Cross-Modal Dynamic Aggregation with Adaptive Relevance Modulation Fusion Network for Remote Sensing Visual Question Answering

Zihua Zuo and Chao Li(✉)

Department of Computer Science and Technology, Harbin Engineering University. Harbin, China
lichao006@hrbeu.edu.cn

Abstract. The fundamental objective of the Remote Sensing Visual Question Answering (RS VQA) paradigm is to deduce precise linguistic responses for natural language inquiries grounded in Earth observation imagery[1]. Nevertheless, the inherent disparity separating rudimentary image characteristics from abstract semantic concepts presents a substantial hurdle in decoding intricate textual queries. Moreover, the lack of dynamic modulation mechanisms for integrating visual and textual features impedes the balanced interpretation of image content and textual semantics. In response to these identified limitations, this study introduces a novel architecture termed the Cross-Modal Dynamic Aggregation with Adaptive Relevance Modulation Fusion Network (CDAR-Net), tailored specifically for the RS VQA domain. Specifically, we propose a Cross-Modal Dynamic Aggregation Module (CMDA) that employs a multi-level attention mechanism to iteratively fuse text-guided visual features, enabling a progressive transition and dynamic integration from low-level visual features to high-level semantic information. We further introduce an Adaptive Correlation Modulation Fusion Module (ACMF) that dynamically adjusts visual and textual feature weights based on questions context, enhancing the representation of relevant information. Experimental results demonstrate that CDAR-Net outperforms existing state-of-the-art methods on three RS VQA datasets.

Keywords: remote sensing, visual question answering (VQA), cross-modal fusion, dynamic aggregation.

1 Introduction

The domain of Remote Sensing Visual Question Answering (RS VQA) constitutes a formidable research endeavor. Specifically, it necessitates the profound comprehension and inferential analysis of Earth observation data in order to synthesize precise linguistic responses. This technology holds significant potential for advancing environmental monitoring and disaster management by facilitating more effective analysis of environmental changes and disaster scenarios[2]. Early research[3] focused on extracting visual features using CNNs and text features using RNNs, ultimately integrating these features into a unified vector representation.

With the continuous advancement of research, the rich spatial and semantic information in RS images has driven RS VQA methods to focus increasingly on enhancing visual feature extraction. Yuan et al.[4] introduced a CNN-based multi-level feature learning approach that jointly extracts language-guided image features and progressively increases sample difficulty, thereby improving semantic comprehension. Zheng et al.[5] employed CNNs to capture spatial information, combining them with word embeddings for semantic representation and leveraging attention mechanisms and bilinear techniques to enhance feature encoding. A novel framework designated as SHRNet was introduced by Zhang et al.[6]. To derive multi-resolution visual representations, this architecture seamlessly fuses convolutional neural networks with hashing paradigms. Furthermore, the predictive precision of the model is significantly augmented through the explicit encoding of spatial contextual information. Nevertheless, conventional approaches frequently fail to reconcile the inherent disparity between rudimentary visual representations and abstract semantic concepts. Consequently, this disconnect fosters a disproportionate dependency on raw image data, thereby severely constraining the model's proficiency in decoding intricate inquiries.

In recent years, numerous RS VQA models have increasingly emphasized the interaction between visual and textual features. Chappuis et al. introduced Prompt-RSVQA[7], which extracts image features using CNNs and processes them with distilBERT[8], but it lacks a dynamic mechanism for fusing visual and textual information. Leonard et al.'s LiT-4-RSVQA[9] improves RS VQA efficiency and accuracy with a simplified Transformer architecture while its fusion mechanism remains rudimentary. Bazi et al.[10] proposed a Transformer-based visual-linguistic model leveraging the pre-trained CLIP model[11] for feature extraction from both modalities, incorporating a co-attention mechanism for interaction modeling. Nevertheless, it struggles to dynamically balance the importance of visual and textual features during inference. Similarly, Feng et al.'s UCAGAN[12] mitigates attention misalignment and class imbalance with an enhanced SAMFPN module[13], yet it relies on fixed visual feature extraction and fails to effectively adapt the fusion of visual and semantic information.

More recently, the emergence of remote-sensing-oriented multimodal large language model (MLLM) frameworks has further pushed RSVQA toward open-world and knowledge-enhanced reasoning, but also raises new challenges in grounding and efficiency[14,15].

