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

**Junjue Wang[1], Zhuo Zheng[2], Zihang Chen[1], Ailong Ma[1], Yanfei Zhong[1]***

[1]LIESMARS, Wuhan University, 430074, China
[2]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
{kingdrone, chenzihang, maailong007, 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; Rahnemoonfar 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 categories[1]. 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 generae 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


[1]basic judging, basic counting, relational-based judging, relational-based counting, object situation analysis, and comprehensive analysis.

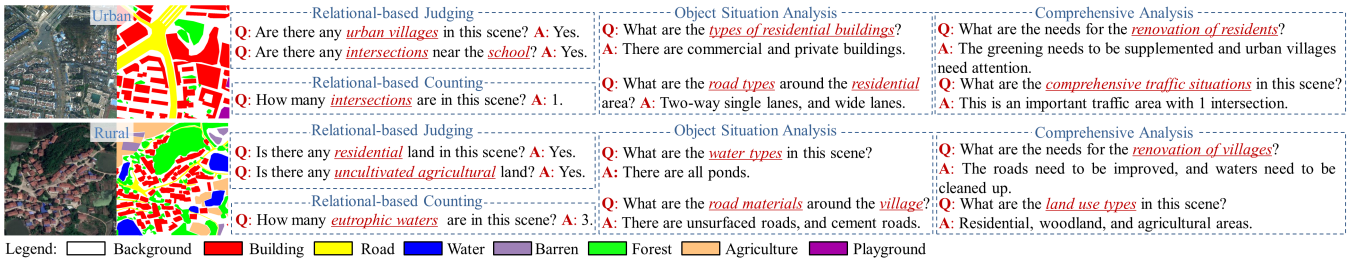



Figure 1: Urban and rural samples (image-mask-QA pairs) from the EarthVQA dataset. The QA pairs are designed to based on city planning needs, including judging, counting, object situation analysis, and comprehensive analysis types. This multi-modal and multi-task dataset poses new challenges, requiring object-relational reasoning and knowledge summarization.

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.

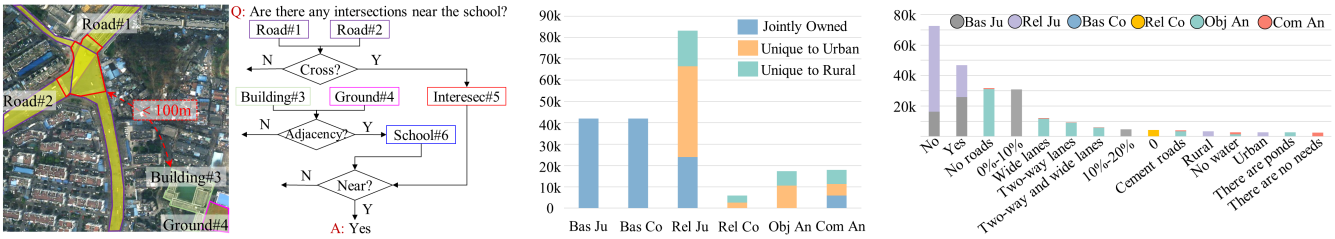
**Remote sensing visual question answering.** The remote sensing community has some early explorations including both datasets and methods. The QA pairs of the RSVQA dataset (Lobry et al. 2020) are queried from OSM, and images are obtained from Sentinel-2 and other sensors. RSIVQA dataset (Zheng et al. 2021) is automatically generated from the existing classification and object detection datasets, *i.e.*, AID (Xia et al. 2017), HRRSD (Zhang et al. 2019), etc. The FloodNet (Rahnemoonfar et al. 2021) dataset was designed for disaster assessment, mainly concerned with the inundation of roads and buildings.

Compared with these datasets, the EarthVQA dataset has two advantages: **1) Multi-level annotations.** The annotations include pixel-level semantic labels, object-level analysis questions, and scene-level land use types. Supervision from different perspectives advances a comprehensive understanding of complex scenes. **2) Complex and practical questions.** The existing datasets focus on counting and judging questions, which only involve simple relational reasoning about one or two types of objects. In addition to counting and judging, EarthVQA also contains various object analysis and comprehensive analysis questions. These promote complex relational reasoning by introducing spatial or semantic analysis of more than three types of objects. Only basic judging and counting answers are auto-generated from the LoveDA masks. Other reasoning answers (Figure 1) are manually annotated (reasoning distances, layouts, topologies, sub-properties, etc) for city planning needs.

