Change Detection Meets Visual Question Answering

Zhenghang Yuan, *Student Member, IEEE*, Lichao Mou, Zhitong Xiong, *Member, IEEE*,
and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*—The Earth's surface is continually changing, and identifying changes plays an important role in urban planning and sustainability. Although change detection techniques have been successfully developed for many years, these techniques are still limited to experts and facilitators in related fields. In order to provide every user with flexible access to change information and help them better understand land-cover changes, we introduce a novel task: change detection-based visual question answering (CDVQA) on multitemporal aerial images. In particular, multitemporal images can be queried to obtain high-level change-based information according to content changes between two input images. We first build a CDVQA dataset, including multitemporal image–question–answer triplets using an automatic question–answer generation method. Then, a baseline CDVQA framework is devised in this work, and it contains four parts: multitemporal feature encoding, multitemporal fusion, multimodal fusion, and answer prediction. In addition, we also introduce a change enhancing module to multitemporal feature encoding, aiming at incorporating more change-related information. Finally, the effects of different backbones and multitemporal fusion strategies are studied on the performance of CDVQA task. The experimental results provide useful insights for developing better CDVQA models, which are important for future research on this task. The dataset will be available at https://github.com/YZHJessica/CDVQA.**

***Index Terms*—Change detection, deep learning, multitemporal aerial images, visual question answering (VQA).**

I. Introduction

THE Earth's surface is continually changing by man-made and natural influences. These changes are closely involved in human and social development and also guide urban planning and sustainability [1]. Change detection, aiming at detecting differences of the same region at different times, has become a research priority in recent decades [2], [3], [4]. Timely and effective change information can be used for many practical applications, such as environmental management [5], [6], [7], natural disasters monitoring [8], [9], urban land use [10], [11], and agriculture production [12].

Nowadays, change detection technology has been developed significantly, and there are various algorithms with great performance improvement for remote sensing data [13], [14], [15]. Change detection methods can be divided into two main categories, depending on whether or not the types of changes are detected. One is binary change detection that only detects changed regions but ignores the type of changes, e.g., the object-oriented key point vector distance for detecting binary land-cover changes [16] and the end-to-end 2-D convolutional neural network (CNN) for hyperspectral image change detection [17]. Change maps obtained by such methods are visualized by binary values to depict change information at the pixel level. The other is semantic change detection, for instance, using an asymmetric Siamese network for identifying changes via feature pairs [18] and reasoning bitemporal semantic correlations [19]. These methods not only detect changed regions but also identify change types.

Although change detection has great application value, the specialized nature of this task makes change information limited to researchers. It is still difficult for end users to access and understand much of important change information. For instance, ordinary users are interested in a certain change type in a certain region, but it is inconvenient and ineffective for them to find it on change maps in practical applications. Considering this problem, efficient and effective change information interaction with end users becomes important. In this context, natural language processing (NLP) enables computers to understand the text in almost the same way as humans. It is user-friendly and can greatly improve the interactivity between image analysis systems and end users. Therefore, in order to alleviate the interaction issue, the integration of computer vision and NLP [20] has gradually become a hot research topic in the machine learning community. In particular, tasks, such as visual description generation [21], visual storytelling [22], visual question answering (VQA) [23], [24], and visual dialog [25], have been fully and successfully conducted in computer vision. Similarly, tasks of integrating remote sensing imagery and NLP, such as image captioning and VQA, have also become an active research topic in the field of remote sensing [26], [27]. Captioning for remote sensing images was first proposed in [28], and Lu *et al.* [29] further explored captioning methods using both handcrafted and convolutional

Manuscript received 25 November 2021; revised 6 April 2022 and 21 June 2022; accepted 1 August 2022. Date of publication 23 September 2022; date of current version 29 September 2022. This work was supported in part by the China Scholarship Council; in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme (Acronym: So2Sat) under Grant ERC-2016-StG-714087; in part by the Helmholtz Association through the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100; in part by the German Federal Ministry of Education and Research (BMBF) in the Framework of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001; and in part by the German Federal Ministry for Economic Affairs and Climate Action in the Framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Corresponding authors: Lichao Mou; Xiao Xiang Zhu.)*

The authors are with the Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: zhenghang.yuan@tum.de; lichao.mou@tum.de; zhitong.xiong@tum.de; xiaoxiang.zhu@tum.de).

Digital Object Identifier 10.1109/TGRS.2022.3203314

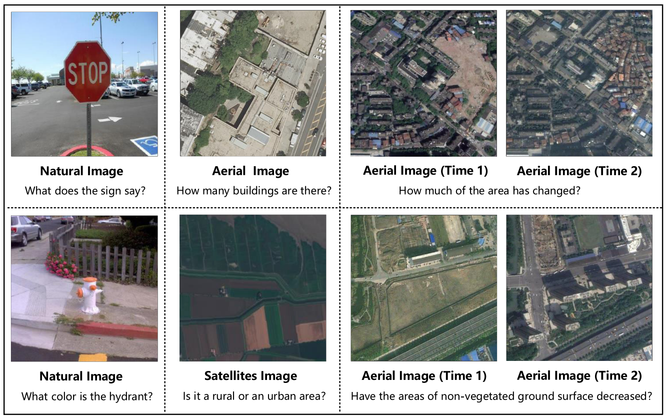

Natural Image
What does the sign say?
Aerial Image
How many buildings are there?
Aerial Image (Time 1)
Aerial Image (Time 2)
How much of the area has changed?
Natural Image
What color is the hydrant?
Satellites Image
Is it a rural or an urban area?
Aerial Image (Time 1)
Aerial Image (Time 2)
Have the areas of non-vegetated ground surface decreased?

Fig. 1. Examples of questions for natural imagery, aerial imagery, and multitemporal aerial images in VQA tasks.

features and proposed a new dataset. Recently, a multilayer aggregated Transformer was utilized to extract information for caption generation [30]. Regarding VQA for remote sensing data (RSVQA), Lobry *et al.* [31] first introduced this task, built two datasets, and used a hybrid CNN–recurrent neural network (RNN) model to extract features, and Yuan *et al.* [32] proposed a self-paced curriculum learning-based model trained from easy to hard questions gradually.

Compared to natural images, aerial images are more specialized due to the top-view perspective and complicated background. As shown in Fig. 1, answers to questions about natural images [23] are more obvious than answers to questions about aerial images [31] in many cases in VQA tasks. Besides, Fig. 1 shows that answers to questions about the comparison of multitemporal aerial images require careful observation and even calculation, which is unfriendly to ordinary users. Though VQA for natural images has been studied for many years and VQA for remote sensing data has also gradually become a research focus, VQA for change detection based on multitemporal images is underexplored. Considering the significance of change detection task and its values in practical applications, it is vital to investigate how to improve the friendliness and accessibility of change detection systems to end users. Hence, there is also a greater need to develop end user accessible VQA algorithms for multitemporal remotely sensed data.

In this article, we introduce the task of change detection-based visual question answering (CDVQA) on multitemporal aerial images. Specifically, given two aerial images captured at different times and a natural language question about them, the CDVQA task aims to provide an answer in natural language by comparing the content of two images. To this end, we create a CDVQA dataset by an automatic generation method, which contains 2968 pairs of multitemporal images and more than 122 000 question–answer pairs. The questions are carefully designed to cover various types of changes. Moreover, we propose a baseline method for the CDVQA task, as shown in Fig. 2. To sum up, the main contributions of this work are summarized as follows.

1) We design an automatic question–answer generation method and create a new CDVQA dataset. Specifically, the proposed dataset contains 2968 pairs of aerial images and more than 122 000 corresponding question–answer pairs.
2) A baseline framework for CDVQA task is proposed, and it includes four parts: multitemporal feature encoding, multitemporal fusion, multimodal fusion, and answer prediction. In addition, a change enhancing module is proposed to incorporate more change-related information into visual features.
3) Extensive experiments have been conducted to study the effects of different network parts on the CDVQA performance. In particular, different backbones and multitemporal fusion strategies are investigated. The results provide useful insights on the CDVQA task.

The rest of this article is organized as follows. The detailed information for the construction of CDVQA dataset is introduced in Section II. Section III presents the methodology. Experimental results and discussion are shown in Section IV. Finally, this article is concluded in Section V.

II. Dataset

Different from the traditional VQA task, CDVQA involves multitemporal aerial images and requires time series analysis. Taking this into account, we choose the existing semantic change detection dataset SECOND [18] as the basic data to automatically generate a CDVQA dataset. The SECOND dataset collects bitemporal high-resolution optical (RGB) images by several different aerial platforms and sensors, with spatial resolution varying from 0.5 to 3 m [19]. Geographical positions include several cities in China, such as Shanghai, Hangzhou, and Chengdu. It has 4662 pairs of aerial images with the size of 512×512 pixels, and 2968 pairs are publicly available. Each pair consists of a preevent aerial image and a postevent image of the same location at different times. Besides, each pair has two labeled semantic change maps at the pixel level, before and after the change. In each semantic change map, nonchange region and six land-cover classes related to changes, including nonvegetated ground (NVG) surface, buildings, playgrounds, water, low vegetation, and trees, are annotated. The authors of the SECOND dataset declare in their paper that semantic change maps in this dataset are labeled by a team of experts in Earth vision applications and high accuracy is guaranteed. Therefore, the generated question–answer pairs in this work are highly relevant to the content of image pairs. Overall, this dataset has critical semantic change information of main land-cover classes at the pixel level, which provides sufficient information for generating question–answer pairs for the CDVQA task. In this case, we use the 2968 openly available pairs as our basic data for further dataset construction.