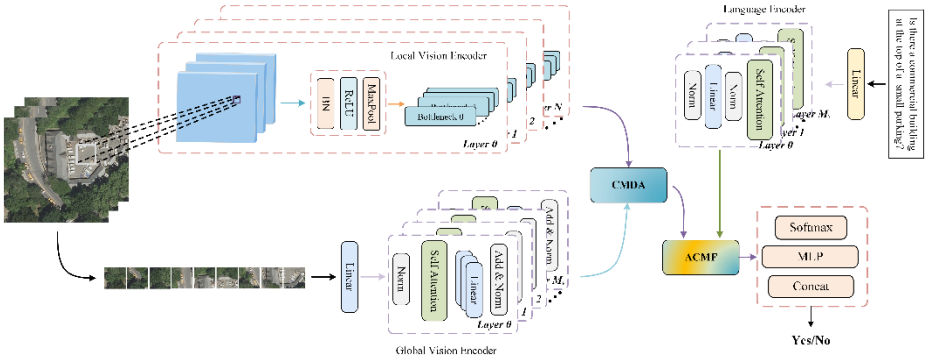

Fig. 1. The detailed architecture of CDAR-Net, consists of two vision encoders, a text encoder, a classifier, and two specialized modules: CMDA for visual fusion and ACMF for vision-text fusion.

While existing RS VQA methods have advanced feature extraction and fusion, they continue to face two critical challenges: (1) an inadequate ability to bridge the semantic gap between low-level visual features and high-level semantic representations, and (2) the lack of an effective dynamic mechanism to optimally balance the contributions of visual and textual features.

To address these challenges, we introduce the Cross-Modal Dynamic Aggregation with Adaptive Relevance Modulation Fusion Network (CDAR-Net) for RS VQA. CDAR-Net integrates the ResNet[16] for extracting low-level visual features and the Vision Transformer (ViT)[17], which captures high-level semantic information and contextual details via global self-attention mechanisms. Central to our approach is the Cross-Modal Dynamic Aggregation (CMDA) module, which utilizes textual information to guide the hierarchical fusion of visual features. Employing a multi-level attention mechanism, the CMDA module adaptively adjusts feature weights, effectively enhancing the complementarity between different semantic levels, enabling a seamless and dynamic transition from low-level features to high-level understanding. Furthermore, we introduce the Adaptive Correlation Modulation Fusion (ACMF) module, which dynamically balances the fusion weights of visual and textual features based on query context, improving the representation of relevant information in the fused features.

Comprehensive empirical evaluations were conducted across three open-source RS VQA benchmarks, namely RSVQA-LR, RSVQA-HR[3], and RSIVQA[5]. The resulting metrics substantiate that CDAR-Net consistently eclipses the capabilities of current state-of-the-art (SOTA) paradigms. Through a series of ablation studies, we further validate the effectiveness of our approach.

The primary innovations introduced in this manuscript are delineated below:

- We propose the CDAR-Net to address the RS VQA task, combining ResNet for local feature extraction and ViT for capturing high-level semantic and contextual information.

- We introduce the CMDA module, which utilizes a multi-level attention mechanism to hierarchically fuse visual features guided by textual information, dynamically reconciling the inherent representational disparity separating rudimentary visual cues from abstract semantic concepts.
- We propose the ACMF module, which adaptively balances the weights of visual and textual features based on query context, improving the relevance and effectiveness of the fused features.
- Comprehensive empirical evaluations executed across a triad of open-access RS VQA benchmarks substantiate that the proposed architecture consistently eclipses contemporary state-of-the-art (SOTA) baselines.

2 Methods

In this section, as shown in Fig. 1, we comprehensively introduce the architecture of CDAR-Net. CDAR-Net includes two independent vision encoders and the CMDA described in Sec. 2.2, along with a language encoder and the ACMF elaborated in Sec. 2.3. Before delving into these components, we formally define the problem addressed in this study in Sec. 2.1.

2.1 Preliminary

We formulate RS VQA as a classification task, where each image-text pair is associated with a single correct answer selected from a set of A candidate answers. The dataset D consists of N image-text pairs, each with its corresponding correct answer. Our goal is to train an RS VQA model capable of accurately predicting the correct answer given a RS image and its associated question. To achieve this, we aim to maximize the log-likelihood of the correct answers during training, which can be formally expressed as:

$$L = \frac{1}{N}\sum_{i=1}^{N} \log P\left(y_i \middle| I_i, q_i; \theta\right) \tag{0}$$

Within this formulation, L stands for the objective loss to be minimized. For any given data index i, the corresponding ground-truth annotation, satellite image, and textual inquiry are denoted by y_i, I_i, and q_i, respectively. Furthermore, the entire set of optimizable weights within the network architecture is parameterized by θ.

