# CHANGE-AWARE VISUAL QUESTION ANSWERING

*Zhenghang Yuan*[1], *Lichao Mou*[1,2], *Xiao Xiang Zhu*[1,2]

[1] Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

## ABSTRACT

Change detection has been a hot research topic in the field of remote sensing, and it can provide information on observing changes of Earth's surface. However, segmentation-based change results are not very friendly to end users. Thus, in order to improve user experience and offer them high-level semantic information on change detection, we introduce a new task: change-aware visual question answering (VQA) on multi-temporal aerial images. Specifically, given a pair of multi-temporal aerial images and questions, this task aims to automatically provide natural language answers. By doing so, end users have better access to easy-to-understand change information through natural language. Besides, we also create a dataset made of multi-temporal image-question-answer triplets and a baseline method for this task. Experimental results offer valuable insights for the further research on this task.

***Index Terms*—** visual question answering (VQA), change detection, aerial images, natural language, deep learning

## 1. INTRODUCTION

Change detection aims to identify differences of the same area at different times, and it is significant to understand and explore the relationships and interactions between humans and nature [1]. A lot of change detection algorithms have been developed and find great use in real-world applications [2, 3, 4, 5]. However, segmentation-based change results, i.e., binary change maps [6] and semantic change maps [7], are not very intuitive and friendly to end users. For example, it is difficult to observe with naked eyes what is the smallest change in the given change maps. If we can provide an answer to the above question in natural language instead of maps, it will make users understand better. In this context, tasks of combining remote sensing images and natural language, such as image captioning [8, 9] and visual question answering (VQA) [10, 11], attract a lot of attention in the remote sensing community nowadays.

With the advent of deep learning, various multi-modal learning networks [12, 13] are proposed for both the computer vision and remote sensing communities [14, 15]. VQA for natural images has been developed for many years [16, 17], and VQA for remote sensing images has also been a hot research topic in recent years[18, 19]. However, change detection-based VQA for multi-temporal aerial images has been neglected. Considering the above issues, in order to enhance user experience and provide them with easy-to-understand information on change detection, we introduce a new task: change-aware VQA on multi-temporal aerial images. As shown in Fig. 1, given a pair of multi-temporal aerial images and corresponding questions about images, this task aims to automatically provide accessible natural language answers. By doing so, it can greatly improve human-computer interaction and allow end users to have better access to change information contained in aerial images.

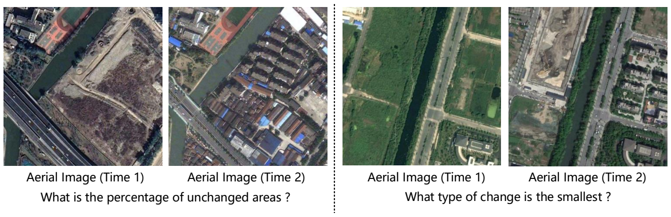

**Fig. 1**. Examples of multi-temporal aerial images and questions.

Datasets are critical for the comparison and evaluation of different approaches in remote sensing community [20]. In this paper, we create a dataset by automatically generating question-answer pairs for the change-aware VQA task. To be specific, the dataset consists of 2,968 pairs of multi-temporal aerial images and more than 122,000 question-answer pairs. In addition, we also propose a baseline method for this task. The main architecture of it is shown in Fig. 2, which includes three parts: feature encoding, feature fusion, and answer prediction.

The rest of the paper is organized as follows. We introduce the dataset in Section 2 and the baseline method in Section 3. Experimental results and some discussion are described in Section 4. Finally, we conclude this paper in Section 5.

