EarthVQA: Towards Queryable Earth via Relational Reasoning-Based Remote Sensing Visual Question Answering

Junjue Wang¹, Zhuo Zheng², Zihang Chen¹, Ailong Ma¹, Yanfei Zhong¹*

¹LIESMARS, Wuhan University, 430074, China
²Department of Computer Science, Stanford University, Stanford, CA 94305, USA
{kimgdrone, chenzihang, mailong007, zhongyanfei}@whu.edu.cn, zhuozheng@cs.stanford.edu

Abstract

Earth vision research typically focuses on extracting geospatial object locations and categories but neglects the exploration of relations between objects and comprehensive reasoning. Based on city planning needs, we develop a multi-modal multi-task VQA dataset (EarthVQA) to advance relational reasoning-based judging, counting, and comprehensive analysis. The EarthVQA dataset contains 6000 images, corresponding semantic masks, and 208,593 QA pairs with urban and rural governance requirements embedded. As objects are the basis for complex relational reasoning, we propose a Semantic OBject Awareness framework (SOBA) to advance VQA in an object-centric way. To preserve refined spatial locations and semantics, SOBA leverages a segmentation network for object semantics generation. The object-guided attention aggregates object interior features via pseudo masks, and bidirectional cross-attention further models object external relations hierarchically. To optimize object counting, we propose a numerical difference loss that dynamically adds difference penalties, unifying the classification and regression tasks. Experimental results show that SOBA outperforms both advanced general and remote sensing methods. We believe this dataset and framework provide a strong benchmark for Earth vision’s complex analysis. The project page is at https://Junjue-Wang.github.io/homepage/EarthVQA.

1 Introduction

High-spatial resolution (HSR) remote sensing images can assist us in quickly obtaining essential information (Zvonkov et al. 2023; Xiao et al. 2023). Most research focuses on the perception of object categories and locations, deriving related tasks such as semantic segmentation (Liu et al. 2023), species detection (Zhao et al. 2022), and urban understanding (Shi et al. 2023). However, the existing methods and datasets ignore the relations between the geospatial objects, thus limiting their ability to knowledge reasoning in complex scenarios. Especially in city planning (Bai, Shi, and Liu 2014), the relations between the transportation hubs and schools, water situations around the farmland, and greenery distributions in residential areas are also significant and urgent to be analyzed. Hence, it is necessary to go beyond object perception and explore object relations, bridging the gap between information and comprehensive knowledge (Li and Krishna 2022).

Visual question answering (VQA) aims to answer customized questions by searching for visual clues in the provided image. Since linguistic questions determine the task properties, the algorithms are flexible and can be developed for reasoning required answers. Recently, preliminary VQA datasets and methods have emerged in the remote sensing field (Lobry et al. 2020; Zheng et al. 2021; Rahmenoonfar et al. 2021). However, most of these researches have the following drawbacks: 1) As for most datasets, QA pairs are automatically labelled based on existing data, such as Open Street Map (OSM) and classification datasets. Most tasks are simple counting and judging questions with no relational reasoning required. The automatic QA pairs do not match actual needs, limiting their practicalities. 2) The development of the remote sensing VQA model lags, and most research directly fuses the global visual and language features to predict the final answers. They ignore the local semantics and relations, which are unsuitable for the complex reasoning of multiple geospatial objects. To this end, we propose a multi-modal multi-task VQA dataset and a semantic object awareness framework to advance complex remote sensing VQA tasks. Main contributions are as follows:

1) We propose the EarthVQA dataset with triplet samples (image-mask-QA pairs). The 208,593 QA pairs encompass six main categories¹. EarthVQA features diverse tasks from easy basic judging to complex relation reasoning and even more challenging comprehensive analysis. Specifically, the residential environments, traffic situations, and renovation needs of waters and unsurfaced roads are explicitly embedded in various questions.

2) To achieve relational reasoning-based VQA, we propose a semantic object awareness framework (SOBA). SOBA utilizes segmentation visual features and pseudo masks to generate pixel-level features with accurate locations. The object awareness-based hybrid attention models the relations for object-guided semantics and bidirectionally aggregates multi-modal features for answering.