A. Multitemporal Image–Question–Answer Triplets Construction

Formally, in each pair of multitemporal aerial images, let $\boldsymbol{x}_{t_1} \in \mathbb{R}^{3\times H\times W}$ be the image at time T_1 and $\boldsymbol{x}_{t_2} \in \mathbb{R}^{3\times H\times W}$ be the image at time T_2. $\boldsymbol{s}_{t_1} \in \mathbb{R}^{H\times W}$ and $\boldsymbol{s}_{t_2} \in \mathbb{R}^{H\times W}$ denote semantic change maps of $\boldsymbol{x}_{t_1}$ and $\boldsymbol{x}_{t_2}$, respectively, and

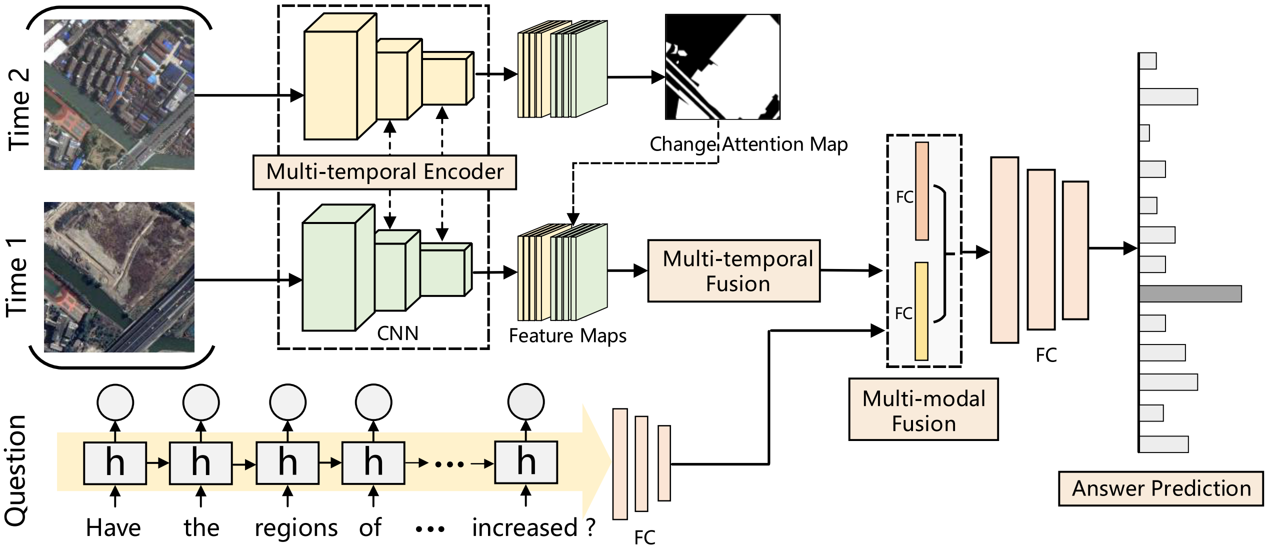


Fig. 2. Main architecture of the proposed CDVQA framework. It contains four main parts: multitemporal feature encoding, multitemporal fusion, multimodal fusion, and answer prediction.

each pixel in $\boldsymbol{s}_{t_1}$ and $\boldsymbol{s}_{t_2}$ indicates one semantic class, ranging from 0 to 6. Semantic change maps show changed regions and provide their change types at the pixel level. Background pixels mean nonchange regions, which are the same in both $\boldsymbol{s}_{t_1}$ and $\boldsymbol{s}_{t_2}$ for an image pair. Foreground pixels indicate changed regions of different land-cover types. Specifically, the value of the pixel in $\boldsymbol{s}_{t_1}$ indicates the semantic class at T_1 and the value of the pixel in $\boldsymbol{s}_{t_2}$ indicates the semantic class at T_2. The main advantage of introducing semantic change maps is that we can access more details about changes, i.e., we are able to know not only where changes happen but also what types they are. In this work, our motivation is to use natural language as queries to obtain these two types of information. Given semantic change information of $\boldsymbol{s}_{t_1}$ and $\boldsymbol{s}_{t_2}$, the following five types of questions are designed in the proposed dataset: change or not, increase/decrease or not, change to what, largest/smallest change, and change ratio. In our case, the smallest/largest change refers to the land-cover class that has the least/most pixels changing. These questions are of great interest to end users in real-world applications. In what follows, a detailed description of the automatic generation of multitemporal image–question–answer triplets for different question types is given.

1) *Change or Not:*

a) *Change or Not for an Image Pair:* The most fundamental yet important information in change detection is about whether a certain land cover has changed. Note that a change occurs regardless of whether the area of a land cover increases or decreases. For each pair of aerial images, the set of changed land-cover classes $\mathcal{L}_{t_1}$ and $\mathcal{L}_{t_2}$ are extracted from $\boldsymbol{s}_{t_1}$ and $\boldsymbol{s}_{t_2}$, respectively. Let l_i be a land-cover class, $l_i \in \mathcal{L}_{t_1}$ or $l_i \in \mathcal{L}_{t_2}$, indicating that the corresponding land-cover type has changed. In this case, the answer should be *yes*. On the contrary, if $l_i \notin \mathcal{L}_{t_1}$ and $l_i \notin \mathcal{L}_{t_2}$, it indicates that the corresponding land cover does not change. Then, the answer should be *no*. All land-cover types are traversed to generate multiple question–answer pairs.

b) *Change or Not for a Single Image:* For change detection tasks, sometimes one want to focus not only on whether a certain land-cover class has changed but also on whether changes have occurred in the preevent image or postevent image. Therefore, we extract semantic change information solely from the first or second image to generate relevant questions and answers. Please note that in this work, the first image in the image pair refers to the preevent/prechange, and the second image means the postevent/postchange. In particular, for the land-cover class l_i, if $l_i \in \mathcal{L}_{t_1}$, it indicates that the corresponding land cover has changed on the preevent image. The answer under this situation should be *yes*. Similarly, if $l_i \in \mathcal{L}_{t_2}$, it means that the area of l_i has changed on the postevent image. The answer will also be *yes*. In other cases, i.e., $l_i \notin \mathcal{L}_{t_1}$ and $l_i \notin \mathcal{L}_{t_2}$, the corresponding answer to the question about whether it has changed on a single image should be *no*.

2) *Increase/Decrease or Not:* Change detection in real-world applications often requires more specific change information, for instance, whether the area of a land cover has increased or decreased. In this context, we denote the area of l_i in $\boldsymbol{s}_{t_1}$ as $\boldsymbol{A}^i_{t_1}$ and the area in $\boldsymbol{s}_{t_2}$ as $\boldsymbol{A}^i_{t_2}$. For increasing-related question–answer pairs, if $\boldsymbol{A}^i_{t_2} - \boldsymbol{A}^i_{t_1} > 0$, the area of l_i increases. Then, the answer to this question should be *yes*. For decreasing-related pairs, the generation process is similar. If $\boldsymbol{A}^i_{t_2} - \boldsymbol{A}^i_{t_1} < 0$, the area of l_i decreases. Note that the area of l_i is defined as all pixels with label l_i in the whole imagery.

3) *Change to What:* This type of question involves more detailed information about changes, i.e., what the land cover at time T_1 mainly becomes at time T_2. Such questions require analyzing the same region in multitemporal images to obtain the change of land-cover types in this region. Although one class may change to more than one class over time, it is more meaningful to focus on the major change. In particular, for a semantic class, we first find its pixel indices in $\boldsymbol{s}_{t_1}$. Then, the indices are used to select the corresponding pixels in $\boldsymbol{s}_{t_2}$. Finally, we count the number of the selected pixels for each land-cover type and choose the type with the largest number as the major change. In this case, the answer to the question what the regions of l_i at time T_1 mainly change to should be the major change type.

