–3796.

[29] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13–23.

[30] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," 2019, *arXiv:1908.07490*.

[31] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.

[32] W. Dai et al., "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 1–14.

[33] V. Dey, Y. Zhang, and M. Zhong, "A review on image segmentation techniques with remote sensing perspective," in *Proc. ISPRS TC VII Symp.—100 Years ISPRS*, vol. 38, Jan. 2010.

[34] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4095–4104.

[35] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet—A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[37] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 84–98, Nov. 2021.

[38] J. Zhuang, J. Yang, L. Gu, and N. Dvornek, "ShelfNet for fast semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 847–856.

[39] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[40] L. Cui, X. Jing, Y. Wang, Y. Huan, Y. Xu, and Q. Zhang, "Improved Swin transformer-based semantic segmentation of postearthquake dense buildings in urban areas using remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 369–385, 2023.
[41] 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.*, Jul. 2021, pp. 1218–1221.
[42] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigearthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5901–5904.
[43] 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.
[44] 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.
[45] 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.
[46] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.
[47] C. Chappuis, V. Zermatten, S. Lobry, B. Le 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 (CVPRW)*, Jun. 2022, pp. 1371–1380.
[48] 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.
[49] A. Sarkar and M. Rahnemoonfar, "VQA-Aid: Visual question answering for post-disaster damage assessment and analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 8660–8663.
[50] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change detection meets visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630613.
[51] L. Tosato, H. Boussaid, F. Weissgerber, C. Kurtz, L. Wendling, and S. Lobry, "Segmentation-guided attention for visual question answering from remote sensing images," in *Proc. IGARSS - IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2024, pp. 2750–2754.
[52] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," 2021, *arXiv:2110.08733*.
[53] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.
[54] Y. Gao, Z. Bai, M. Zhou, B. Jia, P. Gao, and R. Zhu, "Adaptive conditional reasoning for remote sensing visual question answering," *Remote Sens.*, vol. 17, no. 8, p. 1338, Apr. 2025.
[55] A. Sarkar, T. Chowdhury, R. R. Murphy, A. Gangopadhyay, and M. Rahnemoonfar, "SAM-VQA: Supervised attention-based visual question answering model for post-disaster damage assessment on remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4702716.
[56] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[57] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[58] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
[59] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.
[60] J. Devlin, M.-W. 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.*, vol. 1, Jun. 2019, pp. 4171–4186.
[61] F. Li et al., "LLaVA-NeXT-interleave: Tackling multi-image, video, and 3D in large multimodal models," 2024, *arXiv:2407.07895*.
[62] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded large vision-language model for remote sensing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 27831–27840.

Shuyi Wen received the B.S. degree in computer science and technology from Hunan Normal University, Changsha, China, in 2023. She is currently pursuing the M.S. degree in computer science with South China University of Technology, Guangzhou, China.

Her research interests include computer vision and remote sensing.

Aihua Mao received the B.Eng. degree from Hunan University, Changsha, China, in 2002, the M.Sc. degree from Sun Yat-sen University, Guangzhou, China, in 2005, and the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 2009.

He is a Professor with the School of Computer Science and Engineering, South China University of Technology (SCUT), Guangzhou. His research interests include 3-D vision and computer graphics.

Ran Yi received the B.Eng. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2016 and 2021, respectively.

She is an Assistant Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. Her research interests include computer vision, computer graphics, and computational geometry.

Yong-Jin Liu (Senior Member, IEEE) received the B.Eng. degree from Tianjin University, Beijing, China, in 1998, and the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004.

He is a Professor with the Department of Computer Science and Technology, Tsinghua University. His research interests include computer graphics and computer-aided design.