FILTER-GROUP MIXTURE-OF-EXPERTS MODEL FOR REMOTE SENSING FORGED TARGET PERCEPTION

Ziyu Sun[1,*], *Enyuan Zhao*[1,*], *Chengjie Fang*[1], *Wei Xiong*[1], *Heng Pan*[2], *Yanbiao Li*[1,2,†]

[1] Hangzhou Institute for Advanced Study,
University of Chinese Academy of Sciences, Hangzhou, China
[2] Computer Network Information Center, Chinese Academy of Sciences, Beijing, China

ABSTRACT

High-reliability interpretation of remote sensing imagery is critical for national security and resource monitoring. However, forged images often present small object occlusion, noise, and complex textures, leading to weak cues and ambiguous features that hinder tampering detection. Existing methods largely rely on pixel-domain end-to-end learning, lacking the ability to explicitly model the frequency-domain artifacts characteristic of forgeries. To overcome this, we propose a Filter-Group Mixture-of-Experts (FG-MoE) framework that embeds classical signal processing priors into deep models. The architecture builds a heterogeneous expert library with Sobel for edge cues, Wavelet for multi-scale decomposition, Schmid for spatial textures, Gabor for directional features, and a Standard Expert for general-purpose representations. A gating mechanism dynamically weights and combines their outputs, fostering complementary reasoning. Experiments show that filter-driven priors enhance sensitivity to subtle traces under degraded conditions, outperforming conventional baselines.

Index Terms— Copy-Move Forgery, Remote Sensing, Mixture of Experts, Signal Processing Priors, Multimodal

1. INTRODUCTION

Remote sensing image forgery detection is particularly challenging due to the intrinsic characteristics of remote sensing imagery. The unique viewing perspectives, complex surface structures, and specific imaging conditions often render tampering traces weak and ambiguous. At the same time, the top-down perspective, dense distribution of small objects, and complex textures further exacerbate the detection difficulty, requiring region-level fine-grained modeling rather than generic image-level descriptions.

Traditional forgery detection approaches, such as PCA [1], DWT [2], SIFT [3], and ORB [4], are mostly designed for Copy-Move Forgery Detection (CMFD). These methods depend on shallow priors like edges or keypoints for block matching, but cannot capture semantic inconsistencies across forged regions. To overcome these limitations, deep learning methods such as BusterNet [5], STD-Net [6], and CMCF-Net [7] have been proposed, which significantly improve visual feature representation. Nevertheless, their reliance on the visual modality alone constrains their ability to interpret complex tampering scenarios. This has motivated the exploration of multimodal methods integrating textual information. For example, FakeShield [8] adopts a question-driven mechanism to enhance explainability and localization accuracy, highlighting the potential of semantic context in tampering analysis. Given the limitations of unimodal visual analysis, recent research has consequently shifted toward multimodal reasoning frameworks that jointly model visual evidence and linguistic cues. Building upon these multimodal advances, it becomes natural to explore more structured frameworks that can seamlessly integrate semantic reasoning with visual perception.

In this context, the Visual Question Answering (VQA)

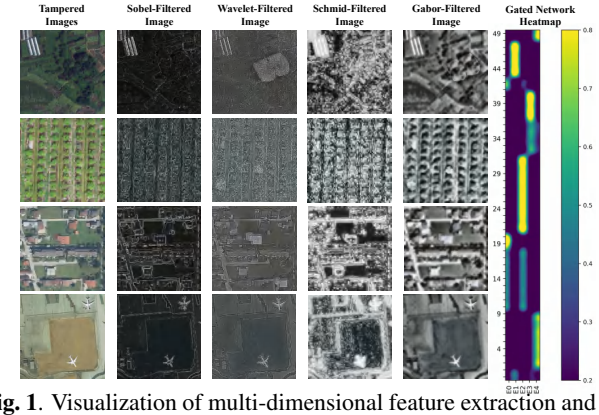


Fig. 1. Visualization of multi-dimensional feature extraction and expert contribution in FG-MoE. The first column shows the input tampered image. Columns two to five depict feature responses from four filter-based branches: Sobel, Wavelet, Schmid, and Gabor. The rightmost heatmap illustrates the expert weights dynamically assigned by the gating network.