4) *Largest/Smallest Change:*

   a) *Largest/Smallest Change for an Image Pair:* Such questions focus on the largest or smallest changes in multitemporal images. For each land-cover type, all changes in the two images should be considered. Therefore, the changed area for the land-cover class l_i is $\boldsymbol{A}_{t_1}^{i} + \boldsymbol{A}_{t_2}^{i}$. By traversing all change types, the maximum and minimum changed regions can be obtained, and the corresponding land-cover classes are answers to this type of question. In this dataset, the smallest change is which has the smallest changed area, and the unchanged type is not considered.

   b) *Largest/Smallest Change for a Single Image:* To extract more detailed information about changes, we also analyze the maximum and minimum changed regions for the preevent and postevent images, respectively. The maximum and minimum changed regions at time T_1 can be easily obtained by $\arg\max_{l_i}(\boldsymbol{A}_{t_1}^{i})$ and $\arg\min_{li}(\boldsymbol{A}_{t_1}^{i})$, and the selected land cover l_i is the corresponding answer. For time T_2, the generation process is the same. This type of question requires a model to not only identify land-cover changes in bitemporal images but also understand which image (T_1 or T_2) is queried by users. In this context, the question "What is the smallest change in the first image?" is actually asking about the land cover of the smallest changed region in the image captured at an earlier date.

5) *Change Ratio:*

   a) *Change Ratio for All Land Covers:* The percentages of changed regions are also very important information in practical applications. The change ratio can be calculated via dividing the changed area by the total area of the whole map and the same for nonchange ratio. Since proportions are continuous numbers, they cannot be compatible with the classification task. Thus, we discretize ratios into bins. To be more specific, numerical answers are quantized into 11 categories: 0%, 0%–10%, 10%–20%, 20%–30%, 30%–40%, 40%–50%, 50%–60%, 60%–70%, 70%–80%, 80%–90%, and 90%–100%. Notice that in this context, A%-B% means $(A, B]$. In this way, we calculate the change percentage for each image pair and gain answers to the change ratio-related questions.

   b) *Change Ratio for Each Land Cover:* In addition to the ratio of all changed regions, we also want to analyze the change ratio for each land-cover class on the preevent or postevent image. Similarly, numerical answers are also quantized as above. For each land-cover class l_i, we first calculate its changed regions $\boldsymbol{A}_{t_1}^{i}$ and $\boldsymbol{A}_{t_2}^{i}$ at T_1 and T_2. Then, the change ratio for l_i on the preevent image is calculated via dividing $\boldsymbol{A}_{t_1}^{i}$ by the total area of the whole image. In the same way, change ratios for different land covers on the postevent image can be obtained.

In practice, we have defined multiple synonymous templates for each type of questions. During the question–answer generation process, for each image pair, question–answer pairs are generated separately for each question type. As more than one template is designed for each question type, we randomly select one of them to generate a sample. To balance the number of samples in each question type, we set different probabilities for generating samples of different question types. Specifically, we set a low probability value for the "yes/no" type and a high probability value for other question types. For each image pair, we generate 16 samples in average.

B. Question and Answer Distributions

As 2968 pairs of images are publicly available, we use these images as the basic data to generate the CDVQA dataset. The whole dataset is split into the training set, validation set, and test set. To better evaluate the robustness and reliability of CDVQA models, we generate two test sets with different distributions of answers. The class distributions of answers in the generated CDVQA dataset are shown in Fig. 3. From this figure, we can see that the training set, validation set, and test set 1 share the same class distribution. The answer distributions of test sets 1 and 2 are different.

As we can see from Fig. 3, answer types in all subsets obey the long-tail distribution. Concretely, answer class *no* dominates answer distributions in all subsets. For example, in the training set, samples with answer *no* occupy 30.9% of all instances. In test set 1, answers *no* account for 31.15% of total answers. In contrast, answers 50%–60% only occupy 0.22% of all answers. The reason for the class imbalance is that there are more questions asking for *yes* or *no*. The answers to questions such as change or not and increase/decrease or not are *yes* or *no*.

The question type distributions of all four subsets are presented in Fig. 4. For simplicity, change ratio for each land-cover is denoted as class change ratio. We can see that distributions of question types are also long-tailed. In addition, question type change or not has the highest frequencies in all subsets. This is the reason why the two most frequent answer types are *yes* and *no*. Similar to answer distributions, the

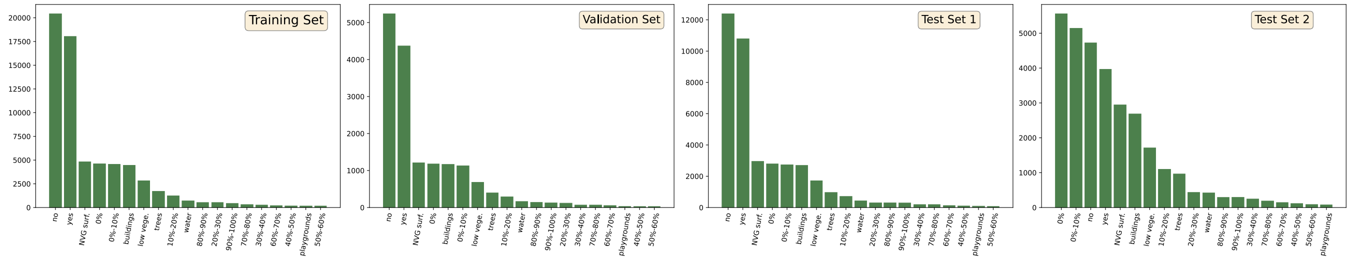


Fig. 3. Visualization of answer distributions of different subsets. From left to right: training set, validation set, test set 1, and test set 2.

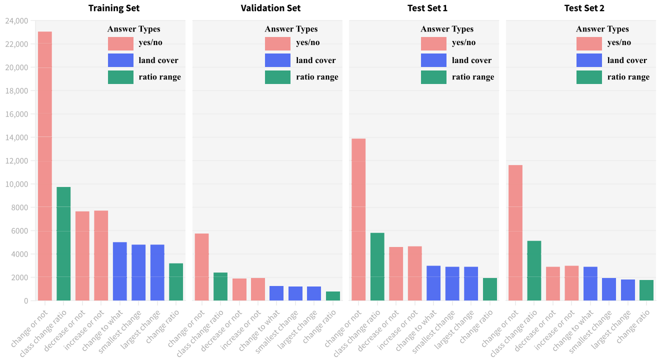


Fig. 4. Visualization of question distributions of different subsets. (From Left to Right) Training set, validation set, test set 1, and test set 2.

distributions of question types for the training set, validation set, and test set 1 are the same, while they are different from the distribution of test set 2. Specifically, the proportions of questions about "change to what," "change ratio," and "class change ratio" increase in test set 2 compared to test set 1. Since questions of these types are more difficult, test set 2 is more difficult than test set 1. Visualization examples of the generated CDVQA dataset are shown in Fig. 5.

III. Methodology

In this work, the CDVQA task is deemed as a classification task. Note that semantic change maps are only used to generate question–answer pairs in the dataset preparation phase, and in CDVQA, only image pairs, questions, and the corresponding answers are used for training and evaluating a model. As shown in Fig. 2, our CDVQA model takes as input two aerial images and a question. The output of the model is an answer predicted by the network. In particular, the whole network architecture consists of four parts. The first component is a multitemporal visual feature learning module, which is used to encode the input images into deep features. The second part, named multitemporal fusion, is responsible for fusing the features of the two images. The third one is a multimodal fusion module that aims at fusing the image and question features. The fourth is an answer prediction part, which takes the fused multimodal feature as input to predict the answer. In addition, for the CDVQA task, we design a change enhancing module to encourage the model to focus on changed pixels of the input images. The proposed modules in our CDVQA framework will be described in detail in the following.

A. Multitemporal Encoder

Different from tasks such as image classification, object detection, and semantic segmentation, change analysis involves two input images of the same location but at different times. Similarly, a CDVQA system takes as input multitemporal inputs. In order to identify changes between two images, temporal differences should be extracted and analyzed.

