Remote Sensing Image Change Captioning with Dual-branch Transformers: A New Method and a Large Scale Dataset

Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi*, *Member, IEEE*

***Abstract*—Analyzing land cover changes with multi-temporal remote sensing (RS) images is crucial for environmental protection and land planning. In this paper, we explore Remote Sensing Image Change Captioning (RSICC), a new task aiming at generating human-like language descriptions for the land cover changes in multi-temporal RS images. We propose a novel Transformer-based RSICC model (RSICCformer). It consists of three main components: 1) a CNN-based feature extractor to generate high-level features of RS image pairs, 2) a dual-branch Transformer encoder to improve the feature discrimination capacity for the changes, and 3) a caption decoder to generate sentences describing the differences. The dual-branch Transformer encoder consists of a hierarchy of processing stages to capture and recognize multiple changes of interest. Concretely, we use the bi-temporal feature differences as keys to enhance image features (queries) from each temporal image in the dual-branch Transformer encoder. To explore the RSICC task, we build a large-scale dataset named LEVIR-CC, which contains 10077 pairs of bi-temporal RS images and 50385 sentences describing the differences between images. We benchmark existing state-of-the-art synthetic image change captioning methods on the LEVIR-CC dataset, and our RSICCformer outperforms previous methods with a significant margin (+4.98% on BLEU-4 and +9.86% on CIDEr-D). The attention visualization results also suggest that our model can focus on changes of interest and ignore irrelevant changes.**

***Index Terms*—Remote sensing images, change detection, image captioning, change captioning, Transformer.**

I. Introduction

THE development of human society and the evolution of the natural environment are accelerating global surface changes. The availability of multi-temporal remote sensing (RS) images provides an opportunity to study surface changes [1]–[3]. The RS image change detection (RSICD) determining pixel-level changed regions has become an emerging remote sensing image interpretation task [3]–[6]. However, the pixel-level changes cannot directly reveal high-level semantic information such as object attributes and the relationship between objects in change regions [7], which requires much human effort to interpret. Therefore, it is necessary to explore methods to describe high-level semantic changes automatically. In this paper, we explore a new task in the remote sensing field, namely Remote Sensing Image Change Captioning (RSICC), which aims to describe the changes in natural language by comparing the RS images captured at different time points. The task can be significant for many applications, such as damage assessment [1], [2], environmental protection [3], [4], and land planning [5], [6].

Image change captioning (CC), which aims at understanding high-level semantic changes in the images and describing them with human language, is an emerging research topic in the computer vision (CV) field. The task involves both vision and language. At the visual analysis stage, the task requires the model to determine whether changes have occurred and to locate and recognize the changed objects and the change types, such as appearing, disappearing, increasing, and decreasing. At the language generation stage, the task requires the model to translate visual features into grammatically compliant sentences. The task has been recently studied in some specific application scenarios, such as monitoring scene CC [7]–[9], 3D scene CC [10], [11], and synthetic image CC [12]–[17]. For instance, Oluwasanmi *et al.* [8] use a siamese Convolutional neural networks (CNNs) to extract the feature discrepancies of an image pair, and then combine a soft-attention mechanism [18] and Long Short-Term Memory (LSTM) [19] to generate semantically associated sentences. Park *et al.* [12] proposed a Dual Dynamic Attention (DUDA) model in which a spatial attention module localizes change regions to enhance image features, and a dynamic attention module utilizes generated words to adaptively focuses on "before", "after", or "difference" feature representations for captioning. Kim *et al.* [16] designed a difference encoder to model viewpoint change, and designed a cycle consistency module fusing the generated caption and before image features. They then minimized the loss between the resulting features and the after image features.

Recently, some large synthetic datasets (e.g., the CLEVR-Change dataset [12], CLEVR-Multi-Change [17], and CLEVR-DC dataset [16]) have provided fundamentals and testbeds for designing and evaluating change captioning methods. However, compared to the synthetic scenes, the real remote sensing scenes have a broader scale range, richer object categories, and more complex ground details. That increases the difficulty of the change captioning with remote sensing

The work was supported by the National Natural Science Foundation of China under the Grant 62125102, and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Chenyang Liu, Rui Zhao, Hao Chen, and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, also with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China. Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China.