[*] These authors contributed equally. [†] Corresponding author: lybmath@cnic.cn. The source code is available at `https://github.com/hlcy-404/FG-MoE`.

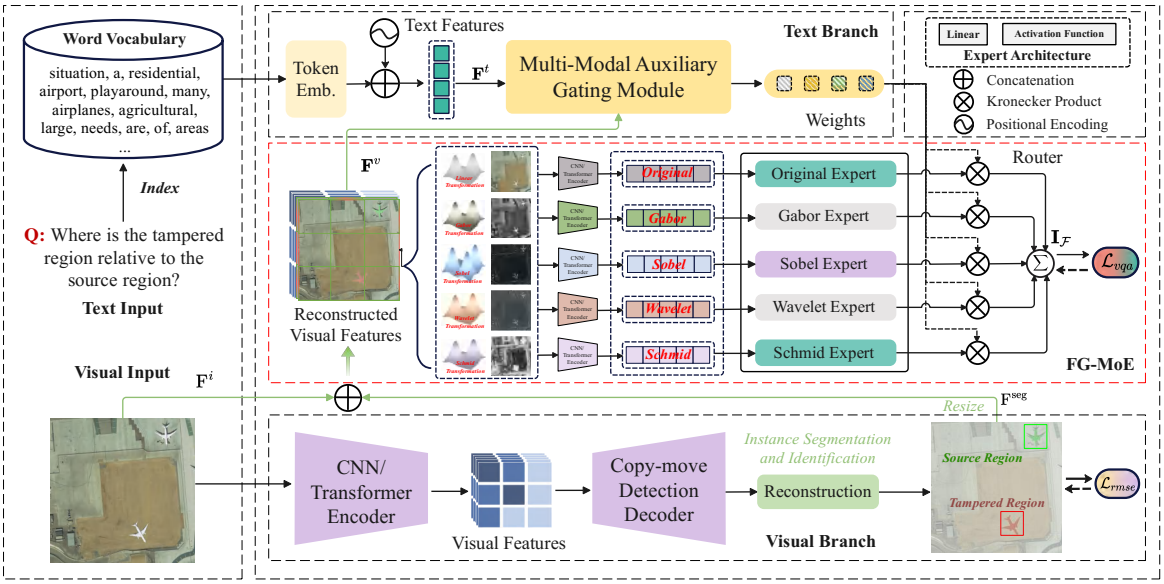


Fig. 2. An illustration of the proposed FG-MoE model.

framework has emerged as a promising direction. Since Lobry et al. [9] introduced VQA to remote sensing, subsequent works such as X-PFA [10], EarthVQA [11], GeoPixel [12], and CDVQA [13] have extended its scope. More recently, Zhang et al. [14] introduced Remote Sensing Copy-Move Question Answering (RSCMQA) to address forgery interpretation. Following this, Zhao et al. [15] proposed the Real-RSCM dataset and a multimodal mixture-of-experts model (CM-MMoE). However, their homogeneous, data-driven expert structure struggles to capture frequency-domain perturbations and texture artifacts, making it difficult to address small-object occlusion, noise interference, and artifact misguidance.

To address these challenges, this paper introduces the Filter-Group Mixture-of-Experts (FG-MoE) framework, which integrates signal processing priors into deep multimodal learning. As illustrated in Fig. 1 and Fig. 2, FG-MoE incorporates four classes of classical filters—Sobel, Wavelet, Schmid, and Gabor—into its expert library, each emphasizing different tampering traces (e.g., edges, textures, directional patterns, and multi-scale structures). These heterogeneous experts provide structured physical priors, which are dynamically combined through a Multi-Modal Auxiliary Gating Module. The gating mechanism fuses semantic cues from text queries with visual features, adaptively weighting expert contributions to achieve semantic-guided modeling of physical clues. The heatmap in Fig. 1 illustrates that the gating network can effectively assign expert weights based on contextual semantics, enabling a transition from passive pixel-level perception to active, semantically-driven reasoning.