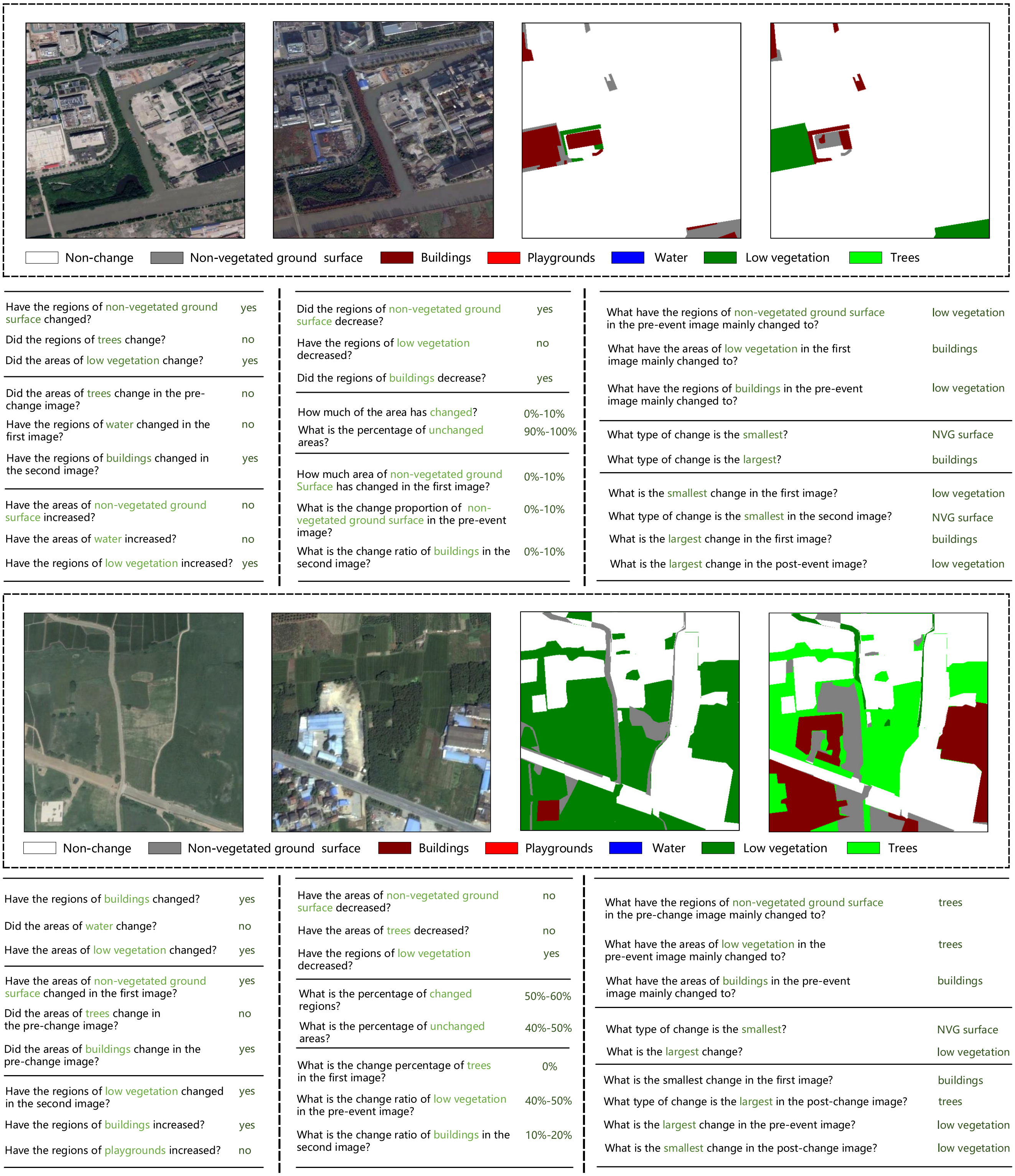
In respect of multiple inputs, Siamese networks are commonly used in many vision tasks. We denote the feature of the image of time T_1 as $\boldsymbol{F}_1 = f_1(\boldsymbol{x}_{t_1})$. Likewise, $f_2(\cdot)$ is used to obtain the encoded representation for the image of time T_2. For Siamese networks, we set the network architecture and parameters of f_1 and f_2 to be the same.

In this work, we explore the effects of different encoder networks on CDVQA. For visual feature extraction, CNNs are usually used to learn feature representations, and ResNet [33] is an important milestone in the development of CNN architectures. Thus, different scales of ResNets, e.g., ResNet-18, ResNet-101, and ResNet-152, are employed as the multitemporal encoder of our CDVQA model, aiming at studying the effects of different scales of CNNs on CDVQA.

Recently, Transformer architecture [34] has achieved excellent performance on NLP tasks [35]. Designed for sequence modeling tasks, Transformer has the significant advantage of using attention to learn long-range dependencies in data. Considering its great success in the language modeling domain, it has also been applied to computer vision tasks, to name a few, image classification [36], [37], object detection [38], and semantic segmentation [39]. In this work, the Transformer-based encoder for multitemporal images is also used.

B. Change Enhancing Module

Change detection is a fundamental task in remote sensing and also the core of CDVQA task. To answer change-related questions, a model needs to focus on changed regions and further analyze semantic information. In a number of computer vision tasks, self-attention mechanism [40], [41], [42] is used to boost the performance by focusing on important parts of data samples. However, there are two input images in our case, where the self-attention mechanism is not applicable.



Have the regions of non-vegetated ground surface changed?	yes
Did the regions of trees change?	no
Did the areas of low vegetation change?	yes
Did the areas of trees change in the pre-change image?	no
Have the regions of water changed in the first image?	no
Have the regions of buildings changed in the second image?	yes
Have the areas of non-vegetated ground surface increased?	no
Have the areas of water increased?	no
Have the regions of low vegetation increased?	yes

Did the regions of non-vegetated ground surface decrease?	yes
Have the regions of low vegetation decreased?	no
Did the regions of buildings decrease?	yes
How much of the area has changed?	0%-10%
What is the percentage of unchanged areas?	90%-100%
How much area of non-vegetated ground Surface has changed in the first image?	0%-10%
What is the change proportion of non-vegetated ground surface in the pre-event image?	0%-10%
What is the change ratio of buildings in the second image?	0%-10%

What have the regions of non-vegetated ground surface in the pre-event image mainly changed to?	low vegetation
What have the areas of low vegetation in the first image mainly changed to?	buildings
What have the regions of buildings in the pre-event image mainly changed to?	low vegetation
What type of change is the smallest?	NVG surface
What type of change is the largest?	buildings
What is the smallest change in the first image?	low vegetation
What type of change is the smallest in the second image?	NVG surface
What is the largest change in the first image?	buildings
What is the largest change in the post-event image?	low vegetation

Have the regions of buildings changed?	yes
Did the areas of water change?	no
Have the areas of low vegetation changed?	yes
Have the areas of non-vegetated ground surface changed in the first image?	yes
Did the areas of trees change in the pre-change image?	no
Did the areas of buildings change in the pre-change image?	yes
Have the regions of low vegetation changed in the second image?	yes
Have the regions of buildings increased?	yes
Have the regions of playgrounds increased?	no

Have the areas of non-vegetated ground surface decreased?	no
Have the areas of trees decreased?	no
Have the regions of low vegetation decreased?	yes
What is the percentage of changed regions?	50%-60%
What is the percentage of unchanged areas?	40%-50%
What is the change percentage of trees in the first image?	0%
What is the change ratio of low vegetation in the pre-event image?	40%-50%
What is the change ratio of buildings in the second image?	10%-20%

What have the regions of non-vegetated ground surface in the pre-change image mainly changed to?	trees
What have the areas of low vegetation in the pre-event image mainly changed to?	trees
What have the areas of buildings in the pre-event image mainly changed to?	buildings
What type of change is the smallest?	NVG surface
What is the largest change?	low vegetation
What is the smallest change in the first image?	buildings
What type of change is the largest in the post-change image?	trees
What is the largest change in the pre-event image?	low vegetation
What is the smallest change in the post-change image?	low vegetation

Fig. 5. Visualization examples of the generated CDVQA dataset. Here, we show two data samples, and each one contains bitemporal images, questions, and the corresponding answers. Best viewed in color and zoomed in.

Hence, in this work, we propose a change enhancing module to enhance the CDVQA model in terms of the capability of detecting changes.

We denote that the encoded deep features for the input two images are $\boldsymbol{F}_1 \in \mathbb{R}^{N\times C\times H\times W}$ and $\boldsymbol{F}_2 \in \mathbb{R}^{N\times C\times H\times W}$, where N is the batch size, C is the number of channels, and H and W are the height and width of feature maps, respectively. The conventional self-attention model [34] first transforms the input feature into three independent features, i.e., the query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$. In contrast, for the proposed change enhancing module, we treat the feature representations $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ as the query and key, respectively, and compute their similarity $\boldsymbol{F}_s \in \mathbb{R}^{N\times C\times H\times W}$ as follows:

$$\boldsymbol{F}_s = |f_q(\boldsymbol{F}_1) - f_k(\boldsymbol{F}_2)| \tag{1}$$

where $f_q(\cdot)$ and $f_k(\cdot)$ are 1×1 convolutions for the purpose of feature transformation. Next, a change enhancing map $\boldsymbol{M}_{\text{ce}} \in \mathbb{R}^{N\times H\times W}$ can be obtained by

$$\boldsymbol{M}_{\text{ce}} = \sigma(f_c(\boldsymbol{F}_s)) \tag{2}$$

where $f_c(\cdot)$ is a 1×1 convolution layer for predicting the change enhancing map and $\sigma(\cdot)$ is the ReLU activation function. The map $\boldsymbol{M}_{\text{ce}}$ is used to encourage the model to focus on regions where differences between $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are large. To this end, we scale $\boldsymbol{M}_{\text{ce}}$ with a parameter θ and add it with an identity matrix $\boldsymbol{I}$. θ is a learnable parameter with an initial value of 0 and is optimized during training in an end-to-end manner. Then, we multiply the transformed $\boldsymbol{M}_{\text{ce}}$ and two encoded features, respectively

$$\begin{aligned}\boldsymbol{F}_{c1} &= (\boldsymbol{I} + \theta\boldsymbol{M}_{\text{ce}})\boldsymbol{F}_1\\ \boldsymbol{F}_{c2} &= (\boldsymbol{I} + \theta\boldsymbol{M}_{\text{ce}})\boldsymbol{F}_2\end{aligned} \tag{3}$$

where $\boldsymbol{F}_{c1}$ and $\boldsymbol{F}_{c2}$ are the final encoded features corresponding to two input images. $\boldsymbol{F}_{c1}$ and $\boldsymbol{F}_{c2}$ will then be fused by the multitemporal feature fusion module, which will be introduced in Section III-C.