3) To add distance sensitivity for regression questions, we

*Corresponding author
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹basic judging, basic counting, relational-based judging, relational-based counting, object situation analysis, and comprehensive analysis.
propose a numerical difference (ND) loss. The dynamic ND penalty is seamlessly integrated into cross-entropy loss for the regression task. ND loss introduces the sensitivity of numerical differences into the model training.

2 Related Work

General visual question answering. The vanilla VQA model (Antol et al. 2015) includes three parts: a convolutional neural network (CNN), a long-short term memory (LSTM), and a fusion classifier. Specifically, CNN extracts visual features for input images, and LSTM embeds the language features for the questions. Global features are interacted in the fusion classifier and finally generate the answer. Based on this architecture, more powerful encoders and fusion modules were proposed. To obtain local visual features, the bottom-up top-down attention (BUTD) mechanism (Anderson et al. 2018) introduced objectness features generated by Faster-RCNN (Ren et al. 2015) pretrained on Visual Genome (Krishna et al. 2017) data. For computational efficiency, a recurrent memory, attention, and composition (MAC) cell (Hudson and Manning 2018) was designed to explicitly model the relations between image and language features. Similarly, the stacked attention network (SAN) (Yang et al. 2016) located the relevant visual clues guided by question layer-by-layer. By combining objectness features with attention, the modular co-attention network (MCAN) (Yu et al. 2019) adopted a transformer to model intra- and inter-modality interactions. To alleviate language biases, D-VQA (Wen et al. 2021) applied an unimodal bias detection module to explicitly remove negative biases. BLIP-2 (Li et al. 2023) and Instruct-BLIP (Dai et al. 2023) bridge the large pre-trained vision and language models using the Q-Former, addressing VQA as a generative task. Besides, many advanced VQA methods (Marino et al. 2021) eliminate statistical bias by accessing external databases.

Remote sensing visual question answering. The remote sensing community has some early explorations including both datasets and methods. The QA pairs of the RSVQA dataset (Lobry et al. 2020) are queried from OSM, and each rural image has 29 QAs. Following the balanced distribution of QA pairs.

The EarthVQA dataset was designed for disaster assessment, mainly concerned with the inundation of roads and buildings.

Remote sensing algorithms are mainly modified from general methods, for example, RSVQA is based on vanilla VQA (Antol et al. 2015). RSIVQA (Zheng et al. 2021) designed a mutual attention component to improve interactions for multi-modal features. CDVQA (Yuan et al. 2022) introduced VQA into change detection task. We novelly introduce pixel-level features for the guidance of VQA tasks, making it suitable for scenes with compact objects.

3 EarthVQA Dataset

The EarthVQA dataset was extended from the LoveDA dataset (Wang et al. 2021), which encompasses 18 urban and rural regions from Nanjing, Changzhou, and Wuhan. LoveDA dataset provides 5987 HSR images and semantic masks with seven common land-cover types. There are three significant revisions: 1) Quantity expansion. 8 urban and 5 rural samples are added to expand capacity to 6000 images (WorldView-3 0.3m). 2) Label refinement. ‘playground’ class was added as an important artificial facility, and some errors were revised for semantic labels. 3) Addition of QA pairs. We added 208,593 QA pairs to introduce VQA tasks for city planning. Each urban image has 42 QAs and each rural image has 29 QAs. Following the balanced distribution (Wang et al. 2021), train set includes 2522 images with 88166 QAs, val set includes 1669 images with 57202.
QAs, and test set includes 1809 images with 63225 QAs. **Annotation procedure.** EarthVQA currently does not involve ambiguous questions such as geographical orientations. As for ‘Are there any intersections near the school?’ in Figure 2(a), by judging the topology, the recognized Road#1 and Road#2 firstly form Intersection#5. Similarly, Ground#4 and Building#3 jointly form the scene of School#6. We use the ArcGIS toolbox to calculate the polygon-to-polygon distance between School#6 and Intersection#5, and obtain 94.8m < 100m. Hence, the final answer is ‘Yes’. Each step has fixed thresholds and conditions. **Statistics for questions.** As is shown in Figure 2(b), urban and rural scenes have common and unique questions according to the city planning demands. The number of questions for urban and rural is balanced, eliminating geographical statistical bias. Basic questions involve the statistics and inference of a certain type of object, i.e., ‘What is the area of the forest?’. Relational-based questions require semantic or spatial relational reasoning between different objects. Comprehensive analysis focuses on more than three types of objects, including a summarization of traffic facilities, water sources around agriculture, land-use analysis, etc.