The main contributions of this paper are as follows:

- We introduce FG-MoE, a heterogeneous mixture-of-experts framework that embeds classical signal-processing priors (Sobel, Schmid, Gabor, Wavelet) into deep multimodal learning via a semantic-guided gating mechanism.
- A task-structured expert paradigm is proposed to leverage complementary priors—boundaries, textures, directions, and scales to enhance the detection of subtle forgeries.
- Extensive experiments on the Real-RSCM benchmark show that FG-MoE achieves competitive results, outperforming or matching most state-of-the-art methods in Overall Accuracy (OA) and Average Accuracy (AA).

2. METHODOLOGY

2.1. Problem Formulation

To accurately identify the source and tampered regions and to conduct precise reasoning by integrating foreground and background knowledge, a Filter-Group Mixture-of-Experts (FG-MoE) model was developed. The FG-MoE model employs a two-stage training process. First, the forgery detection network is trained to generate hierarchical representations of the source region, tampered region, and background. Second, corresponding experts are designed using a variety of differentiated filters to capture subtle visual cues.

2.2. Forgery Segmentation and Identification

In scenarios where potential tampered regions exist, a novel reconstruction network is employed to generate visual representations. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, instance segmentation and identification are performed, resulting in a hierarchical visual representation $\mathbf{F}^{\text{seg}} \in \mathbb{R}^{H'\times W'\times C}$ that includes both the source and tampered regions. Here, C denotes the feature dimension. The final step involves fusing the obtained $\mathbf{F}^{\text{seg}}$ with the original image $\mathbf{F}^{i}$ to generate the final

prompted visual representation $\mathbf{F}^v$. The process is detailed as follows:

$$\mathbf{F}^v = \mathrm{FFN}(\mathrm{Cat}(\mathbf{F}^i, \mathbf{F}^{seg})). \quad (1)$$

Here, FFN denotes a feedforward neural network, and Cat represents the concatenation operation.

2.3. Filter-Group Mixture-of-Experts

The MoE consists of a multimodal gating network $G_{\mathcal{M}}$ and N filter-based experts E_N^{filter}. In FG-MoE, visual features $\mathbf{F}^v$ and textual features $\mathbf{F}^t$ compute similarity scores via cross-attention, which the gating network uses to route experts. The gating weights are defined as:

$$G_{\mathcal{M}}(F^v, F^t) = \mathrm{Softmax}(\mathrm{topK}(\mathrm{Att}_{\mathcal{C}}(F^v, F^t) \cdot W_{\mathcal{M}})), \quad (2)$$

where $W_{\mathcal{M}}$ is a learnable matrix and the top K outputs are normalized by $softmax$. The visual input $\mathbf{F}^v$ is processed by specialized filters: Sobel (boundaries), Schmid (textures), Gabor (directionality), and Wavelet (multi-scale objects). Each expert E_i^{filter} applies a shared two-layer structure with linear layers and customized activations—Mish (Sobel), SiLU (Schmid/Wavelet), and GeLU (Gabor)—to encourage complementary learning. The final FG-MoE output $\mathbf{I}_{\mathcal{F}}$ is a weighted sum of expert outputs:

$$I_{\mathcal{F}} = \sum_{i=1}^{N} G_{\mathcal{M}}(F^v, F^t)_i E_i^{\mathrm{filter}}(F^v). \quad (3)$$