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 improves 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 division (Wang et al. 2021), `train` set includes 2522 images with 88166 QAs, `val` set includes 1669 images with 57202




(a) Annotation procedure of relational-based QA. (b) Statistics of questions. (c) Distributions of 15 most frequent answers.

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.

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 novelly adopt a segmentation network for refined guidance. For an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we utilize the encoder outputs $\mathbf{F}^v \in \mathbb{R}^{H'\times W'\times C}$ as the visual features. $C$ denotes feature dimension and $H' = \frac{H}{32}, W' = \frac{W}{32}$ according to common settings. Pseudo semantic output $\mathbf{M}^v \in \mathbb{R}^{H\times W}$ 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 $\mathbf{M}^v$ (including categories and boundaries), it is adopted to explicitly enhance visual features. OGA is proposed to dynamically weight $\mathbf{F}^v$ and $\mathbf{M}^v$ from the channel dimension. Using the nearest interpolation, $\mathbf{M}^v$ is firstly resized into the same size as $\mathbf{F}^v$. 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 $\mathbf{F}_g^v$ 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 $\mathbf{F}_g^v$ 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 $N_e$ 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 $\mathbf{X} \in \mathbb{R}^{P\times d_m}$. $P = \frac{H}{32} \times \frac{W}{32}$ denotes token size and $d_m$ is hidden size.

At each block $i$, features are transformed into a triplet: $\mathbf{Q} = \mathbf{X}^{i-1}\mathbf{W}^q, \mathbf{K} = \mathbf{X}^{i-1}\mathbf{W}^k, \mathbf{V} = \mathbf{X}^{i-1}\mathbf{W}^v$, where $\mathbf{W}^q$, $\mathbf{W}^k$, $\mathbf{W}^v \in \mathbb{R}^{d_m\times d_v}$ 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: $Att(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_v}})\mathbf{V}$. MSA repeats the attention operation $M$ times in parallel and concatenates outputs. Finally, outputs are fused by a linear projection. Formally, $MSA(\mathbf{Q},\mathbf{K},\mathbf{V}) = Concat(h_1,...,h_M)\mathbf{W}^O$, where

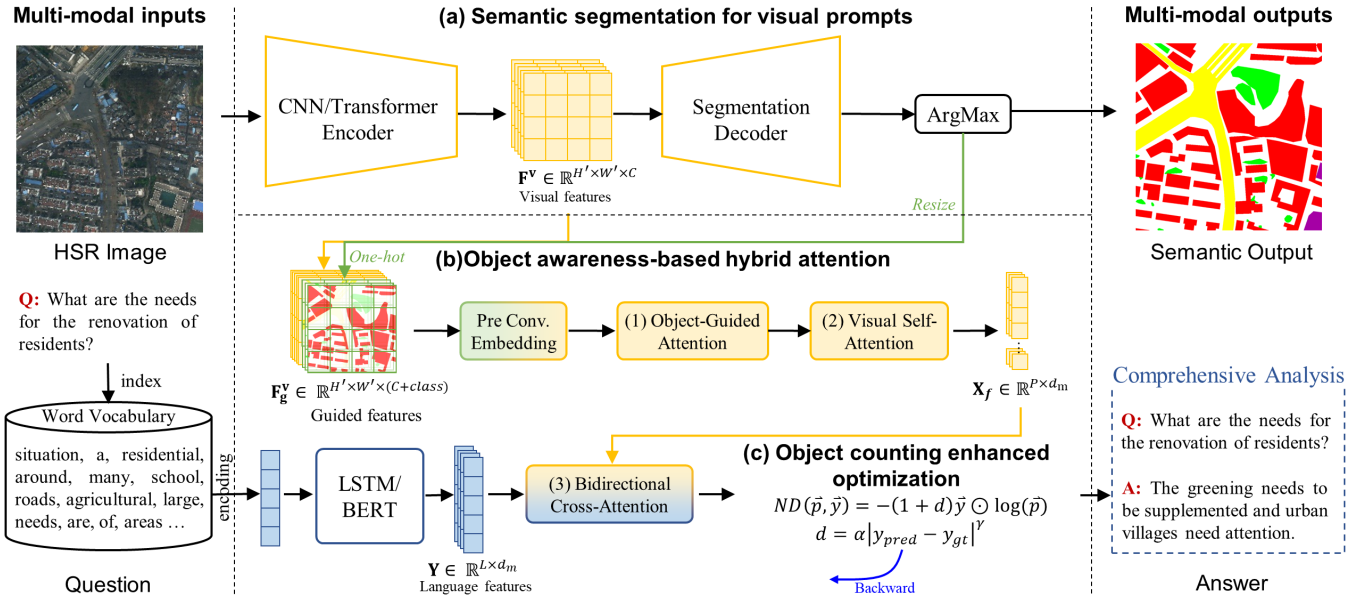



Figure 3: The architecture of SOBA includes (a) deep semantic segmentation for visual features, (b) object awareness-based hybrid attention, and (c) object counting enhanced optimization.