**Statistics for answers.** As shown in Figure 2(c), we selected the top 15 most frequent answers from 166 unique answers in the dataset. Similar to the common VQA datasets, the imbalanced distributions of answers bring more challenges when faced with the actual Earth environment.

### 4 Semantic Object Awareness Framework
To achieve efficient relational reasoning, we design the SOBA framework for complex city scenes. SOBA includes a two-stage training: 1) semantic segmentation network training for generating visual features and pseudo masks; and 2) hybrid attention training for reasoning and answering.

#### 4.1 Semantic Segmentation for Visual Features
Faced with HSR scenes containing multiple objects, we newly adopt a segmentation network for refined guidance. For an input image \( I \in \mathbb{R}^{H \times W \times 3} \), we utilize the encoder outputs \( F^v \in \mathbb{R}^{H' \times W' \times C} \) as the visual features. \( C \) denotes feature dimension and \( H' = \frac{H}{32}, W' = \frac{W}{32} \) according to common settings. Pseudo semantic output \( M^v \in \mathbb{R}^{H \times W} \) is also adopted for object awareness. Compared with the existing Faster-RCNN based algorithms (Yu et al. 2019; Anderson et al. 2018) which averages box features in one vector, the pixel-level visual features preserve the locations and semantic details inside objects. This contributes to the modeling of various compact objects in HSR scenes.

#### 4.2 Object Awareness-Based Hybrid Attention
Guided by questions and object masks, Object awareness-based hybrid attention reasons visual cues for final answers. As is shown in Figure 3, there are three components: 1) object-guided attention (OGA), 2) visual self-attention (VSA), and 3) bidirectional cross-attention (BCA).

**OGA for object aggregation.** Because segmentation output has object details \( M^v \) (including categories and boundaries), it is adopted to explicitly enhance visual features. OGA is proposed to dynamically weight \( F^v \) and \( M^v \) from the channel dimension. Using the nearest interpolation, \( M^v \) is firstly resized into the same size as \( F^v \). One-hot encoding followed with a pre-convolutional embedding then serializes the object semantics. The embedding contains a \( 3 \times 3 \) convolution, a batch normalization, and a ReLU. They are concatenated to obtain object-guided features \( F^v_q \) as inputs for OGA. The reduction and reverse projections further refine the features dimensionally. After activation, we use the refined features to calibrate subspaces of \( F^v_q \) from the channel dimension.

**VSA for feature enhancement.** To capture long-distance relations between geospatial objects, VSA (Dosovitskiy et al. 2021) hierarchically transforms the refined features. VSA includes \( N_c \) transformer blocks, and each includes a multi-head self-attention (MSA) and a feed-forward network (FFN). The refined features are reduced by a \( 1 \times 1 \) convolution and reshaped to generate patches \( X \in \mathbb{R}^{H' \times W' \times d_m}, P = \frac{H'}{32} \times \frac{W'}{32} \) denotes token size and \( d_m \) is hidden size.

At each block \( i \), features are transformed into a triplet: \( Q = X^{i-1} W^q, K = X^{i-1} W^k, V = X^{i-1} W^v \), where \( W^q, W^k, W^v \in \mathbb{R}^{d_m \times d_v} \) denote the weights of three linear projections and \( d_v = d_m / M \) is the reduction dim of each head. The self-attention firstly calculates the similarities between each patch and then weight their values: \( Att(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_v}}\right)V \). MSA repeats the attention operation \( M \) times in parallel and concatenates outputs. Finally, outputs are fused by a linear projection. Formally, \( MSA(Q, K, V) = \text{Concat}(h_1, ..., h_M)W^c \), where

![Figure 2: Details of questions and answers in EarthVQA dataset. Each urban image has a set of 42 questions and each rural image has a set of 29 questions, ensuring relatively balanced for each question. The imbalanced distributions of answers bring more challenges when faced with the actual Earth environment.](image-url)
4.3 Object Counting Enhanced Optimization