2.4. Loss Function

The overall loss $\mathcal{L}$ consists of three components: a tamper detection loss $\mathcal{L}_{\text{rmse}}$, a VQA loss $\mathcal{L}_{\text{vqa}}$, and an expert balance regularization $\mathcal{L}_{\text{balance}}$. The RMSE loss measures reconstruction error between predicted and ground-truth masks:

$$\mathcal{L}_{rmse} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{F}}^o - \mathbf{F}^o)^2}, \quad (4)$$

where n is the number of samples, $\hat{\mathbf{F}}^o$ the ground truth, and $\mathbf{F}^o$ the prediction. The VQA loss is computed with standard cross-entropy:

$$\mathcal{L}_{vqa} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i), \quad (5)$$

where y_i and $\hat{y}_i$ denote ground-truth labels and predicted probabilities.To encourage balanced expert utilization, we add:

$$\mathcal{L}_{balance} = CV(\mathrm{Importance}(\mathbf{F}^v, \mathbf{F}^t))^2, \quad (6)$$

where *Importance* is the cumulative contribution of each expert across a batch, and CV is the coefficient of variation.The final loss is:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{rmse} + (1-\alpha) \cdot \mathcal{L}_{vqa} + \mathcal{L}_{balance}. \quad (7)$$

with trade-off coefficient α.

3. EXPERIMENTS

3.1. Experimental Setup and Comparative Experiments

We evaluate model performance using Overall Accuracy (OA) and Average Accuracy (AA), the latter mitigating class imbalance effects. Experiments were conducted on the Real-RSCM dataset with a 70/15/15 train/validation/test split, kept consistent across all methods. Our model was implemented in PyTorch with BERT as the text encoder and ViT-B as the visual backbone. Training lasted 30 epochs, with a tampered mask reconstruction module active for the first 20 epochs under a cosine annealing learning rate schedule. To ensure fairness, all baselines were initialized with the same pretrained weights and trained under identical settings. We compared FG-MoE with thirteen state-of-the-art models, including five general-purpose VQA approaches—SAN [16], MAC [17], MCAN [18], DVQA [19], and BLIP-2 [20]—and eight remote sensing-specific VQA models: RSVQA [9], RSIVQA [21], FEH [22], MQVQA [23], SGA [24], EarthVQA [11], STMA [14], and CM-MMoE [15].

As shown in Table 1, all methods achieve strong performance on factual questions (Q1–Q3), but accuracy drops significantly for independent (Q4–Q10) and related (Q11–Q14) questions involving subtle tampering cues. Although STMA and CM-MMoE demonstrate robustness in modeling source–tamper discrepancies, their performance remains limited when handling complex reasoning. In contrast, FG-MoE achieves competitive results, ranking first in 12 of 14 categories and achieving an overall accuracy of 92.95% and an average accuracy of 91.47%. These results highlight that integrating structured signal processing priors into a mixture-of-experts framework enhances sensitivity to physical cues, yielding both stronger discriminative ability and improved interpretability for complex forgery analysis.

3.2. Ablation Experiments

Expert Structure Ablation. We ablated expert number, composition, and prior integration (Fig. 3). Accuracy of homogeneous experts (E_0) peaked at three, with larger counts causing redundancy, showing capacity is not the bottleneck. Heterogeneous configurations outperformed five homogeneous variants (e.g., sobel$\times$5, schmid$\times$5), validating the benefit of complementary priors. Replacing standard with filter-based experts yielded non-monotonic gains, suggesting improvements stem from domain priors rather than model size.

Loss Function Ablation. Training used $\mathcal{L}_{vqa}$, $\mathcal{L}_{rmse}$, and $\mathcal{L}_{balance}$, with $\mathcal{L}balance$ converging quickly and left unweighted. A single hyperparameter α adjusted the trade-off between $\mathcal{L}_{vqa}$ and $\mathcal{L}_{rmse}$; results were stable across settings, with $\alpha = 0.3$ giving the best OA and AA (Table 2).

Top-K Gating. We varied the number of activated experts K (Table 3). Too few caused under-utilization, too many redundancy; $K = 3$ achieved the best balance.