The overall architecture of the proposed CDAR-Net is depicted in **Fig. 1.** Given an RS image I, it is processed through both a CNN and a ViT to extract local image features and global image features, respectively. Simultaneously, a question q is passed through BERT to extract textual features, resulting in the textual representation $\mathbf{h}_{\mathrm{t}}$. After obtaining these three types of features, we first apply the proposed CMDA to perform text-guided cross-modal fusion of $\mathbf{h}_{\mathrm{CNN}}$ and $\mathbf{h}_{\mathrm{ViT}}$, yielding the fused visual features h_v^{fusion}. Next, the textual feature $\mathbf{h}_{\mathrm{t}}$ and the fused visual features

h_v^{fusion} are combined, and we use the ACMF module to perform text-context-based adaptive fusion of the visual and textual features, resulting in the final fused image-text feature h^{fusion}. Lastly, the fused feature h^{fusion} is passed through an answer classifier to generate the final output $\hat{y}$.

$$\hat{y} = \text{softmax}(W_{out} \cdot GELU(Norm(W_1 h^{fusion} + b_1)W_2 + b_2) + b_{out}) \quad (2)$$

where W_1, W_2, and W_{out} are learnable weight matrices, b_1, b_2, and b_{out} represent bias terms.

2.2 Cross-Modal Dynamic Aggregation Network

After extracting local visual features $\mathbf{h}_{\text{CNN}}$ from the CNN and global visual features $\mathbf{h}_{\text{ViT}}$ from the ViT, we propose the CMDA, as depicted in Fig. 2. CMDA leverages the textual feature $\mathbf{h}_{\text{t}}$ to guide the fusion of visual features, integrating an attention mechanism to adaptively adjust the modality feature weights. This improves the model's flexibility and adaptability, facilitates the capture of multi-scale visual information, enables deep fusion across visual modalities, and mitigates issues stemming from modality discrepancies.

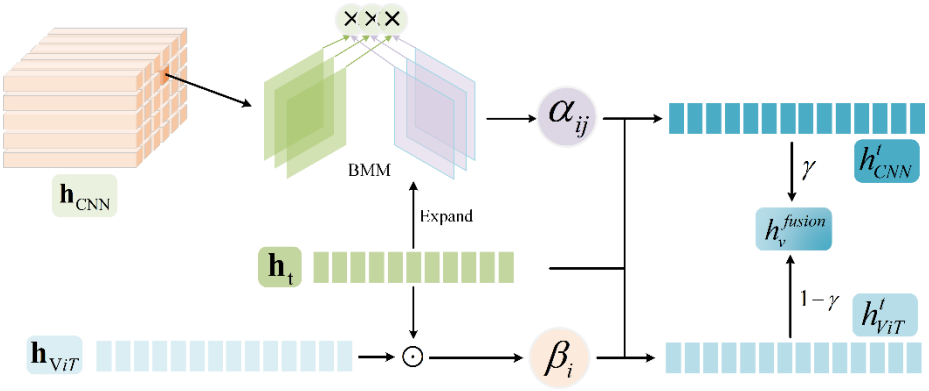

Fig. 2. The architecture of CMDA, where "Expand" refers to the dimension expansion of text embeddings, "BMM" denotes batch matrix multiplication, and ⊙ represents element-wise multiplication.

First, we compute the attention weights of the textual features with respect to both the local and global visual features, specifically the correlation scores between low-level and high-level semantic information. For each $\mathbf{h}_{\text{t}}$, we calculate its correlation α_{ij} with the j-th local visual feature $\mathbf{h}_{\text{CNN},j}$ and its correlation β_i with the i-th dimension of $\mathbf{h}_{\text{ViT}}$, formalized as follows:

$$\alpha_{ij} = \frac{exp(\mathbf{h}_t^{\cdot} \mathbf{h}_{CNN,j})}{\sum_{k=1}^{N} exp(\mathbf{h}_t^{\cdot} \mathbf{h}_{CNN,k})}, j = 1,2,...,N$$
$$\beta_i = \frac{\exp\left(\mathbf{h}_{t,i}^{\cdot} \mathbf{H}_{\mathrm{VIT,i}}\right)}{\sum_{k=1}^{M} \exp\left(\mathbf{h}_{t,k}^{\cdot} \mathbf{H}_{\mathrm{ViT,k}}\right)}, i = 1,2,...,M \quad (3)$$