VQA tasks include both classification and regression (object counting) questions. However, existing methods regard them as a multi-classification task, which is processed with cross-entropy (CE) loss. Eq. (2) represents that CE loss is insensitive to the distance between predicted value and true value, and is therefore not suitable for the regression task.

\[ CE(\hat{p}, \hat{y}) = -\hat{y} \odot \log(\hat{p}) + \gamma (\hat{p} - 1 + \hat{y}) \]

For regression tasks, a different penalty is required. We introduce the following penalty for the regression task.

\[ \text{diff} = \frac{1}{1 + \exp(y_{\text{pred}} - y_{\text{gt}})} \]

\[ d = a \cdot \text{diff} \]

\[ ND(\hat{p}, \hat{y}) = -(1 + d) \hat{y} \odot \log(\hat{p}) \]

This formulation unifies classification and regression objectives into one optimization framework. \( \alpha \) controls the overall penalty for regression tasks compared to classification tasks. \( \gamma \) determines the sensitivity of regression penalty to numerical differences. As the \( \alpha \) increases, the overall penalty increases, meaning that optimization focuses more on regression tasks. With \( \alpha = 0 \), the ND loss degenerates into the original CE loss and the penalty is constant (d = 0 when \( |y_{\text{diff}}| \in [0, +\infty) \)). The sensitivity of the regression penalty increases as \( \gamma \) increases, and when \( \gamma > 1 \), the penalty curve changes from concave to convex.
5 Experiments

Evaluation metrics. Following common settings (Yu et al. 2019), we adopt the classification accuracy and root-mean-square error (RMSE) as evaluation metrics. Especially, RMSE is used to evaluate counting tasks. We use mean Union over Intersection (mIoU) to report semantic segmentation performance. All experiments were performed under PyTorch framework using one RTX 3090 GPU.

Experimental settings. For comparison, we selected eight general (SAN (Yang et al. 2016), MAC (Hudson and Manning 2018), BUTD (Anderson et al. 2018), BAN (Kim, Jun, and Zhang 2018), MCAN (Yu et al. 2019), D-VQA (Wen et al. 2021), BLIP-2 (Li et al. 2023), Instruct-BLIP (Dai et al. 2023)) and two remote sensing (RSVQA (Lobby et al. 2020), RISVQA (Zheng et al. 2021)) VQA methods. Because MCAN, BUTD, BAN, and D-VQA need object guidance, we adopt visual features from Semantic-FPN (Kirillov et al. 2019) fairly. All VQA models were trained for 40k steps with a batch size of 16. We set the two-layer LSTM with the hidden size of 384 and ResNet50 as default. As for large vision-language models, BLIP-2 and Instruct-BLIP trained Q-Former following their original settings. The vision encoder adopts ViT-g/14 and language decoder is FlanT5x. Following (Wang et al. 2021), Semantic-FPN was trained for 15k steps using the same batch size, generating visual features and semantic masks. Segmentation augmentations include random flipping, rotation, scale jittering, and cropping for 512 × 512 patches. We used Adam solver with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was set to $5e^{-5}$, and a ‘poly’ schedule with a power of 0.9 was applied. The hidden size of the language and image features was $d_m = 384$. The number of heads $M$ is set to 8, and the numbers of layers in self- and cross-attention modules are $N_E = N_D = 3$. We set $\alpha = 1$ and $\gamma = 0.5$ for ND loss.

5.1 Comparative Experiments

Main comparative results. Thanks to the diverse questions, EarthVQA can measure multiple perspectives of VQA models. Table 1 shows that all methods achieve high accuracies on basic judging questions. The models with pixel-level visual features obtain higher accuracies, especially for the counting tasks. This is because the semantic locations provide more spatial details, which benefits the object statistics.