C. Multitemporal Fusion

After the feature encoding and change enhancing processing, we need to fuse features of time T_1 and time T_2 to obtain the final visual feature $\boldsymbol{F}_v$. For the fusion of multiple feature maps, elementwise subtraction, multiplication, summation, and concatenation are commonly used methods. Given two feature maps $\boldsymbol{F}_{c1} \in \mathbb{R}^{N\times C\times H\times W}$ and $\boldsymbol{F}_{c2} \in \mathbb{R}^{N\times C\times H\times W}$, the aforementioned fusion methods can be formulated as

$$\begin{aligned}\boldsymbol{F}_{v1} &= \boldsymbol{F}_{c1} \ominus \boldsymbol{F}_{c2}\\ \boldsymbol{F}_{v2} &= \frac{\boldsymbol{F}_{c1}}{||\boldsymbol{F}_{c1}||_2} \ominus \frac{\boldsymbol{F}_{c2}}{||\boldsymbol{F}_{c2}||_2}\\ \boldsymbol{F}_{v3} &= \boldsymbol{F}_{c1} {}^\frown \boldsymbol{F}_{c2}\\ \boldsymbol{F}_{v4} &= \boldsymbol{F}_{c1} \oplus \boldsymbol{F}_{c2}\\ \boldsymbol{F}_{v5} &= \boldsymbol{F}_{c1} \otimes \boldsymbol{F}_{c2}\end{aligned} \tag{4}$$

where $\ominus$ denotes the elementwise subtraction operation. Note that we normalize the two features before the elementwise subtraction operation for computing $\boldsymbol{F}_{v2}$. $\oplus$ and $\otimes$ denote elementwise summation and multiplication operations, respectively. $^\frown$ stands for the concatenation operation along the channel dimension. To study the effects of different fusion strategies, we compare and analyze their performance in Section IV.

D. Multimodal Fusion

Since CDVQA involves both visual features and language representations, we need to fuse multimodal features. After the multitemporal feature fusion, the final visual representation $\boldsymbol{F}_v \in \mathbb{R}^{N\times C\times H\times W}$ can be obtained. Meanwhile, an RNN is used to encode the question into feature vector $\boldsymbol{V}_q \in \mathbb{R}^{N\times L}$. As the skip-thoughts model has been applied in many remote sensing image-based NLP tasks [31], [43], we choose to use the pretrained skip-thoughts model [44] for the language feature extraction part. Specifically, skip-thought vectors are modeled with an encoder–decoder architecture, and both are constructed with RNNs. The encoder transforms the input sentence into a vector, and two decoders are used to decode the vector into the previous and the next sentence. In this work, we use the encoder of skip-thoughts for generating language embeddings.

Before fusing features of two modalities, we first transform the visual feature $\boldsymbol{F}_v$ into a feature vector $\boldsymbol{F}_{vt} \in \mathbb{R}^{N\times L}$. Then, the two feature vectors have the same size, and we can fuse $\boldsymbol{F}_{vt}$ and $\boldsymbol{V}_q$ together. As how to fuse them is not the main research content of this work, we simply merge them into a multimodal feature by concatenation

$$\boldsymbol{F}_m = \boldsymbol{V}_q {}^\frown \boldsymbol{F}_{vt} \tag{5}$$

where $\boldsymbol{F}_m \in \mathbb{R}^{N\times 2L}$ is the fused multimodal representation.

Finally, as the answer prediction is modeled as a classification task in this work, the feature $\boldsymbol{F}_m$ is used to predict the answer by passing through a classifier, i.e., two fully connected layers. The answer is given by selecting the answer class with the highest probability. The output dimension of the first layer is 256 and the final output dimension of the classifier is 19, as there are 19 answer types. Specifically, the possible answers include no, yes, 0%–10%, 0, NVG surface, buildings, low vegetation, 10%–20%, trees, 20%–30%, water, 80%–90%, 30%–40%, 90%–100%, 70%–80%, 40%–50%, 60%–70%, 50%–60%, and playgrounds (sorted by the number of samples).

IV. Experiments

A. Datasets

The CDVQA dataset is publicly available in 2968 image pairs with the size of 512×512. Based on these image pairs, there are more than 122 000 question–answer pairs generated in total. The training, validation, and test sets are split based on image pairs captured at different geographical positions. In particular, the training set contains 65 967 question–answer pairs, which are generated from 1600 (53.91%) image pairs. There are 16 441 question–answer pairs in the validation set, which are produced based on 400 (13.48%) image pairs. Besides, we use the left 968 (32.61%) image pairs to generate two test sets with 39 686 (test set 1) and 31 036 (test set 2) question–answer pairs for more comprehensive model evaluation. Note that there is an overlap between the two test sets.

B. Implementation Details

The generated dataset for CDVQA follows the same format as the work of RSVQA [31]. Regarding training parameters,

TABLE I

Numerical Results of Using Different Backbone Networks on Test Set 1 of CDVQA Dataset

Question Types	ResNet-18	ResNet-101	ResNet-152	ViT-B16
change ratio	0.3455	0.3388	0.3476	**0.3600**
class change ratio	0.7200	0.7134	0.7115	**0.7231**
change or not	0.8379	0.8387	0.8374	**0.8401**
change to what	0.5710	0.5737	0.5770	**0.5820**
increase or not	0.6913	0.6902	0.6854	**0.6952**
decrease or not	**0.7303**	0.7243	0.7275	0.7185
smallest change	0.2627	**0.2758**	0.2734	0.2710
largest change	0.4603	0.4576	**0.4669**	0.4648
Average Accuracy	0.5773	0.5766	0.5783	**0.5818**
Overall Accuracy	0.6771	0.6763	0.6766	**0.6800**

TABLE II

Numerical Results of Using Different Backbone Networks on Test Set 2 of CDVQA Dataset

Question Types	ResNet-18	ResNet-101	ResNet-152	ViT-B16
change ratio	0.3444	0.3465	0.3588	**0.3651**
class change ratio	0.7131	0.7158	0.7142	**0.7174**
change or not	0.7818	0.8363	**0.8403**	0.8353
change to what	0.5316	0.5693	0.5705	**0.5790**
increase or not	0.7003	0.6880	**0.7012**	0.6847
decrease or not	0.7142	**0.7331**	0.7264	0.7303
smallest change	0.2000	**0.2803**	0.2568	0.2592
largest change	0.3793	0.4607	0.4637	**0.4665**
Average Accuracy	0.5456	0.5787	0.5790	**0.5797**
Overall Accuracy	0.6165	0.6333	0.6336	**0.6341**

the Adam optimizer is used with an initial learning rate of 1e^{-4}. For all ResNet-based models, the batch size is set to 70, and the size of the input image is scaled to 256×256. Since the used ViT [37] model requires the input size to be 384×384, we have to reduce the batch size to 32 considering GPU memory limit. For all experiments, 50 epochs are used to train models. We utilize accuracy as a measurement for each question type. In addition, average accuracy and overall accuracy are also reported.

C. Effects of Different Backbones

The backbone network of the visual encoder is an important component. Therefore, we compare four different backbones: three ResNets (ResNet-18, ResNet-101, and ResNet-152) and a vision Transformer model ViT. In all experiments, we fuse multitemporal visual features by feature concatenation for all backbone networks.

The results on two different test sets are shown in Tables I and II. From the results, we can see that compared to ResNet-18 and ResNet-101, ResNet-152 does not show a significant performance advantage. For example, on test set 1, ResNet-18 and ResNet-152 deliver very close average and overall accuracies. This indicates that merely improving the capability of backbone network for visual learning only yields a limited gain. However, when we change the network architecture of the backbone from ResNet to Transformer, the performance can be further improved. The reason for this improvement is that the self-attention mechanism of Transformer networks is beneficial for learning more representative features. Note that the parameters of backbone networks are fixed during the training stage. In Figs. 6 and 7, we also visualize training and validation losses of models with different backbones. It can be seen that the ViT backbone has significantly lower losses than ResNet-based networks. Note that we omit the first five epochs to better compare the final convergence state.