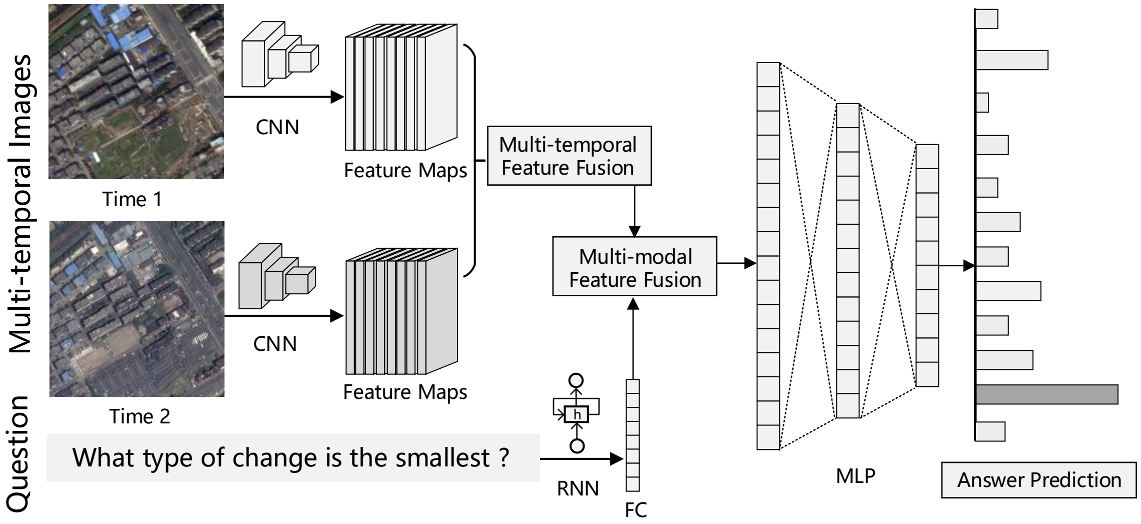

**Fig. 2**. The main architecture of the proposed baseline model.

## 2. DATASET

In this work, the dataset is automatically generated based on the semantic change detection dataset SECOND [21]. There are 2,968 pairs of publicly available multi-temporal aerial images in the SECOND dataset, and each pair has two semantic change maps at the pixel level. Particularly, 6 land-cover classes about changes are labeled in the semantic change maps, including buildings, playgrounds, low vegetation, trees, non-vegetated ground (NVG) surface, and water. In general, five question types are designed based on the semantic change information of land-cover classes.

1. Change or not. This question type focuses on whether a land-cover has changed. The answer is *yes* if the corresponding land-cover has changed, and vice versa.
2. Increase/decrease or not. The point of this question type is whether the area of a certain land-cover has increased or decreased. Based on this, the answers is *yes* or *no*.
3. Change to what. This question type pays attention to what the land-cover at time $t_1$ mainly changed to at time $t_2$. Answers are corresponding land-cover classes.
4. Largest/smallest change. What is the largest or smallest change in multi-temporal images is also a common question type. Answers are also land-cover classes.
5. Change ratio. This type focuses on the ratio of changed regions or unchanged regions. Answers are quantized numbers: 0%, 0%-10%, 10%-20%,..., 90%-100%. In this case, A%-B% means (A, B].

There are question templates for these question types listed above. To add randomness and variety, templates are randomly selected to generate question-answer pairs. Overall, there are more than 122,000 question-answer pairs in this dataset.

## 3. METHODOLOGY

We formulate the change-aware VQA task as a classification task. The overall architecture of the proposed baseline method for it is shown in Fig. 2. The inputs to the model are two multi-temporal aerial images and the corresponding question, and the output is an answer. To be specific, the architecture consists of three parts: feature encoding, feature fusion, and answer prediction.

### 3.1. Feature Encoding

Feature encoding part is used to extract visual and language features from the input images and question, respectively. In this work, we compare three backbones including ResNet-18, ResNet-101, and ResNet-152 as encoders to learn visual features. Let $\boldsymbol{x}_{t_1} \in \mathbb{R}^{3\times H\times W}$ and $\boldsymbol{x}_{t_2} \in \mathbb{R}^{3\times H\times W}$ denote the image at time $t_1$ and time $t_2$, respectively. $H, W$ are the height and width of them. $\boldsymbol{F}_1 = f_1(\boldsymbol{x}_{t_1})$ and $\boldsymbol{F}_2 = f_2(\boldsymbol{x}_{t_2})$ are visual features learned from two input images. As we use Siamese networks, $f_1$ and $f_2$ share the same network architecture and parameters. As for the input question, recurrent neural network (RNN) is used to encode it as the language feature $\boldsymbol{V}_q \in \mathbb{R}^{N\times L}$. $N$ is the batch size and $L$ is the dimension of feature vector.

### 3.2. Feature Fusion

In the feature fusion part, we need to make multi-temporal fusion and multi-modal fusion. Since how to fuse features is