Table 1. Performance comparison with state-of-the-art methods on the Real-RSCM test set. The best results are in **bold** and the second-best are underlined.

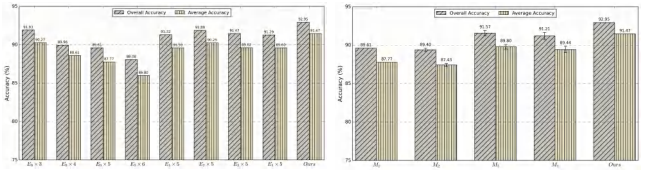
Method	Basic Questions			Independent Questions						Related Questions					OA	AA
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14		
*General VQA Methods																
SAN (CVPR, 2016)[16]	85.57	96.31	98.38	34.75	65.40	31.39	65.21	89.97	94.00	36.03	65.16	40.72	58.87	55.56	68.48	65.51
MAC (ICLR, 2018)[17]	87.94	**98.62**	99.15	49.54	84.27	64.10	85.56	94.06	95.28	82.12	91.12	77.74	88.16	55.05	84.80	82.34
MCAN (CVPR, 2019)[18]	84.44	96.39	98.75	26.18	68.19	35.15	75.23	90.13	92.99	34.86	73.26	71.43	61.74	54.60	71.84	68.81
DVQA (NeurIPS, 2021)[19]	90.47	97.54	99.21	45.84	77.96	73.31	87.93	93.93	97.58	78.65	88.27	71.19	85.39	56.18	84.04	81.67
BLIP-2 (ICML, 2023)[20]	90.92	96.76	99.17	48.24	74.58	48.32	78.52	88.50	90.04	47.93	61.11	58.67	65.38	58.36	74.04	71.89
*Remote Sensing VQA Methods																
RSVQA (TGRS, 2020)[9]	88.63	96.29	98.25	48.19	80.98	60.66	84.86	92.54	95.73	74.04	86.52	42.91	61.60	54.46	78.77	76.12
RSIVQA (TGRS, 2021)[21]	84.16	95.18	98.18	30.03	68.61	25.23	68.39	91.22	94.83	41.19	59.46	34.72	49.46	54.89	67.01	63.97
FEH (TGRS, 2022)[22]	93.23	97.57	99.33	59.29	85.46	78.83	92.72	93.00	96.11	84.28	92.10	61.95	82.21	56.14	86.05	83.73
MQVQA (TGRS, 2023)[23]	89.89	95.26	98.38	61.25	81.29	82.28	90.88	95.18	97.05	62.77	78.63	46.80	62.79	55.51	80.49	78.42
EarthVQA (AAAI, 2024)[11]	87.33	94.76	97.01	56.95	82.37	61.77	83.62	92.79	95.05	84.11	91.12	79.35	87.15	54.11	84.16	81.96
SGA (IGARSS, 2024)[24]	89.40	95.83	96.97	57.45	83.33	63.01	85.08	94.79	97.24	84.52	92.63	79.22	87.84	57.53	85.38	83.21
STMA (ArXiv, 2024)[14]	91.79	97.78	99.05	68.82	87.13	82.28	92.25	97.52	97.90	79.79	90.08	68.85	82.78	55.86	87.08	85.12
CM-MMOE (ICME, 2025)[15]	93.66	97.69	99.34	83.07	92.12	89.57	95.46	98.01	98.69	91.18	95.57	84.25	90.80	**59.57**	92.01	90.64
FG-MoE (Ours)	**94.88**	98.02	**99.41**	**84.62**	**94.14**	**91.57**	**95.71**	**98.23**	**98.83**	**94.35**	**96.28**	**87.31**	**91.11**	56.12	**92.95**	**91.47**

Table 2. Ablation study on the loss function hyperparameter.

α	0.1	0.2	0.3	0.4	0.5	0.6
OA	81.88	91.76	**92.95**	91.75	92.28	91.58
AA	79.06	90.01	**91.47**	90.25	90.78	90.08

Table 3. Ablation Study on the Top-K hyperparameter.