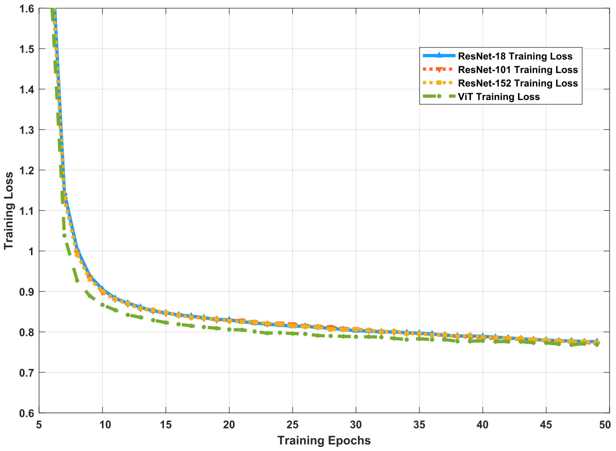


Fig. 6. Visualization of training losses. Four different backbone networks are compared.

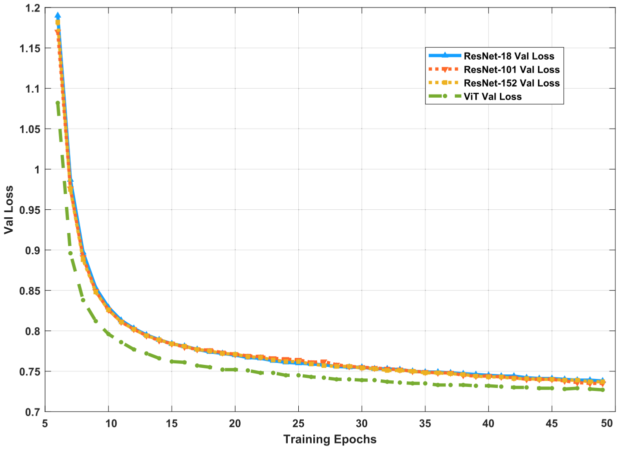
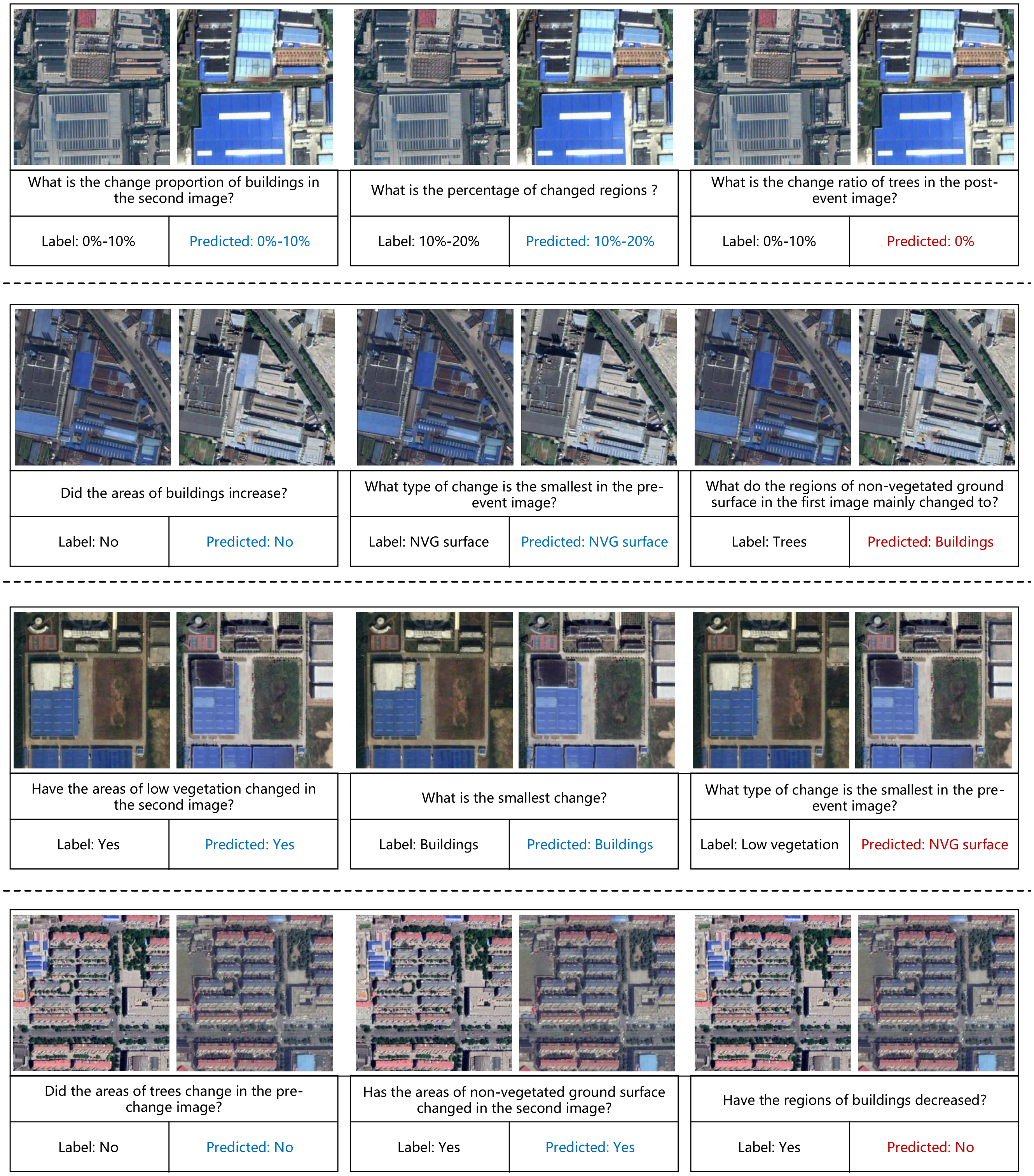


Fig. 7. Visualization of validation losses. Four different backbone networks are compared.

From the results, we can see that different backbone networks have very little impact on the performance of our framework. This is because visual feature learning may not be the key to improving accuracy. Other parts of the model, such as multitemporal fusion and change analysis part, may be more critical for the performance improvement of the CDVQA task.

D. Effects of Different Multitemporal Fusion Strategies

In this section, we quantitatively compare five commonly used feature fusion operations, namely, concatenation (Concat), summation (Sum), subtraction (Sub), normalized subtraction (NSub), and multiplication (Mul). The numerical results on two test sets are presented in Tables III and IV.



What is the change proportion of buildings in the second image?

Label: 0%-10% | Predicted: 0%-10%

What is the percentage of changed regions ?

Label: 10%-20% | Predicted: 10%-20%

What is the change ratio of trees in the post-event image?

Label: 0%-10% | Predicted: 0%

Did the areas of buildings increase?

Label: No | Predicted: No

What type of change is the smallest in the pre-event image?

Label: NVG surface | Predicted: NVG surface

What do the regions of non-vegetated ground surface in the first image mainly changed to?

Label: Trees | Predicted: Buildings

Have the areas of low vegetation changed in the second image?

Label: Yes | Predicted: Yes

What is the smallest change?

Label: Buildings | Predicted: Buildings

What type of change is the smallest in the pre-event image?

Label: Low vegetation | Predicted: NVG surface

Did the areas of trees change in the pre-change image?

Label: No | Predicted: No

Has the areas of non-vegetated ground surface changed in the second image?

Label: Yes | Predicted: Yes

Have the regions of buildings decreased?

Label: Yes | Predicted: No

Fig. 8. Visualization examples of CDVQA results. Each row presents three different questions and the same input image pair. Correctly predicted results are shown in blue and wrong answers are shown in red.

The results in these tables show that concatenation is the best. The concatenation operation first concatenates two inputs together, and then, several fully connected layers are used to fuse these inputs by learnable weights. This makes it a more flexible and general fusion strategy.