here, N and M represent the number of local visual features and the dimensionality of the features, respectively $\mathbf{h}_{\mathrm{t},k}$ and $\mathbf{h}_{\mathrm{ViT},k}$ denote the components of the textual and global visual features in the k-th feature dimension. Next, based on α_{ij} and β_i, we compute the weighted representations of $\mathbf{h}_{\mathrm{CNN}}$ and $\mathbf{h}_{\mathrm{ViT}}$ guided by the text, denoted as $\mathbf{h}_{\mathrm{CNN}}^{\mathrm{t}}$ and $\mathbf{h}_{\mathbf{ViT}}^{\mathrm{t}}$:

$$h_{CNN}^{t} = \sum_{j=1}^{N} \alpha_{ij} \mathbf{h}_{\mathrm{CNN},j}, \quad h_{ViT}^{t} = \beta_i \square \ \mathbf{h}_{ViT} \quad (4)$$

Finally, we fuse the local visual features, weighted by the text-guided attention, with the global visual features to obtain the fused visual feature h_v^{fusion}:

$$h_v^{fusion} = \gamma h_{CNN}^{t} + (1-\gamma) h_{ViT}^{t} \quad (5)$$

where γ is a learnable parameter that controls the fusion ratio between local and global features.

Table 1. Comparison of CDAR-Net with existing SOTA methods on the RSVQA-LR and RSVQA-HR datasets. Bolded values indicate the best performance in each category or metric.

Dtasets	Type	RSVQA[3]	EasyToHard[4]	Bi-Modal[10]	UCAGAN[12]	SAMFPN[13]	SHRNet[6]	MADNet[18]	CDAR-Net
RSVQA-LR	Count	67.01%	69.22%	72.22%	70.68%	69.51%	73.87%	72.85%	**75.30%**
	Presence	87.46%	90.66%	91.06%	91.17%	91.21%	91.03%	90.96%	**91.46%**
	Comparison	81.50%	87.49%	91.16%	89.08%	91.01%	90.48%	91.68%	**92.09%**
	Rural/Urban	90.00%	91.67%	92.66%	91.00%	92.67%	94.00%	95.00%	**96.00%**
	AA	81.49%	84.76%	86.78%	85.48%	86.10%	87.34%	87.62%	**88.71%**
	OA	79.08%	83.09%	85.56%	83.95%	84.76%	85.85%	85.97%	**87.13%**
RASVQA-HR Test Set 1	Count	68.63%	69.06%	69.80%	**71.53%**	69.19%	70.04%	70.02%	69.56%
	Presence	90.43%	91.39%	92.03%	91.10%	91.99%	**92.45%**	92.36%	92.02%
	Comparison	88.19%	89.75%	91.83%	88.56%	91.44%	91.68%	91.87%	**91.94%**
	Area	85.24%	85.92%	86.27%	86.83%	85.84%	86.35%	86.58%	**90.81%**
	AA	83.12%	83.97%	84.98%	84.51%	84.61%	85.13%	85.21%	**86.08%**
	OA	83.23%	84.16%	85.30%	84.56%	84.94%	85.39%	85.51%	**86.01%**
RASVQA-HR Test Set 2	Count	61.47%	61.95%	63.06%	63.28%	62.58%	**63.42%**	63.38%	62.44%
	Presence	86.26%	87.97%	89.37%	88.51%	89.04%	**89.81%**	89.69%	89.44%

	Comparison	85.94%	87.68%	89.62%	87.76%	88.95%	89.44%	89.82%	**89.89%**
	Area	76.33%	78.62%	80.12%	78.47%	78.36%	80.37%	80.58%	**88.42%**
	AA	77.50%	79.06%	80.54%	79.51%	79.73%	80.76%	80.87%	**82.55%**
	OA	78.23%	79.29%	81.23%	80.21%	80.54%	81.37%	81.51%	**82.33%**

2.3 Adaptive Correlation Modulation Fusion Module

To guarantee that the computational model concurrently assimilates the visual context and strictly interprets the linguistic semantics of the query, we formulate the ACMF. This component functions by adaptively apportioning the significance of textual and visual characteristics amidst the aggregation procedure. By executing a context-aware calibration of multi-modal feature coefficients, this approach continuously tailors the integration process to the specific prompt. Such adaptive scaling ensures the optimal retention of pertinent details within the final aggregated embedding. Specifically, as illustrated in Fig.3, ACMF generates adaptive weights by computing the correlations between the visual and textual features and the question context, and subsequently performs weighted fusion of the features.