<table>
<thead>
<tr>
<th>Method</th>
<th>Promp.</th>
<th>$\uparrow$Accuracy(%)</th>
<th>$\uparrow$OA(%)</th>
<th>$\downarrow$RMSE</th>
<th>$\downarrow$OR</th>
<th>Param. (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAN</td>
<td>×</td>
<td>87.59</td>
<td>76.26</td>
<td>59.23</td>
<td>55.00</td>
<td>42.72</td>
</tr>
<tr>
<td>MAC</td>
<td>×</td>
<td>82.89</td>
<td>72.53</td>
<td>55.86</td>
<td>46.32</td>
<td>42.30</td>
</tr>
<tr>
<td>BUTD</td>
<td>✓</td>
<td>90.01</td>
<td>77.16</td>
<td>60.95</td>
<td>56.29</td>
<td>42.29</td>
</tr>
<tr>
<td>BAN</td>
<td>✓</td>
<td>89.81</td>
<td>77.58</td>
<td>63.71</td>
<td>55.67</td>
<td>45.06</td>
</tr>
<tr>
<td>MCAN</td>
<td>✓</td>
<td>89.65</td>
<td>79.83</td>
<td>63.16</td>
<td>57.28</td>
<td>43.71</td>
</tr>
<tr>
<td>D-VQA</td>
<td>✓</td>
<td>89.73</td>
<td>77.38</td>
<td>63.99</td>
<td>55.14</td>
<td>43.20</td>
</tr>
<tr>
<td>BLIP-2</td>
<td>✓</td>
<td>88.13</td>
<td>70.26</td>
<td>58.58</td>
<td>42.72</td>
<td>28.54</td>
</tr>
<tr>
<td>Instruct- BLIP</td>
<td>✓</td>
<td>89.67</td>
<td>76.96</td>
<td>63.34</td>
<td>59.72</td>
<td>45.68</td>
</tr>
<tr>
<td>RSVQA</td>
<td>×</td>
<td>82.43</td>
<td>70.68</td>
<td>55.53</td>
<td>42.45</td>
<td>35.46</td>
</tr>
<tr>
<td>RISVQA</td>
<td>×</td>
<td>85.32</td>
<td>75.01</td>
<td>56.63</td>
<td>51.55</td>
<td>39.25</td>
</tr>
<tr>
<td>SOBA (ours)</td>
<td>✓</td>
<td>89.63</td>
<td>82.64</td>
<td>80.17</td>
<td>67.86</td>
<td>61.40</td>
</tr>
</tbody>
</table>

Table 1: Compared results with other VQA methods on EarthVQA^best

Compared with advanced methods, SOBA achieves the best overall performances with similar or lower complexity.

<table>
<thead>
<tr>
<th>Object Guidance</th>
<th>Att. Type</th>
<th>$\uparrow$OA(%)</th>
<th>$\downarrow$OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only Concat</td>
<td>-</td>
<td>77.61</td>
<td>0.856</td>
</tr>
<tr>
<td>+SA</td>
<td>S</td>
<td>77.72</td>
<td>0.861</td>
</tr>
<tr>
<td>+SCSE</td>
<td>C&amp;S</td>
<td>77.89</td>
<td>0.854</td>
</tr>
<tr>
<td>+CBAM</td>
<td>C&amp;S</td>
<td>77.95</td>
<td>0.857</td>
</tr>
<tr>
<td>+SE</td>
<td>C</td>
<td>78.02</td>
<td>0.853</td>
</tr>
<tr>
<td>+GC</td>
<td>C</td>
<td>78.03</td>
<td>0.847</td>
</tr>
<tr>
<td>+OGA (ours)</td>
<td>C</td>
<td>78.14</td>
<td>0.839</td>
</tr>
</tbody>
</table>

Table 2: Compared results with other attention mechanisms. ‘C’ and ‘S’ denote channel and spatial attention.

Object guided attention. OGA introduces object semantics into visual features and we compare it with related variants. Table 2 shows compared results for spatial, channel, and combined attentions, i.e., SA(Woo et al. 2018), SCSE(Roy, Navab, and Wachinger 2018), CBAM(Woo et al. 2018), SE(Hu, Shen, and Sun 2018), GC(Cao et al. 2019). Channel attentions bring more stable improvements than spatial attentions. Because pseudo masks and visual features are concatenated dimensionally, spatial attentions are hard to calibrate the subspaces of visual features and object masks. Channel attentions enhance key object semantics and weaken uninterested background features. Hence, our OGA abandoned spatial attention and achieved the best accuracies.