$h_i = Att(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)$ and $\mathbf{W}^O \in \mathbb{R}^{Md_v \times d_m}$ denotes projection weights. MSA models long-distance dependency by calculating the similarities between each geospatial object. FFN consists of two linear transformation layers, and a GELU to improve visual representations. The formulation is shown as $FFN(\mathbf{X}^{i-1}) = GELU(\mathbf{X}^{i-1}\mathbf{W}_1)\mathbf{W}_2$, where $\mathbf{W}_1 \in \mathbb{R}^{d_m \times d_f}, \mathbf{W}_2 \in \mathbb{R}^{d_f \times d_m}$ represent the learnable projection parameters. $d_f$ denotes the hidden size of FFN.

**BCA for multi-modal interaction.** BCA advances the interaction with visual and language features via a bidirectional fusion mechanism. BCA consists of two series of $N_d$ transformer blocks. The first stage aggregates useful language features to enhance visual features $\mathbf{X}_f$ and second stage implicitly models object external relations according to keywords, boosting language features $\mathbf{Y}_f$. The implementation can be formulated as follows:

$$\begin{aligned} \mathbf{Q}_{\mathrm{V}} &= \mathbf{X}\mathbf{W}^q, \mathbf{K}_{\mathrm{L}} = \mathbf{Y}\mathbf{W}^k, \mathbf{V}_{\mathrm{L}} = \mathbf{Y}\mathbf{W}^v \\ \mathbf{X}_f &= \mathrm{Att}(\mathbf{Q}_{\mathrm{V}}, \mathbf{K}_{\mathrm{L}}, \mathbf{V}_{\mathrm{L}}) \\ \mathbf{Q}_{\mathrm{L}} &= \mathbf{Y}\mathbf{W}^q, \mathbf{K}_{\mathrm{V}} = \mathbf{X}_f\mathbf{W}^k, \mathbf{V}_{\mathrm{V}} = \mathbf{X}_f\mathbf{W}^v \\ \mathbf{Y}_f &= \mathrm{Att}(\mathbf{Q}_{\mathrm{L}}, \mathbf{K}_{\mathrm{V}}, \mathbf{V}_{\mathrm{V}}) \end{aligned} \tag{1}$$

Finally, the fused $\mathbf{X}_f$ and $\mathbf{Y}_f$ are used for the final analysis. Compared with previous research (Cascante-Bonilla et al. 2022) which only uses one-way cross-attention, bidirectional attention mechanism hierarchically aggregates multi-modal features by simulating the human process of finding visual cues (Savage 2019). Besides, we have also conducted comparative experiments with alternative cross-attention variants in Table 3 and Table 6.

## 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.

$$CE(\vec{p}, \vec{y}) = -\vec{y} \odot log(\vec{p}) = \sum_{i=1}^{class} -y_i log(p_i) \tag{2}$$

where $\vec{y}$ specifies one-hot encoded ground truth and $\vec{p}$ denotes predicted probabilities. To introduce difference penalty for the regression task, we add a modulating factor $d = \alpha|\mathbf{y}_{diff}|^\gamma = \alpha|\mathbf{y}_{pr} - \mathbf{y}_{gt}|^\gamma$ to CE loss. $\mathbf{y}_{pr}$ and $\mathbf{y}_{gt}$ represent the predicted and ground truth number, respectively. $\alpha \geq 0$ and $\gamma \geq 0$ are tunable distance awareness factors. $d$ represents the distance penalty $d \propto \mathbf{y}_{diff}$. Finally, we design the numerical difference (ND) loss as follows:

$$\begin{aligned} ND(\vec{p}, \vec{y}) &= -(1+d)\vec{y} \odot log(\vec{p}) \\ &= -(1+\alpha|\mathbf{y}_{diff}|^\gamma)\vec{y} \odot log(\vec{p}) \\ &= -(1+\alpha|\mathbf{y}_{pr} - \mathbf{y}_{gt}|^\gamma) \sum_{i=1}^{class} y_i log(p_i) \end{aligned} \tag{3}$$