First, we extract a deeper semantic contextual representation $C \in \mathbb{R}^{d_c}$ from the question and use it to compute the correlation scores R_V between h_v^{fusion} and the context C, as well as the correlation scores R_T between $\mathbf{h}_t$ and C:

$$\begin{aligned} R_V &= \sigma\left(W_V[h_v^{fusion}; C] + b_V\right) \\ R_T &= \sigma\left(W_T[\mathbf{h}_t; C] + b_T\right) \end{aligned} \tag{6}$$

here, $[h_v^{fusion}; C]$ and $[\mathbf{h}_t; C]$ denote the concatenation of the features, with W_V, W_T, b_V, and b_T are learnable parameters, and σ representing the activation function. Subsequently, we transform the correlation scores into weights using the Softmax function, ensuring that the weights sum to 1:

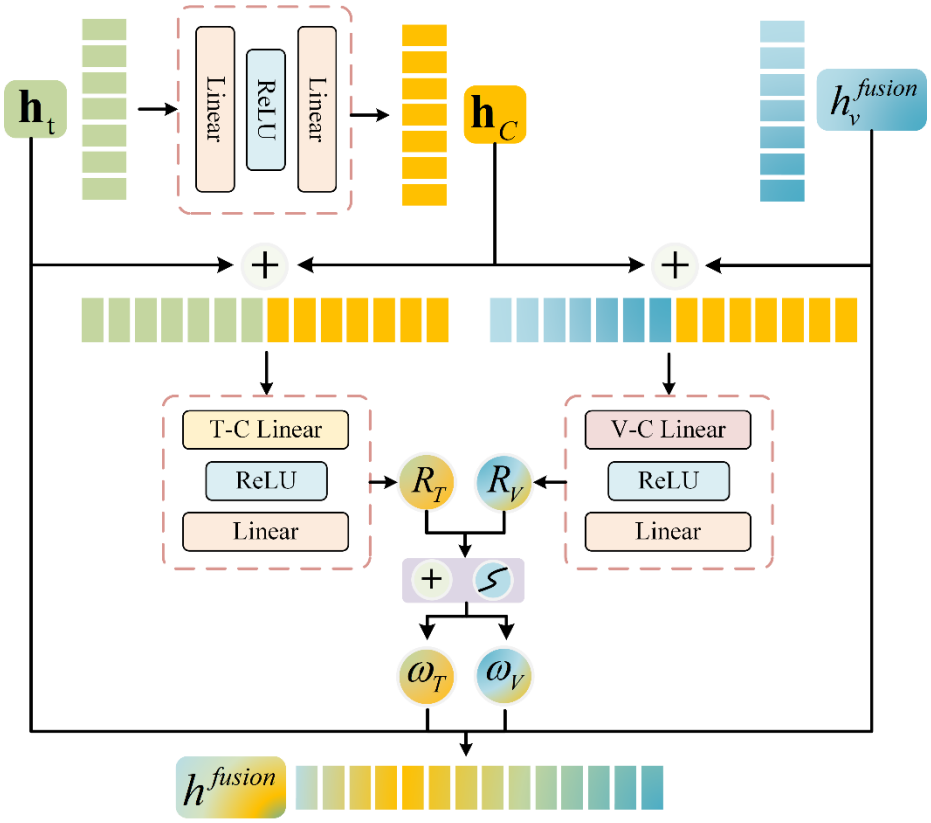

Fig. 3. The architecture of ACMF module, where T-C Linear and V-C Linearare employed for the linear transformations of text-context and visual-context, respectively. The input dimensions correspond to the hidden layer sizes $\mathbf{h}_t$ and $\mathbf{h}_c$, as well as $\mathbf{h}_c$ and h_v^{fusion}.

$$w_V = \frac{\exp(R_V)}{\exp(R_V)+\exp(R_T)}$$
$$w_T = \frac{\exp(R_T)}{\exp(R_V)+\exp(R_T)} \tag{7}$$

Finally, based on the calculated weights, we modulate and combine the visual and textual features to obtain the final fused representation h^{fusion}:

3 Experiments

3.1 Datasets

To rigorously assess the efficacy of the proposed architecture, we conduct empirical validations across three open-source RS VQA benchmarks: RSVQA-LR[3], RSVQA-HR[3], and RSIVQA[5].