<table>
<thead>
<tr>
<th>Cross-Attention</th>
<th>Query</th>
<th>$\uparrow$OA(%)</th>
<th>$\downarrow$OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-way (vanilla)</td>
<td>LLLLL</td>
<td>77.11</td>
<td>0.977</td>
</tr>
<tr>
<td></td>
<td>VVVVV</td>
<td>77.53</td>
<td>0.880</td>
</tr>
<tr>
<td>Bidirectional</td>
<td>LLL-VVV</td>
<td>77.57</td>
<td>0.867</td>
</tr>
<tr>
<td></td>
<td>VVV-LLL</td>
<td>78.14</td>
<td>0.839</td>
</tr>
</tbody>
</table>

Table 3: Compared results between one-way (vanilla) and bidirectional cross-attention. ‘V’ and ‘L’ denote visual and language features, respectively.

One-way vs. bidirectional cross-attention. Existing transformer-based methods (Yu et al. 2019; Cascante-Bonilla et al. 2022) utilize one-way (vanilla) attention to perform interactions, where visual features are only treated...
as queries. In contrast, we further gather enhanced visual features via the keywords (language features as queries), simulating the human process of finding visual cues. As cross-attention consists of six transformer blocks, we compare the different combinations. Table 3 shows that in one-way attention, querying visual features outperforms querying the language features. This is because visual features are more informative, and their enhancement brings more improvements. Bidirectional attention outperforms one-way structure due to more comprehensive interactions.

<table>
<thead>
<tr>
<th>VSA</th>
<th>BCA</th>
<th>Promp.</th>
<th>OGA</th>
<th>ND</th>
<th>↑OA(%)</th>
<th>↓OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>72.55</td>
<td>1.509</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>73.78</td>
<td>1.520</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>74.91</td>
<td>1.128</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>77.30</td>
<td>0.866</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>77.54</td>
<td>0.859</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>78.14</td>
<td>0.839</td>
</tr>
</tbody>
</table>

Table 4: Architecture ablation study

5.2 Module Analysis

Architecture of SOBA. SOBA was disassembled into five sub-modules: 1) VSA, 2) BCA, 3) semantic features, 4) OGA, and 5) ND loss. Table 4 shows that each module enhances the overall performance in distinct ways. BCA produces a more significant improvement than VSA, and they complement each other (jointly obtaining OA=74.91%). OGA further improves the OA by explicitly adding the objectness semantics. ND loss significantly boosts the counting performance from the aspect of optimization. All modules are compatible with each other within the SOBA framework.

Encoder variants. Table 5 shows the effects brought by segmentation networks with advanced CNN and Transformer encoders, i.e., HRNet (Wang et al. 2020), Swin Transformer (Liu et al. 2021), Mix Transformer (Xie et al. 2021), ConvNeXt (Liu et al. 2022). SOBA is compatible with the mainstream encoders and VQA performance is stable at a high level (OA>77.22%). Although MiT-B3 achieves lower segmentation accuracies than HR-W40, their features provide similar VQA performances. As for similar segmentation architectures, larger encoders (Swin-S and ConvX-S) outperform better than smaller encoders (Swin-T and ConvX-T) in segmentation and VQA tasks. With Wikipedia’s external knowledge, pretrained BERT-Base (Kenton and Toutanova 2019) brings stable improvements. With abundant computing power and time, larger encoders are recommended.

Bidirectional cross-attention variants. We explored BCA variants with different orders of query, i.e., V and L were processed alternately, cascade, and parallel. Table 6 shows that cascade structure VVV-LVV achieves the best accuracies. VVV hierarchically aggregates language features to enhance visual features, and LLL compresses the visual features to supplement language features. Compared with LLL, first considering VVV retains the most information. Hence, VVV-LLL represents the integration process from details to the whole, which conforms to human perception (Savage 2019). Parallel structure obtains a sub-optimal accuracy, and frequent alternation of cross-attentions may lead to feature confusion.