ND loss 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 $|\mathbf{y}_{diff}| \in [0, +\infty)$). The sensitivity of the regression penalty increases as $\gamma$ increases, and when $\gamma > 1$, the penalty curve changes from concave to convex.

| Method | Promp. | ↑Accuracy(%) | | | | | | ↑OA(%) | ↓RMSE | | ↓OR | Param. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Bas Ju | Rel Ju | Bas Co | Rel Co | Obj An | Com An | | Bas Co | Rel Co | | (M) |
| SAN | × | 87.59 | 81.79 | 76.26 | 59.23 | 55.00 | 43.25 | 75.66 | 1.136 | 1.318 | 1.160 | 32.30 |
| MAC | × | 82.89 | 79.46 | 72.53 | 55.86 | 46.32 | 40.50 | 71.98 | 1.407 | 1.337 | 1.398 | 38.64 |
| BUTD | ✓ | **90.01** | 82.02 | 77.16 | 60.95 | 56.29 | 42.29 | 76.49 | 0.890 | 1.292 | 0.950 | 34.95 |
| BAN | ✓ | 89.81 | 81.87 | 77.58 | 63.71 | 55.67 | 45.06 | 76.74 | 0.819 | 1.241 | 0.883 | 58.73 |
| MCAN | ✓ | 89.65 | 81.65 | 79.83 | 63.16 | 57.28 | 43.71 | 77.07 | 0.816 | 1.230 | 0.879 | 55.17 |
| D-VQA | ✓ | 89.73 | 82.12 | 77.38 | 63.99 | 55.14 | 43.20 | 76.59 | 0.916 | 1.238 | 0.962 | 37.79 |
| BLIP-2 | × | 88.13 | 81.92 | 70.26 | 58.58 | 42.72 | 28.34 | 71.07 | 1.879 | 1.320 | 1.818 | ≈4B |
| Instruct-BLIP | × | 89.67 | 79.69 | 76.96 | 63.34 | 59.72 | 45.68 | 75.25 | 0.799 | 1.217 | 0.862 | ≈4B |
| RSVQA | × | 82.43 | 79.34 | 70.68 | 55.53 | 42.45 | 35.46 | 70.70 | 1.733 | 1.359 | 1.691 | 30.21 |
| RSIVQA | × | 85.32 | 80.44 | 75.01 | 56.63 | 51.55 | 39.25 | 73.71 | 1.718 | 1.346 | 1.676 | 41.41 |
| **SOBA** (ours) | ✓ | 89.63 | **82.64** | **80.17** | **67.86** | **61.40** | **49.30** | **78.14** | **0.785** | **1.145** | **0.839** | 40.46 |

Table 1: Compared results with other VQA methods on EarthVQA$^{test}$

# 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 (Lobry et al. 2020), RSIVQA (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 FlanT5$_{XL}$. 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 $512 \times 512$ 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. Compared with advanced methods, SOBA achieves the best overall performances with similar or lower complexity.

| Object Guidance | Att. Type | ↑OA(%) | ↓OR |
|---|---|---|---|
| Only Concat | - | 77.61 | 0.856 |
| +SA | S | 77.72 | 0.861 |
| +SCSE | C&S | 77.89 | 0.854 |
| +CBAM | C&S | 77.95 | 0.857 |
| +SE | C | 78.02 | 0.853 |
| +GC | C | 78.03 | 0.847 |
| +**OGA** (ours) | C | **78.14** | **0.839** |

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.

| Cross-Attention | Query | ↑OA(%) | ↓OR |
|---|---|---|---|
| One-way (vanilla) | LLLLLL | 77.11 | 0.977 |
| | VVVVVV | 77.53 | 0.880 |
| Bidirectional | LLL-VVV | 77.57 | 0.867 |
| | VVV-LLL | **78.14** | **0.839** |

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.

| VSA | BCA | Promp. | OGA | ND | ↑OA (%) | ↓OR |
|---|---|---|---|---|---|---|
| ✓ | | | | | 72.55 | 1.509 |
| | ✓ | | | | 73.78 | 1.520 |
| ✓ | ✓ | | | | 74.91 | 1.128 |
| ✓ | ✓ | ✓ | | | 77.30 | 0.866 |
| ✓ | ✓ | ✓ | ✓ | | 77.54 | 0.859 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **78.14** | **0.839** |

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-LLL` 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.