RSVQA-LR dataset[3]: Derived from Sentinel-2 imagery (10m resolution), this dataset comprises 772 images (256×256) spanning ~6.55 km^2, along with 77,232 image-question-answer triplets across four question types: Count, Presence, Comparison, and Rural/Urban (specific to low resolution). The entire dataset was partitioned into three distinct subsets, with 77.8% allocated for training purposes, and the remaining data divided equally (11.1% each) between validation and testing.

RSVQA-HR dataset[3]: Derived from USGS High Resolution Orthoimagery with a 15 cm ground resolution, this dataset consists of 10,659 images (512×512) spanning ~5,898 m^2 and 1,066,316 triplets. It supports Count, Presence, Comparison, and Area questions, with the Area type unique to its high resolution. The split allocates 61.5%, 11.2%, and 27.3% to training, validation, and test sets, with test set 2 (6.8%) featuring unseen sensors to evaluate model generalization.

RSIVQA dataset[5]: This dataset aggregates imagery from AID[19], UC-Merced[20],Sydney[21], DOTA[22], and HRRSD[23], comprising 111,134 image-question pairs across Yes/No, Number, and Others question types. The official split assigns 80%, 10% and 10% to training, validation, and test sets, respectively.

3.2 Implementation Details

The implementation is based on Python 3.8 and PyTorch 2.0, executed on two NVIDIA RTX 4090 GPUs. The visual encoder employs pre-trained weights from ViT-B[17] and ResNet-50[16], with input images resized to 224×224 through random cropping and augmented using RandAugment[24]. The language encoder utilizes pre-trained BERT weights[25]. Optimization of the network was executed over a span of 20 epochs, utilizing mini-batches of 64 samples and driven by the AdamW algorithm[26].The initial learning rate of $2e^{-5}$ follows a cosine decay schedule down to $1e^{-8}$, with a weight decay of 0.05.

3.3 Comparison with the State-of-the-Art Methods

In this section, we conduct extensive experiments on three publicly available datasets and compare our approach with current SOTA methods, as shown in TABLEs 1 and 2. As shown in TABLE 1, on the RSVQA-LR dataset, our method outperforms all compared approaches across every question type and evaluation metric. In particular, for the challenging "Count" type questions, the proposed approach achieves an accuracy of 75.30%, surpassing the previous best result (73.87% by SHRNet) by more than 1.4%. For "Presence" and "Comparison" questions, our model attains 91.46% and 92.09% accuracy, respectively, consistently improving upon SOTA baselines such as

MADNet (90.96% and 91.68%) and SHRNet (91.03% and 90.48%). In the Rural/Urban classification task, our method achieves a remarkable 96.00% accuracy, surpassing MADNet (95.00%) . Consequently, our approach attains average accuracy (AA) and overall accuracy (OA) of 88.71% and 87.13%, respectively.

Table 2. Comparison of CDAR-Net with existing SOTA methods on the RSIVQA dataset.

Method	Yes/No	Number	Others	AA	OA
MAIN[5]	92.82%	56.71%	54.50%	68.01%	77.39%
EasyToHard[4]	95.49%	49.03%	63.65%	69.39%	79.70%
SHRNet[6]	97.64%	57.89%	84.60%	80.04%	84.46%
CDAR-Net	**97.85%**	**70.56%**	**94.83%**	**87.75%**	**89.74%**

Secondly, on the RSVQA-HR Test Set 1, as shown in TABLE 1, our model demonstrates a significant improvement over existing approaches, particularly when considering the OA and AA. Specifically, for the challenging "Area" category, our method achieves an accuracy of 90.81%, surpassing the best competing model (UCAGAN at 86.83%) by more than 4%. This, combined with strong results in "Presence" (92.02%) and "Comparison" (91.94%), leads to a notable increase in both AA and OA. With these improvements, our final AA and OA reach 86.08% and 86.01%, respectively, significantly higher than the previous performance level, which hovered around 85%. In Test Set 2, the trends observed in Test Set 1 largely persist. For the "Area" category, our approach decisively outperforms all prior methods, attaining 88.42% accuracy and exceeding the previous best method, MADNet, by 7.84%. Meanwhile, our model achieves an AA of 82.55% and an OA of 82.33%, surpassing the previously leading methods, which had hovered around 81.5%.