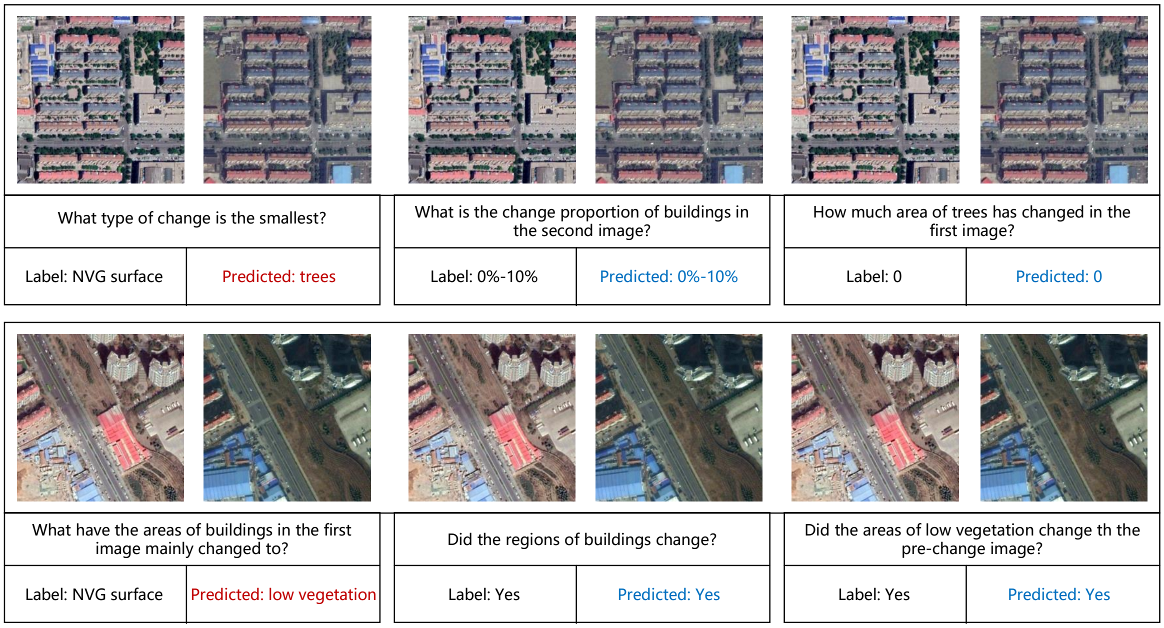
What type of change is the smallest?
Label: NVG surface
Predicted: trees
What is the change proportion of buildings in the second image?
Label: 0%-10%
Predicted: 0%-10%
How much area of trees has changed in the first image?
Label: 0
Predicted: 0
What have the areas of buildings in the first image mainly changed to?
Label: NVG surface
Predicted: low vegetation
Did the regions of buildings change?
Label: Yes
Predicted: Yes
Did the areas of low vegetation change th the pre-change image?
Label: Yes
Predicted: Yes

**Fig. 3**. Visualization examples of experimental results.

not the main study of this work, we simply fuse them by concatenation in both cases. The multi-temporal feature fusion can be formulated as:

$$\boldsymbol{F}_v = \boldsymbol{F}_1 \frown \boldsymbol{F}_2, \tag{1}$$

where $\boldsymbol{F}_v$ is the fused multi-temporal feature, and $\frown$ denotes the concatenation operation. In addition, we resize the feature $\boldsymbol{F}_v$ into $\boldsymbol{F}_{vt} \in \mathbb{R}^{N \times L}$. Then, $\boldsymbol{F}_{vt}$ and $\boldsymbol{V}_q$ have the same size. The multi-modal feature fusion can be described as:

$$\boldsymbol{F}_m = \boldsymbol{V}_q \frown \boldsymbol{F}_{vt}, \tag{2}$$

where $\boldsymbol{F}_m$ denotes the fused multi-modal feature.

### 3.3. Answer Prediction

The third part is the answer prediction. As we formulate it as a classification task, the final answer can be obtained by selecting the answer class with the largest probability from a pre-defined answer pool. In this work, we use fully connected layers as the classifier to get the answer.

## 4. EXPERIMENTS

### 4.1. Implementation Details

Adam optimizer is used in this work and the initial learning rate is 1e-4. The batch size is set to 70, and the input image size is scaled to 256 × 256 for all ResNet-based models. 50 epochs are used to train models in all experiments. In addition, average accuracy and overall accuracy are adopted to measure the performance of models.

### 4.2. Results and Discussion

Experiments are conducted on the proposed dataset. We compare three backbones (ResNet-18, ResNet-101, and ResNet-152) and report the accuracy of the baseline method on the test set in Table 1. Note that change ratio denotes change ratio for all land-covers, while class change ratio means change ratio for each land-cover. Some visualization examples are shown in Fig. 3. From numerical results in Table 1, we can see that different question types vary greatly in accuracy. For example, the baseline method achieves less than 30% accuracy for the question type of smallest change, while obtains approximately 83% accuracy for the question type of change or not. This is mainly because the difficulty of different questions varies significantly.