For change analysis tasks, intuitively, subtraction should be the best fusion method, as it can better highlight changed regions. However, it can be seen from the results that the subtraction operation cannot outperform others. Considering that the direct subtraction of two features may undermine their

TABLE III

NUMERICAL RESULTS OF USING DIFFERENT FUSION STRATEGIES ON TEST SET 1 OF CDVQA DATASET

ResNet-101	Concat	Sub	NSub	Mul	Sum
	$\frown$	$\ominus$	$\ominus$	$\otimes$	$\oplus$
change ratio	**0.3388**	0.3182	0.3151	0.3047	0.3352
class change ratio	0.7134	0.7129	0.7130	0.7131	**0.7167**
change or not	**0.8387**	0.8249	0.8245	0.8271	0.8378
change to what	**0.5737**	0.5269	0.5543	0.5693	0.5690
increase or not	0.6902	0.6897	0.6919	**0.6967**	0.6891
decrease or not	**0.7243**	0.6983	0.7056	0.7020	0.7239
smallest change	0.2758	0.2778	0.2789	**0.2799**	0.2703
largest change	**0.4576**	0.4473	0.4411	0.4290	0.4476
Average Accuracy	**0.5766**	0.5620	0.5655	0.5652	0.5737
Overall Accuracy	**0.6763**	0.6631	0.6657	0.6665	0.6746

TABLE IV

NUMERICAL RESULTS OF USING DIFFERENT BACKBONE NETWORKS ON TEST SET 2 OF CDVQA DATASET

ResNet-101	Concat	Sub	NSub	Mul	Sum
	$\frown$	$\ominus$	$\ominus$	$\otimes$	$\oplus$
change ratio	**0.3465**	0.3223	0.3243	0.3119	0.3409
class change ratio	0.7158	0.7097	0.7103	0.7104	**0.7169**
change or not	**0.8363**	0.8195	0.8228	0.8148	0.8347
change to what	**0.5693**	0.5275	0.5549	0.5396	0.5690
increase or not	0.6880	0.6986	**0.7146**	0.7036	0.6886
decrease or not	0.7331	0.7029	0.7046	0.7172	**0.7423**
smallest change	**0.2803**	0.2792	0.2796	0.2799	0.2679
largest change	**0.4607**	0.4442	0.4428	0.4290	0.4486
Average Accuracy	**0.5787**	0.5630	0.5692	0.5632	0.5761
Overall Accuracy	**0.6333**	0.6199	0.6244	0.6196	0.6313

TABLE V

ABLATION STUDY ON TEST SET 1 OF CDVQA DATASET FOR RESNET-101 BACKBONE

Question Types	w/o CEM	w/ CEM
change ratio	0.3388	**0.3854**
class change ratio	**0.7134**	0.7071
change or not	**0.8387**	0.8292
change to what	0.5737	**0.5804**
increase or not	0.6902	**0.7443**
decrease or not	0.7243	**0.7697**
smallest change	0.2758	**0.3127**
largest change	0.4576	**0.4777**
Average Accuracy	0.5766	**0.6008**
Overall Accuracy	0.6763	**0.6903**

representability, we normalize two input features by using ℓ_2 normalization before the subtraction operation. Nevertheless, the normalized subtraction operation is still no better than concatenation and summation. This indicates that directly subtracting two inputs is not useful to CDVQA tasks, and a specific change analysis module should be designed.

E. Effect of Change Enhancing Module

It is critical to obtain semantic change information from multitemporal images. However, there are no pixelwise ground-truth change labels available in this task. To incorporate change information into the model, we propose a change enhancing module to highlight changed regions in the input images. To validate the effectiveness of the module, we conduct an ablation study, and numerical results are shown in Tables V and VI. In the two tables, change enhancing module is abbreviated as CEM for the sake of simplification. The experimental results on both test sets indicate that the proposed change enhancing module is beneficial to the CDVQA task.

TABLE VI

ABLATION STUDY ON TEST SET 2 OF CDVQA DATASET FOR RESNET-101 BACKBONE

Question Types	w/o CEM	w/ CEM
change ratio	0.3465	**0.3904**
class change ratio	0.7158	**0.7714**
change or not	**0.8363**	0.8123
change to what	0.5693	**0.6511**
increase or not	0.6880	**0.7534**
decrease or not	0.7331	**0.7868**
smallest change	**0.2803**	0.2431
largest change	**0.4607**	0.3766
Average Accuracy	0.5787	**0.5982**
Overall Accuracy	0.6333	**0.6513**

TABLE VII

EXPERIMENTAL RESULTS IN THE CROSS-DATASET TEST SETTING

Question Types	Random init.	Ours
change ratio	0.0442	**0.0615**
class change ratio	0.0362	**0.0751**
change or not	0.0730	**0.2042**
change to what	0.0722	**0.0834**
increase or not	0.0382	**0.4713**
decrease or not	0.0837	**0.4659**
smallest change	0.0434	**0.1413**
largest change	0.0724	**0.1557**
Average Accuracy	0.0601	**0.1969**
Overall Accuracy	0.0559	**0.1560**

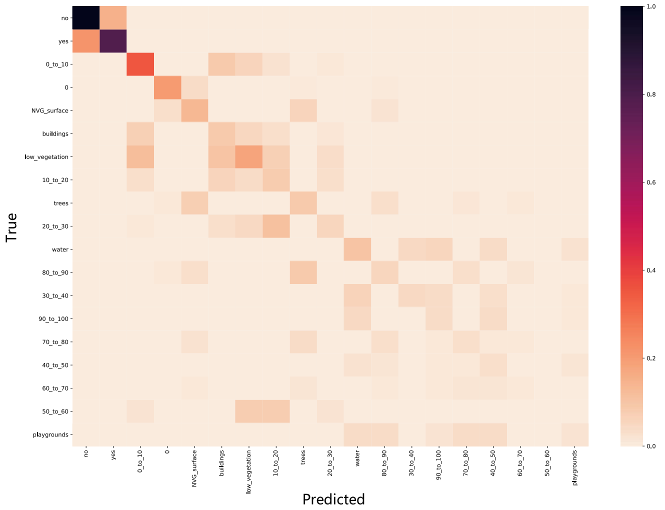


Fig. 9. Normalized confusion matrix for our CDVQA dataset on test set 1 (ResNet-152 is used as the backbone).

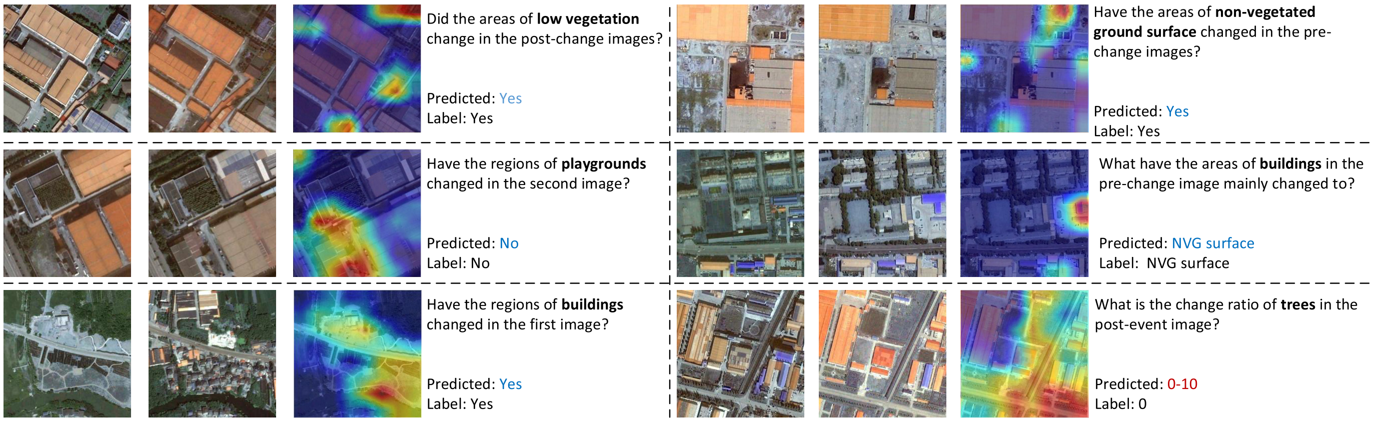

Did the areas of **low vegetation** change in the post-change images?
Predicted: Yes
Label: Yes
Have the areas of **non-vegetated ground surface** changed in the pre-change images?
Predicted: Yes
Label: Yes
Have the regions of **playgrounds** changed in the second image?
Predicted: No
Label: No
What have the areas of **buildings** in the pre-change image mainly changed to?
Predicted: NVG surface
Label: NVG surface
Have the regions of **buildings** changed in the first image?
Predicted: Yes
Label: Yes
What is the change ratio of **trees** in the post-event image?
Predicted: 0-10
Label: 0

Fig. 10. Visualization examples of change attention maps. Best viewed in color.

In particular, from the results in Tables V and VI, it can be seen that the proposed module can consistently improve both the average accuracy and overall accuracy.

F. Cross-Dataset Evaluation

In order to explore the generalization ability of the model, we construct another CDVQA dataset as an additional test set. Specifically, we collect 138 image pairs of size 256 × 256 from the HTCD [45] dataset (only binary change maps are available) and manually annotate semantic change maps. Then, 3303 question–answer pairs are generated and used for the cross-dataset test setting. To show the effectiveness of the proposed method, we compare the performance of a model trained on the CDVQA dataset and another model with randomly initialized weights. Table VII shows the numerical results. We can see that the model trained on our CDVQA dataset can be transferred to unseen scenarios, but its performance is not satisfactory in this cross-dataset test setting. This is mainly because there is a domain gap between the tests set of CDVQA and this one. We see that much more research efforts are needed in this direction.