Lastly, as shown in TABLE 2, our approach consistently outperforms existing SOTA models across all evaluated question types in the RSIVQA dataset. In the "Yes/No" category, our model achieves a near-perfect accuracy of 97.85%, slightly improving upon the previously best-performing SHRNet (97.64%). The improvements are even more pronounced in the "Number" and "Others" categories. Specifically, we increase the "Number" accuracy to 70.56%, representing an improvement of over 12% compared to the SHRNet baseline (57.89%), and achieve 94.83% in the "Others" category, surpassing SHRNet's 84.60% by more than 10%. These advancements result in an AA and OA of 87.75% and 89.74%, respectively. In comparison to prior benchmarks that reported AA and OA values below 80% and 85%, our model sets a new standard, demonstrating both its effectiveness in handling diverse question types within the RSIVQA dataset.

3.4 Ablation Study

The current section is dedicated to a rigorous component-wise evaluation. Utilizing three standard data collections, we deconstruct the model to assess how inherently each constituent element bolsters the overarching performance, with the corresponding statistical findings synthesized in Table 3. The results consistently show that the removal of any module leads to a decline in the model's performance across all datasets

and metrics. Specifically, omitting the CNN backbone reduces the OA on RSVQA-LR from 87.13% to 86.01%. Similarly, removing the ViT component results in a further decrease in both OA and AA, highlighting its critical role in robust visual feature extraction.

Table 3. Ablation study of CDAR-Net under different configurations on the RSVQA-LR, RSVQA-HR, and RSIVQA datasets.

Variant	RSVQA-LR		RSVQA-HR (Test set 1)		RSVQA-HR (Test set 2)		RSIVQA	
	OA	AA	OA	AA	OA	AA	OA	AA
w/o CNN	86.01%	87.16%	85.18%	85.23%	81.17%	81.33%	87.97%	85.88%
w/o ViT	85.23%	85.81%	84.92%	85.02%	80.76%	80.92%	87.21%	85.48%
w/o MLA	86.69%	87.98%	85.48%	85.52%	81.80%	81.75%	88.56%	86.25%
w/o ACMF	86.59%	87.75%	85.64%	85.79%	81.79%	81.88%	88.98%	86.82%
CDAR-Net	**87.13%**	**88.71%**	**86.01%**	**86.08%**	**82.33%**	**82.55%**	**89.74%**	**87.75%**

The multi-level attention mechanism (MLA) module is crucial for improving cross-modal alignment. Without MLA, the OA on RSVQA-LR drops from 87.13% to 86.69%, and similar performance degradations are observed on RSVQA-HR and RSIVQA, further emphasizing MLA's role in hierarchically guided feature integration. Similarly, the ACMF module significantly improves performance. Removing ACMF results in a reduction in accuracy across all benchmarks, with the AA on RSIVQA decreasing from 87.75% to 86.82%. Our full model, which integrates the CNN, ViT, MLA, and ACMF components, achieves the highest OA and AA across all test scenarios.

These findings validate that each component contributes to the final representation and reasoning process, and together, they deliver SOTA performance across multiple challenging RS VQA benchmarks.

4 Conclusion

In this paper, we propose the CDAR-Net for RSVQA, which incorporates two key modules: CMDA and ACMF. The CMDA module facilitates a progressive transition from low-level visual features to high-level semantic understanding, effectively bridging the semantic gap between these two feature domains. Meanwhile, the ACMF module employs a dynamic and context-sensitive feature weighting mechanism to enable efficient dynamic modulation, achieving deep fusion of visual and textual features. Rigorous empirical analyses executed on three established RSVQA benchmarks substantiate that CDAR-Net systematically surpasses contemporary state-of-the-art baselines. Such results compellingly affirm both the framework's efficacy and its robust generalization capacity.