| Img Enc | Lan Enc | ↑mIoU(%) | ↑OA(%) |
|---|---|---|---|
| HR-W40 | LSTM | 57.31 | 77.92 |
| MiT-B3 | LSTM | 56.44 | 77.43 |
| Swin-T | LSTM | 56.89 | 77.22 |
| Swin-S | LSTM | 57.44 | 78.01 |
| ConvX-T | LSTM | 57.17 | 78.24 |
| ConvX-S | LSTM | **57.34** | 78.43 |
| Swin-T | BERT-Base | 56.89 | 77.63 |
| Swin-S | BERT-Base | 57.44 | 78.23 |
| ConvX-S | BERT-Base | **57.34** | **78.65** |

Table 5: Encoder variants analysis

| Query | ↑OA(%) |
|---|---|
| `LV-LV-LV` | 77.51 |
| `VL-VL-VL` | 77.58 |
| `LLL-VVV` | 77.57 |
| `VVV-LLL` | **78.14** |
| Parallel | 77.98 |

Table 6: BCA variants

| Optim. | ↑OA(%) |
|---|---|
| CE | 77.54 |
| DIW | 77.38 |
| Focal | 77.57 |
| OHEM | 77.85 |
| SOM | 77.58 |
| **ND** | **78.14** |

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.

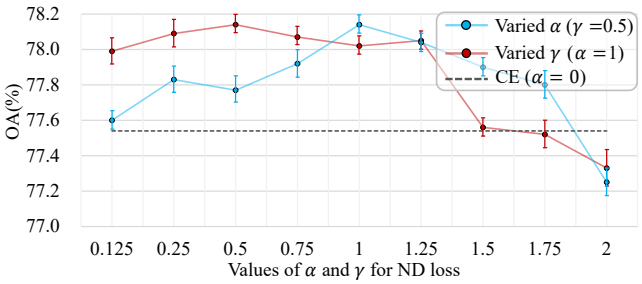
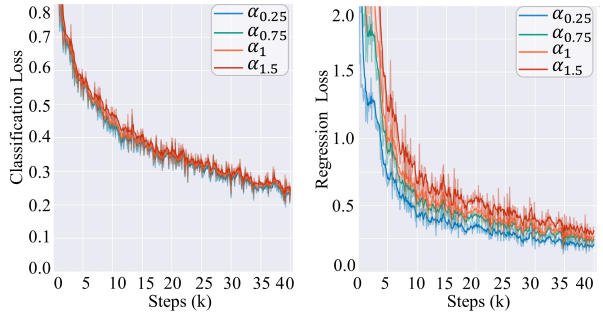



Figure 4: Experimental results with varied $\alpha$ and $\gamma$ for ND loss. The optimal values range from 0.125 to 1.25 with wide ranges of hyperparameters selection. The mean values and standard deviations are reported after five runs.

### 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.




(a) Classification loss ($\gamma_{0.5}$) (b) Regression loss ($\gamma_{0.5}$).

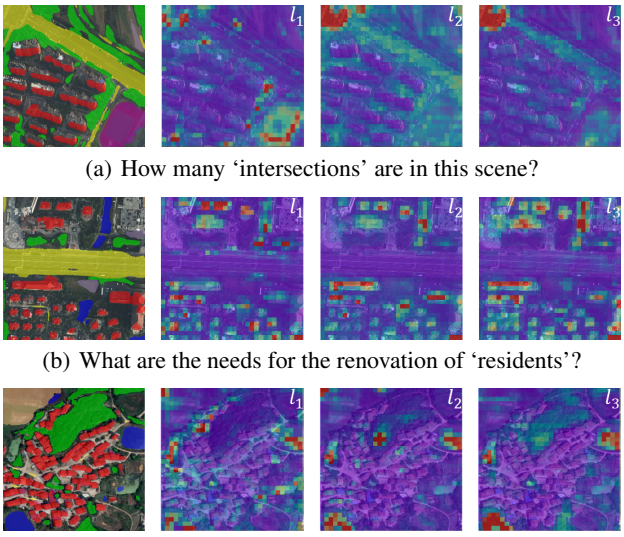
(c) Classification loss ($\alpha_{1.0}$). (d) Regression loss ($\alpha_{1.0}$).

Figure 5: The training losses of classification and regression tasks with different $\alpha$ and $\gamma$. The changes of $\alpha$ and $\gamma$ mainly affect the regression task optimization.