G. Discussion

Regarding numerical results, generally, the average accuracy is lower than 60%, and the overall accuracy is lower than 70%. Some visualization examples of CDVQA results are presented in Fig. 8. Both correctly predicted examples and failures are displayed. From the experimental results, we can conclude that CDVQA is a complex and challenging task. To correctly answer different types of questions, a model first needs to learn multimodal representations for the input images and questions. Visual and language understanding is of great importance for the model. Besides, CDVQA also requires the model to be able to analyze semantic change information, i.e., the model needs to not only locate changed areas but also identify land-cover classes of changed regions to answer some complex questions. Currently, the proposed baseline framework does not make use of semantic change labels. Thus, its performance on questions related to land-cover classes is not that satisfactory. The change ratio for each land cover has higher accuracy than the change ratio for all land covers. This is mainly because the former has more training samples. We also visualize the normalized confusion matrix in Fig. 9. Note that the confusion matrix is normalized along the predicted label axis.

To better understand what the model has learned for making decisions, we visualize attention maps of our model on some examples in Fig. 10. It can be seen that the model learns to focus on the related changed regions to predict answers. From experimental results, we can conclude that more research efforts are needed to reach a satisfactory performance on the challenging CDVQA task. Specifically, more effective change analysis-based visual learning methods should be investigated. We also see that Transformer-based models have great potential for multitemporal and multimodal feature learning in CDVQA tasks. In addition, self-supervised or unsupervised change detection methods need to be studied. How to obtain the semantic change information from multitemporal data in an unsupervised manner is also an important research direction in CDVQA tasks.

V. Conclusion

To provide ordinary end users with flexible access to change information, we introduce a new task named CDVQA with natural language as output. This task takes multitemporal aerial images and a natural question as inputs to predict the corresponding answer. To be specific, we create a new dataset, which contains 2968 pairs of aerial images and more than 122 000 question–answer pairs. In addition, a baseline CDVQA model is devised, and different components of models are evaluated on the generated dataset. The experimental results outline possible problems that are needed to be addressed for the CDVQA task. This work also provides some useful insights for developing better CDVQA models, which are important for future research in this direction.

Acknowledgment

The authors acknowledge the SECOND dataset contributors from Wuhan University, Wuhan, China, and Ca’ Foscari University of Venice, Venice, Italy.

References

[1] S. Saha, F. Bovolo, and L. Bruzzone, "Building change detection in VHR SAR images via unsupervised deep transcoding," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 1917–1929, Mar. 2020.

[2] Y. Ban and O. Yousif, "Change detection techniques: A review," in *Multitemporal Remote Sensing*. Springer, 2016, pp. 19–43.

[3] Y. You, J. Cao, and W. Zhou, "A survey of change detection methods based on remote sensing images for multi-source and multi-objective scenarios," *Remote Sens.*, vol. 12, no. 15, p. 2460, Jul. 2020.

[4] M. Leenstra, D. Marcos, F. Bovolo, and D. Tuia, "Self-supervised pre-training enhances change detection in Sentinel-2 imagery," 2021, *arXiv:2101.08122*.

[5] J. Tian, A. A. Nielsen, and P. Reinartz, "Improving change detection in forest areas based on stereo panchromatic imagery using kernel MNF," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7130–7139, Nov. 2014.

[6] C. Baker, R. L. Lawrence, C. Montagne, and D. Patten, "Change detection of wetland ecosystems using landsat imagery and change vector analysis," *Wetlands*, vol. 27, no. 3, pp. 610–619, Sep. 2007.

[7] A. Schmitt and B. Brisco, "Wetland monitoring using the Curvelet-based change detection method on polarimetric SAR imagery," *Water*, vol. 5, no. 3, pp. 1036–1051, Jul. 2013.

[8] P. Washaya, T. Balz, and B. Mohamadi, "Coherence change-detection with Sentinel-1 for natural and anthropogenic disaster monitoring in urban areas," *Remote Sens.*, vol. 10, no. 7, p. 1026, Jun. 2018.

[9] H. Qiao, X. Wan, Y. Wan, S. Li, and W. Zhang, "A novel change detection method for natural disaster detection and segmentation from video sequence," *Sensors*, vol. 20, no. 18, p. 5076, Sep. 2020.

[10] A.-M. Olteanu-Raimond *et al.*, "Use of automated change detection and VGI sources for identifying and validating urban land use change," *Remote Sens.*, vol. 12, no. 7, p. 1186, Apr. 2020.

[11] B. Mishra and J. Susaki, "Sensitivity analysis for L-band polarimetric descriptors and fusion for urban land cover change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 4231–4242, Oct. 2014.

[12] B. Haack, J. Wolf, and R. English, "Remote sensing change detection of irrigated agriculture in Afghanistan," *Geocarto Int.*, vol. 13, no. 2, pp. 65–75, Jun. 1998.

[13] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag. Replaces Newslett.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[14] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

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

[16] Z. Lv, T. Liu, and J. A. Benediktsson, "Object-oriented key point vector distance for binary land cover change detection using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6524–6533, Sep. 2020.

[17] 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 Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2018.

[18] K. Yang *et al.*, "Asymmetric Siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[19] 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," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[20] A. Mogadala, M. Kalimuthu, and D. Klakow, "Trends in integration of vision and language research: A survey of tasks, datasets, and methods," 2019, *arXiv:1907.09358*.

[21] A. Khamparia, B. Pandey, S. Tiwari, D. Gupta, A. Khanna, and J. J. P. C. Rodrigues, "An integrated hybrid CNN–RNN model for visual description and generation of captions," *Circuits, Syst., Signal Process.*, vol. 39, no. 2, pp. 776–788, Feb. 2020.

[22] H. Ting-Hao *et al.*, "Visual storytelling," 2016, *arXiv:1604.03968*.

[23] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[24] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Understand.*, vol. 163, pp. 21–40, Oct. 2017.

[25] T. Tu, Q. Ping, G. Thattai, G. Tur, and P. Natarajan, "Learning better visual dialog agents with pretrained visual-linguistic representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5622–5631.

[26] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[27] Z. Yuan, L. Mou, and X. X. Zhu, "Self-paced curriculum learning for visual question answering on remote sensing data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 2999–3002.

[28] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2016, pp. 1–5.

[29] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2017.

[30] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[31] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8555–8566, Dec. 2020.

[32] 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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[35] I. V. Tetko, P. Karpov, R. Van Deursen, and G. Godin, "State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis," *Nature Commun.*, vol. 11, no. 1, pp. 1–11, Dec. 2020.

[36] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted Windows," 2021, *arXiv:2103.14030*.

[37] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[39] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," 2021, *arXiv:2105.15203*.

[40] Z. Xiong, Y. Yuan, and Q. Wang, "MSN: Modality separation networks for RGB-D scene recognition," *Neurocomputing*, vol. 373, pp. 81–89, Jan. 2020.

[41] Y. Liu *et al.*, "Dual self-attention with co-attention networks for visual question answering," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107956.

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

[43] C. Chappuis, S. Lobry, B. Kellenberger, B. Le Saux, and D. Tuia, "How to find a good image-text embedding for remote sensing visual question answering?" 2021, *arXiv:2109.11848*.

[44] R. Kiros *et al.*, "Skip-thought vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[45] R. Shao, C. Du, H. Chen, and J. Li, "SUNet: Change detection for heterogeneous remote sensing images from satellite and UAV using a dual-channel fully convolution network," *Remote Sens.*, vol. 13, no. 18, p. 3750, Sep. 2021.

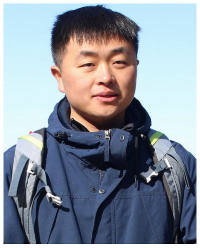
Zhenghang Yuan (Student Member, IEEE) received the B.E. and M.E. degrees from Northwestern Polytechnical University, Xi'an, China, in 2017 and 2020, respectively. She is currently pursuing the Ph.D. degree with the Technical University of Munich, Munich, Germany.

Her research interests include remote sensing, computer vision, and deep learning.

Lichao Mou received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

Since 2019, he has been a Research Scientist with DLR-IMF, Weßling, Germany, and an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU), Munich. In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. He is currently a Guest Professor with the Munich AI Future Lab AI4EO, TUM, and the Head of the Visual Learning and Reasoning Team, Department "EO Data Science," Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany.

Dr. Mou was a recipient of the First Place in the 2016 IEEE GRSS Data Fusion Contest and a finalist for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and 2019 Joint Urban Remote Sensing Event.



Zhitong Xiong (Member, IEEE) received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2021.

He is currently a Post-Doctoral Fellow with the Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany. His research interests include computer vision, machine learning, and remote sensing.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently a Professor with the Data Science in Earth Observation (former: Signal Processing in Earth Observation), TUM, and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School. Since 2019, she has been heading the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has also been the Co-Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, in 2009; Fudan University, Shanghai, China, in 2014; The University of Tokyo, Tokyo, Japan, in 2015; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2016. She is currently a Visiting AI Professor with ESA's Phi-lab, Frascati, Italy. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, e.g., global urbanization, UN's sustainable development goals (SDGs), and climate change.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves on the Scientific Advisory Board of several research organizations, including the German Research Center for Geosciences (GFZ) and Potsdam Institute for Climate Impact Research (PIK). She is also an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She serves as an Area Editor for the Special Issue of *IEEE Signal Processing Magazine*.