References

1. Ke Hu, Wenzhen Zhang, and Shichao Zhang, “Multi-level dynamic gated inter-action fusion network for remote sensing visual question answering,” in Advanced Intelligent Computing Technology and Applications, De-Shuang Huang, Wei Chen, Yijie Pan, and Haiming Chen, Eds., Singapore, 2025, pp. 282–293, Springer Nature Singapore.
2. Congcong Wen, Yuan Hu, Xiang Li, Zhenghang Yuan, and Xiao Xiang Zhu, "Vision-language models in remote sensing: Current progress and future trends," CoRR, vol. abs/2305.05726, 2023.
3. Sylvain Lobry, Begüm Demir, and Devis Tuia, "RSVQA meets bigearthnet: A new, large-scale, visual question answering dataset for remote sensing," in IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2021, Brussels, Belgium, July 11-16, 2021. 2021, pp. 1218-1221, IEEE.
4. Zhenghang Yuan, Lichao Mou, Qi Wang, and Xiao Xiang Zhu, "From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data," IEEE Trans. Geosci. Remote. Sens., vol. 60, pp. 1-11, 2022.
5. Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu, "Mutual attention inception network for remote sensing visual question answering," IEEE Trans. Geosci. Remote. Sens., vol. 60, pp. 1-14, 2022.
6. Zixiao Zhang, Licheng Jiao, Lingling Li, Xu Liu, Puhua Chen, Fang Liu, Yuxuan Li, and Zhicheng Guo, "A spatial hierarchical reasoning network for remote sensing visual question answering," IEEE Trans. Geosci. Remote. Sens., vol. 61, pp. 1-15, 2023.
7. Christel Chappuis, Valérie Zermatten, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia, "Prompt-rsvqa: Prompting visual context to a language model for remote sensing visual question answering," in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022. 2022, pp. 1371-1380, IEEE.
8. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," CoRR, vol. abs/1910.01108, 2019.
9. Leonard W. Hackel, Kai Norman Clasen, Mahdyar Ravanbakhsh, and Begüm Demir, "LIT-4-RSVQA: lightweight transformer-based visual question answering in remote sensing," in IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2023, Pasadena, CA, USA, July 16-21, 2023. 2023, pp. 2231-2234, IEEE.
10. Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Mohamed Lamine Mekhalfi, Mansour Abdulaziz Al Zuair, and Farid Melgani, "Bi-modal transformer-based approach for visual question answering in remote sensing imagery," IEEE Trans. Geosci. Remote. Sens., vol. 60, pp. 1-11, 2022.
11. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Marina Meila and Tong Zhang, Eds. 2021, vol. 139 of Proceedings of Machine Learning Research, pp. 8748-8763, PMLR.
12. Jiangfan Feng, Etao Tang, Maimai Zeng, Zhujun Gu, Pinglang Kou, and Wei Zheng, "Improving visual question answering for remote sensing via alternate-guided attention and combined loss," Int. J. Appl. Earth Obs. Geoinformation, vol. 122, pp. 103427, 2023.
13. Jiangfan Feng and Hui Wang, "A multi-scale contextual attention network for remote sensing visual question answering," Int. J. Appl. Earth Obs. Geoinformation, vol. 126, pp. 103641, 2024.

14. Fang Liu, Licheng Jiao, Xu Liu, Lingling Li, and Zixiao Zhang, "Visual-grounded hierarchical perception for remote sensing visual question answering," IEEE Transactions on Geoscience and Remote Sensing, 2025.
15. Xin Li, Liang Yao, Jiale Zhu, Chuanyi Zhang, Wenwen Dai, and Fan Liu, "Co-llava: Efficient remote sensing visual question answering via model collaboration," Remote Sensing, vol. 17, no. 3, pp. 466, 2025.
16. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. 2016, pp. 770-778, IEEE Computer Society.
17. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. 2021, OpenReview.net.
18. Yangyang Li, Yunfei Ma, Guangyuan Liu, Qiang Wei, Yanqiao Chen, Ronghua Shang, and Licheng Jiao, "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.
19. Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," IEEE Trans. Geosci. Remote. Sens., vol. 55, no. 7, pp. 3965-3981, 2017.
20. Yi Yang and Shawn D. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in 18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings, Divyakant Agrawal, Pusheng Zhang, Amr El Abbadi, and Mohamed F. Mokbel, Eds. 2010, pp. 270-279, ACM.
21. Fan Zhang, Bo Du, and Liangpei Zhang, "Saliency-guided unsupervised feature learning for scene classification," IEEE Trans. Geosci. Remote. Sens., vol. 53, no. 4, pp. 2175-2184, 2015.
22. Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge J. Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. 2018, pp. 3974-3983, Computer Vision Foundation / IEEE Computer Society.
23. Yuanlin Zhang, Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," IEEE Trans. Geosci. Remote. Sens., vol. 57, no. 8, pp. 5535-5548, 2019.
24. Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc Le, "Randaugment: Practical automated data augmentation with a reduced search space," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, Eds., 2020.
25. Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang, "Visualbert: A simple and performant baseline for vision and language," CORR, vol. abs/1908.03557, 2019.

26. Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. 2019, OpenReview.net.