ND loss comprises two components, *i.e.*, the original classification loss and an enhanced regression loss. Figure 5 illustrates the effects of varying $\alpha$ and $\gamma$ on these two types of loss. It is evident that changes have little impact on classification optimization, as the difference penalty is only added to the regression loss. As the values of $\alpha$ and $\gamma$ increase, the regression losses become larger and more unstable. However, as training progresses, the regression losses gradually stabilize and eventually converge. Figure 5 shows that these two parameters control the numerical difference penalty in different ways. This decomposition analysis of training loss can also provide references for tuning $\alpha$ and $\gamma$.

### 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.




(a) How many 'intersections' are in this scene?

(b) What are the needs for the renovation of 'residents'?

(c) What are the 'water' types in this scene?

Figure 6: Visualization of attention maps in BCA with language features as queries. From left to right are the $l_1, l_2$ and $l_3$. Three examples are queried by different keywords: 'intersections', 'residents', and 'water'.

## 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

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Bai, X.; Shi, P.; and Liu, Y. 2014. Society: Realizing China's urban dream. *Nature*, 509(7499): 158–160.

Cao, Y.; Xu, J.; Lin, S.; Wei, F.; and Hu, H. 2019. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Cascante-Bonilla, P.; Wu, H.; Wang, L.; Feris, R. S.; and Ordonez, V. 2022. Simvqa: Exploring simulated environments for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5056–5066.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Hudson, D. A.; and Manning, C. D. 2018. Compositional Attention Networks for Machine Reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6399–6408.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1): 32–73.

Li, F.-F.; and Krishna, R. 2022. Searching for computer vision north stars. *Daedalus*, 151(2): 85–99.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Liu, Y.; Zhong, Y.; Ma, A.; Zhao, J.; and Zhang, L. 2023. Cross-resolution national-scale land-cover mapping based on noisy label learning: A case study of China. *International Journal of Applied Earth Observation and Geoinformation*, 118: 103265.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Lobry, S.; Marcos, D.; Murray, J.; and Tuia, D. 2020. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8555–8566.

Ma, A.; Wang, J.; Zhong, Y.; and Zheng, Z. 2022. FactSeg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Marino, K.; Chen, X.; Parikh, D.; Gupta, A.; and Rohrbach, M. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14111–14121.

Rahnemoonfar, M.; Chowdhury, T.; Sarkar, A.; Varshney, D.; Yari, M.; and Murphy, R. R. 2021. FloodNet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9: 89644–89654.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. 2017. ChexNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Roy, A. G.; Navab, N.; and Wachinger, C. 2018. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE Transactions on Medical Imaging*, 38(2): 540–549.

Savage, N. 2019. How AI and neuroscience drive each other forwards. *Nature*, 571(7766): S15–S15.

Shi, S.; Zhong, Y.; Liu, Y.; Wang, J.; Wan, Y.; Zhao, J.; Lv, P.; Zhang, L.; and Li, D. 2023. Multi-temporal urban semantic understanding based on GF-2 remote sensing imagery: from tri-temporal datasets to multi-task mapping. *International Journal of Digital Earth*, 16(1): 3321–3347.

Shrivastava, A.; Gupta, A.; and Girshick, R. 2016. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 761–769.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; and Zhong, Y. 2021. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Wen, Z.; Xu, G.; Tan, M.; Wu, Q.; and Wu, Q. 2021. Debiased visual question answering from feature and sample perspectives. *Advances in Neural Information Processing Systems*, 34: 3784–3796.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xia, G.-S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; and Lu, X. 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7): 3965–3981.

Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Wang, Y.; and Zhang, L. 2023. From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution. *Information Fusion*, 96: 297–311.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21–29.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6281–6290.

Yuan, Z.; Mou, L.; Xiong, Z.; and Zhu, X. X. 2022. Change detection meets visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Zhang, Y.; Yuan, Y.; Feng, Y.; and Lu, X. 2019. Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8): 5535–5548.

Zhao, H.; Zhong, Y.; Wang, X.; Hu, X.; Luo, C.; Boitt, M.; Piiroinen, R.; Zhang, L.; Heiskanen, J.; and Pellikka, P. 2022. Mapping the distribution of invasive tree species using deep one-class classification in the tropical montane landscape of Kenya. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187: 328–344.

Zheng, X.; Wang, B.; Du, X.; and Lu, X. 2021. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Zvonkov, I.; Tseng, G.; Nakalembe, C.; and Kerner, H. 2023. OpenMapFlow: A Library for Rapid Map Creation with Machine Learning and Remote Sensing Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14655–14663.