Top-K Count	2	3	4	5
Overall Accuracy	91.80	**92.95**	92.11	92.17
Average Accuracy	90.21	**91.47**	90.43	90.52



(a) MoE for Expert Heterogeneity **(b) MoE with Progressive Prior Integration**

Fig. 3. Ablation study on expert architecture. (a) Comparison of expert number and heterogeneity, where E_0 to E_4 represent the Standard, Sobel, Schmid, Wavelet, and Gabor experts. (b) Progressive integration of filter-based priors: M_1 is the baseline with five Standard experts, while M_2 to M_5 show incremental substitution with filter-guided experts.

3.3. Visualization

To complement quantitative evaluations, we present visualization analyses that reveal how FG-MoE perceives and reasons over tampered remote sensing images in Fig. 4. The Single-Expert Heatmaps illustrate that each filter-guided expert responds distinctly to perturbations such as texture variation and high-frequency noise, highlighting the benefits of expert heterogeneity. The Multi-Filter Heatmap Difference Overlay further exposes inter-expert divergences, which act as robust high-order cues for subtle forgery localization.

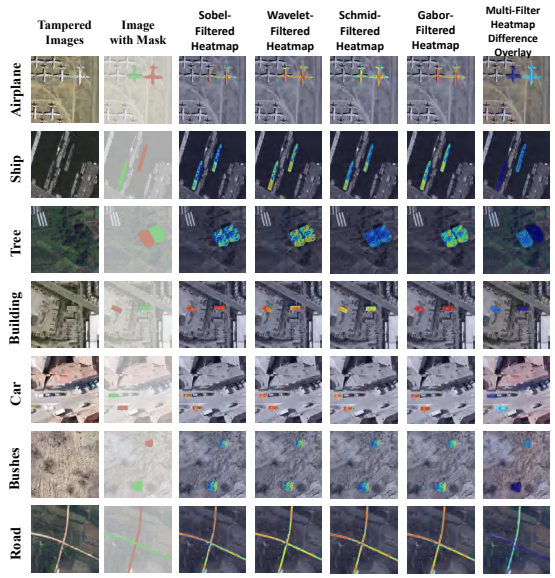


Fig. 4. Visual interpretation of feature-level analysis in FG-MoE. Each row corresponds to a typical tampered case. The first two columns show the tampered image and ground truth. Columns 3–6 present expert-specific response maps (sobel, wavelet, schmid, gabor). The final column overlays the inter-expert discrepancy, revealing subtle inconsistencies indicative of tampering.

4. CONCLUSION

We propose FG-MoE, a mixture-of-experts model that embeds classical signal-processing priors into deep architectures for interpreting remote sensing copy-move forgeries. By leveraging a heterogeneous expert library and a multimodal gating network, the model adaptively reasons over multi-dimensional features. Future work will focus on exploring richer priors and evaluating the framework on broader visual reasoning tasks to further assess its generalizability.

5. ACKNOWLEDGMENTS

This work was supported by the Science and Technology Program of Zhejiang Province (Grant No. 2025C02038), the Natural Science Foundation of Zhejiang Province (Grant No. LTGG23F020004), the Hangzhou Qianjiang Distinguished Expert Program (Grant No. HZ20251217834), and the Research Funds of Hangzhou Institute for Advanced Study, UCAS.

6. REFERENCES

[1] M. Bashar, K. Noda, K. Ohnishi, and N. Mori, "Exploring Duplicated Regions in Natural Images," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2646–2659, 2010.

[2] Richa Singh, Sandeep Verma, Suman Avdhesh Yadav, and S. Vikram Singh, "Copy-move Forgery Detection using SIFT and DWT detection Techniques," in *International Conference on Intelligent Engineering and Management (ICIEM)*, 2022, pp. 338–343.