<table>
<thead>
<tr>
<th>Query</th>
<th>↑OA(%)</th>
<th>Optim.</th>
<th>↑OA(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LV-LV-LV</td>
<td>77.31</td>
<td>CE</td>
<td>77.54</td>
</tr>
<tr>
<td>VL-VL-VL</td>
<td>77.58</td>
<td>DIW</td>
<td>77.38</td>
</tr>
<tr>
<td>LLL-LVV</td>
<td>77.57</td>
<td>Focal</td>
<td>77.57</td>
</tr>
<tr>
<td>VVV-LLV</td>
<td>78.14</td>
<td>OHEM</td>
<td>77.85</td>
</tr>
<tr>
<td>Parallel</td>
<td>77.98</td>
<td>SOM</td>
<td>77.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ND</td>
<td>78.14</td>
</tr>
</tbody>
</table>

Table 6: BCA variants

Table 7: Optim analysis

Optimization analysis. We compare ND loss with similar optimization algorithms designed to address the sample imbalance problem, including 1) dynamic inverse weighting (DIW) (Rajpurkar et al. 2017), 2) Focal loss (Lin et al. 2017), 3) online hard example mining (OHEM) (Shrivastava, Gupta, and Girshick 2016), 4) small object mining (SOM) (Ma et al. 2022). In Table 7, Focal loss obtains better performance by adaptively balancing weights of easy and hard examples. DIW failed to exceed the CE due to its extreme weighting strategies. OHEM dynamic focuses on hard samples during the training, slightly improving OA (±0.31%). These optimization algorithms only focus on sample imbalances but are not sensitive to numerical distance. They inherently cannot contribute to regression tasks. In contrast, ND loss shows excellent performances on both classification and regression tasks.

Figure 4: Experimental results with varied α and γ for ND loss. The optimal values range from 0.125 to 1.25 with wide ranges of hyperparameters selection. The mean values and standard deviations are reported after five runs.
5.3 Hyperparameter Analysis for ND Loss

As ND loss introduces two hyperparameters, $\alpha$ controls overall penalty and $\gamma$ determines sensitivity to numerical differences. In order to evaluate their effects on performances, we individually vary $\alpha$ and $\gamma$ from 0 to 2, and the results are reported in Figure 4. Compared with CE loss, the additional difference penalty can bring stable gains. The suitable value of $\alpha$ ranges from 0.125 to 1.25 and reaches the highest OA at 1. When $\alpha > 1.25$, the performance drops because the large loss will bring instability during training. When $\alpha$ is fixed at 1, the optional $\gamma$ also ranges from 0.125 to 1.25, and OA floats between 77.99% and 78.14%. When $\gamma > 1$, the influence curve changes from concave to convex, resulting in a significant increase in difference penalties. The model performance is not very sensitive to the hyperparameters introduced by ND loss, which reflects high fault tolerance and robustness. Overall, our ND loss is superior to the CE baseline, with wide ranges of hyperparameter selection.

5.4 Visualizations on Cross-Attention

To analyze the mechanism of multi-modal feature interaction, we visualize the attention maps in each layer of BCA according to different queries. The question in Figure 6(a) is ‘How many intersections are in this scene?’, and ‘intersections’ is selected as a query word. The first attention map shows some incorrect activations on the scattered roads and playground tracks. However, as the layer deepens, BCA successfully reasons the right spatial relation for the key roads, and the attention map focuses on the intersection in the upper left corner. Similarly, Figure 6(b) shows another example, which displays the process of gradually attending to the ‘residential’ area. The third example shows a rural scene, and we select ‘water’ to query the visual features. The attention map initially focuses on some trees and waters due to their similar spectral values. Then the correct waters are enhanced, and uninterested trees are filtered out.

6 Conclusion

To go beyond information extraction, we introduce the VQA to remote sensing scene understanding, achieving relational reasoning-based reasoning. Based on the city planning needs, we designed a multi-modal and multi-task VQA dataset named EarthVQA. Besides, a two-stage semantic object awareness framework (SOBA) is proposed to advance useful VQA tasks. Extensive experiments demonstrated the superiority of SOBA. Future work will explore the interactions between segmentation and VQA tasks.
Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant Nos. 42325105, 42071350, and 42171336.

References