From experimental results, it can also be seen that this task is a challenging task. In order to correctly answer different question types, the model needs to learn multi-modal features and analyze semantic change information. To be specific, the model needs not only to locate the area of change according to questions, but also to identify the land-cover category of changed areas in order to answer some complex questions. More research is needed to achieve satisfactory performance on this task.

## 5. CONCLUSION

In this paper, we introduce a new task: change-aware VQA on multi-temporal aerial images. Specifically, given a pair of bi-temporal aerial images and corresponding questions, this task aims to automatically provide natural language an-

**Table 1**. Results on the Test Set of the Proposed Dataset.

| Question Types | ResNet-18 | ResNet-101 | ResNet-152 |
|---|---|---|---|
| change ratio | 0.3455 | 0.3388 | **0.3476** |
| class change ratio | **0.7200** | 0.7134 | 0.7115 |
| change or not | 0.8379 | **0.8387** | 0.8374 |
| change to what | 0.5710 | 0.5737 | **0.5770** |
| increase or not | **0.6913** | 0.6902 | 0.6854 |
| decrease or not | **0.7303** | 0.7243 | 0.7275 |
| smallest change | 0.2627 | **0.2758** | 0.2734 |
| largest change | 0.4603 | 0.4576 | **0.4669** |
| Average Accuracy | 0.5773 | 0.5766 | **0.5783** |
| Overall Accuracy | **0.6771** | 0.6763 | 0.6766 |

swers. To compare and evaluate different methods, we build a dataset for it. Specifically, this dataset consists of 2,968 pairs of multi-temporal aerial images and more than 122,000 question-answer pairs. In addition, we propose a baseline method for change-aware VQA. Experiments are conducted and results offer valuable insights for the further research on this task.

## 6. ACKNOWLEDGMENT

This work is supported by the China Scholarship Council.

## 7. REFERENCES

[1] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *International journal of remote sensing*, vol. 25, no. 12, pp. 2365–2401, 2004.

[2] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: a survey," *Earth Science Informatics*, vol. 12, no. 2, pp. 143–160, 2019.

[3] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, 2018.

[4] S. Ye, J. Rogan, Z. Zhu, and J. R. Eastman, "A near-real-time approach for monitoring forest disturbance using landsat time series: Stochastic continuous change detection," *Remote Sensing of Environment*, vol. 252, p. 112167, 2021.

[5] X. Li, Z. Yuan, and Q. Wang, "Unsupervised deep noise modeling for hyperspectral image change detection," *Remote Sensing*, vol. 11, no. 3, p. 258, 2019.

[6] Z. Yuan, Q. Wang, and X. Li, "Robust PCANet for hyperspectral image change detection," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4931–4934, IEEE, 2018.

[7] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-Temporal semantic reasoning for the semantic change detection in HR remote sensing images," *arXiv preprint arXiv:2108.06103*, 2021.

[8] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization-driven deep remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[9] G. Hoxha and F. Melgani, "A novel svm-based decoder for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[10] Z. Yuan, L. Mou, and X. X. Zhu, "Self-paced curriculum learning for visual question answering on remote sensing data," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 2999–3002, IEEE, 2021.

[11] Z. Yuan, L. Mou, Q. Wang, and X. X. Zhu, "From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[12] Z. Xiong, Y. Yuan, and Q. Wang, "ASK: Adaptively selecting key local features for RGB-D scene recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2722–2733, 2021.

[13] Z. Xiong, Y. Yuan, N. Guo, and Q. Wang, "Variational context-deformable convnets for indoor scene parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3992–4002, 2020.

[14] S. Lobry, B. Demir, and D. Tuia, "RSVQA meets Bigearthnet: A new, large-scale, visual question answering dataset for remote sensing," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 1218–1221, IEEE, 2021.

[15] Z. Xiong, S. Chen, Y. Shi, and X. X. Zhu, "Disentangled latent transformer for interpretable monocular height estimation," *arXiv preprint arXiv:2201.06357*, 2022.

[16] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[17] W. Guo, Y. Zhang, J. Yang, and X. Yuan, "Re-attention for visual question answering," *IEEE Transactions on Image Processing*, vol. 30, pp. 6730–6743, 2021.

[18] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.

[19] S. Lobry, D. Marcos, B. Kellenberger, and D. Tuia, "Better generic objects counting when asking questions to images: A multitask approach for remote sensing visual question answering," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 1021–1027, 2020.

[20] Z. Xiong, W. Huang, J. Hu, Y. Shi, Q. Wang, and X. X. Zhu, "THE benchmark: Transferable representation learning for monocular height estimation," *arXiv preprint arXiv:2112.14985*, 2021.

[21] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang, "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.