[3] Yanfen Gan, Junliu Zhong, and Chiman Vong, "A Novel Copy-Move Forgery Detection Algorithm via Feature Label Matching and Hierarchical Segmentation Filtering," *Information Processing & Management*, vol. 59, no. 1, pp. 102783, 2022.

[4] Gyanda Kaushal and Badal Soni, "ORBcmfd: Oriented FAST and Rotated BRIEF Keypoints Based Image Copy-Move Forgery Detection," in *IEEE Students Conference on Engineering and Systems (SCES)*, 2024, pp. 1–5.

[5] Yue Wu, Wael Abd-Almageed, and Prem Natarajan, "BusterNet: Detecting Copy-Move Image Forgery with Source/Target Localization," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 170–186.

[6] Beijing Chen, Weijin Tan, Gouenou Coatrieux, Yuhui Zheng, and Yun-Qing Shi, "A Serial Image Copy-Move Forgery Localization Scheme With Source/Target Distinguishment," *IEEE Transactions on Multimedia*, vol. 23, pp. 3506–3517, 2021.

[7] Lizhi Xiong, Jianhua Xu, Ching-Nung Yang, and Xinpeng Zhang, "CMCF-Net: An End-to-End Context Multiscale Cross-Fusion Network for Robust Copy-Move Forgery Detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 6090–6101, 2024.

[8] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang, "FakeShield: Explainable Image Forgery Detection and Localization via Multi-modal Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2025.

[9] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia, "RSVQA: Visual Question Answering for Remote Sensing Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.

[10] Ze Zhang, Enyuan Zhao, Ziyi Wan, Xinyue Liang, Min Ye, Jie Nie, and Lei Huang, "Frequency Domain Transfer Learning for Remote Sensing Visual Question Answering," *Expert Systems with Applications*, vol. 291, pp. 128395, 2025.

[11] Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ai, and Yanfei Zhong, "EarthVQA: Towards Queryable Earth via Relational Reasoning-Based Remote Sensing Visual Question Answering," *AAAI Conference on Artificial Intelligence*, vol. 38, pp. 5481–5489, Mar 2024.

[12] Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad Shahbaz Khan, and Salman Khan, "GeoPixel: Pixel Grounding Large Multimodal Model in Remote Sensing," in *International Conference on Machine Learning (ICML)*, 2025.

[13] Zhenghang Yuan, Lichao Mou, Zhitong Xiong, and Xiao Xiang Zhu, "Change Detection Meets Visual Question Answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022.

[14] Ze Zhang, Enyuan Zhao, Di Niu, Jie Nie, Xinyue Liang, and Lei Huang, "Copy-move Forgery Detection and Question Answering for Remote Sensing Image," arXiv preprint arXiv:2412.02575, 2024.

[15] Z. Zhang, E. Zhao, Y. Jiang, J. Nie, and X. Y. Liang, "Challenging Dataset and Multi-Modal Gated Mixture of Experts Model for Remote Sensing Copy-Move Forgery Understanding," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2025.

[16] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola, "Stacked Attention Networks for Image Question Answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 21–29.

[17] Drew A. Hudson and Christopher D. Manning, "Compositional Attention Networks for Machine Reasoning," in *International Conference on Learning Representations (ICLR)*, 2018.

[18] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, "Deep Modular Co-Attention Networks for Visual Question Answering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6274–6283.

[19] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu, "Debiased Visual Question Answering From Feature and Sample Perspectives," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models," in *International Conference on Machine Learning (ICML)*, 2023.

[21] Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu, "Mutual Attention Inception Network for Remote Sensing Visual Question Answering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[22] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[23] Meimei Zhang, Fang Chen, and Bin Li, "Multistep Question-Driven Visual Question Answering for Remote Sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[24] Lucrezia Tosato, Hichem Boussaid, Flora Weissgerber, Camille Kurtz, Laurent Wendling, and Sylvain Lobry, "Segmentation-Guided Attention for Visual Question Answering From Remote Sensing Images," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2024, pp. 2750–2754.