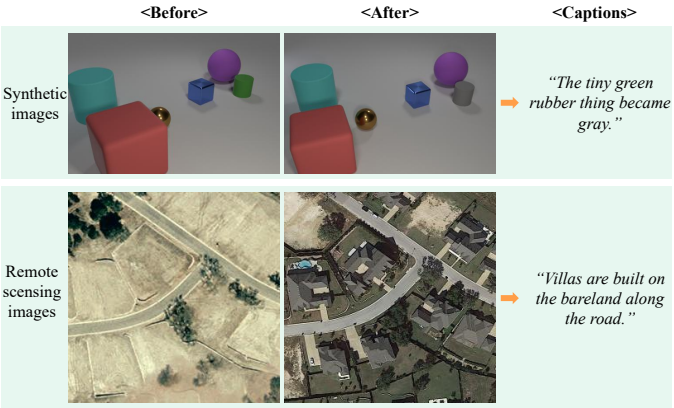


Fig. 1. Illustration of change captioning. The synthetic images of the CLEVR-Change dataset are automatically generated by the CLEVR engine [20], and the object categories are monotonous. However, RS images of our dataset have a broader scale range, richer object categories, and more complex ground details. Besides, unlike the synthetic image CC dataset, the bi-temporal RS images of our dataset are well registered, so there is no viewpoint change.

images. Besides, the remote sensing community currently lacks a publicly available large dataset for the RSICC task. Therefore, we build a large-scale dataset with real remote sensing images and human-annotated sentences, which bridges the gap between synthetic visual scenes and real-world applications. Fig. 1 shows an example comparison between the previous synthetic datasets and our dataset.

Our large-scale LEVIR-CC dataset contains 10077 pairs of bi-temporal RS images and 50385 sentences describing the differences between images. The novel dataset provides an opportunity to explore models that align visual changes and language. In this paper, we refer to relatively small and uninteresting changes as irrelevant changes, such as light change, which bring significant challenges to change captioning. To achieve RSICC, a robust model needs to 1) distinguish the changes of interest and the irrelevant changes in the complex scene, 2) recognize multiple changed objects and corresponding change types, 3) describe visual changes via language.

Two recent research works have explored the RSICC task on small datasets [21], [22]. They used CNNs to extract features of bi-temporal images. Some fusion strategies are designed to fuse the bi-temporal features, and then the fused features are directly sent into recurrent neural networks (RNNs) or support vector machines (SVMs) [23] to generate sentences. However, the feature difference is not sufficiently exploited to capture and recognize changes of interest. In this paper, we propose a novel Transformer-based RSICC (RSICCformer) model to address the problem. Our RSICCformer consists of three main components: a CNN-based feature extractor, a dual-branch Transformer encoder, and a caption decoder. The CNN-based feature extractor generates bi-temporal image features. The dual-branch Transformer-based encoder consists of a hierarchy of processing stages where in each stage, there are three modules: 1) the difference encoding module utilizing bi-temporal features to obtain the semantic features revealing the differences between two images, 2) the siamese cross-encoding (CE) modules utilizing the feature difference to capture and recognize multiple changes of interest, and 3) the multi-stage bi-temporal fusion (MBF) module utilizing the features from different stage siamese CE modules to obtain better high-level semantic feature representations revealing multiple changes of interest and excluding irrelevant changes. For the caption decoder, a Transformer decoder utilizes the features from the dual-branch Transformer encoder to generate sentences describing the differences between images.

On a related line, Qiu *et al.* [17] proposed MCCFormers-S and MCCFormers-D for the synthetic image CC task. To correlate the regions of two images with viewpoint changes, MCCFormers employ the transformer encoder with multi-head attention computing patch-level similarity. Specifically, MCCFormers-S concatenates bi-temporal image features extracted by CNNs and then sends them directly into a Transformer encoder for dense feature interaction. MCCFormers-D uses siamese transformer encoders with the co-attention mechanism [24] to capture relevance between local regions of two images. For the co-attention mechanism, the query is from the features of one image, and the key and value are from the features of the other image. Unlike MCCFormers, since image pairs of our dataset are well registered (without viewpoint change), we can utilize the difference between the corresponding positions of bi-temporal images to capture changes, which is a valuable prior. Therefore, we designed siamese CE modules with the cross-attention mechanism in RSICCformer to utilize the difference information, in which the query is from the features of one image, and the key and value are from the difference features. Experiments show that this strategy effectively improves the model's feature representation and discriminative ability.

Our contributions can be summarized as follows:

- We conduct an in-depth study of an emerging task named Remote Sensing Image Change Captioning (RSICC) that aims at generating human-like language descriptions for the ground feature changes in multi-temporal RS images. The RSICC provides significant application prospects, such as damage assessment, environmental protection, and land planning, and may help explore models to align visual changes and language in RS images.
- We propose a novel Transformer-based RSICC model (RSICCformer). We combine siamese CE modules utilizing the feature difference and MBF modules to obtain high-level semantics revealing multiple changes. Extensive experiments show the effectiveness of our method. The attention visualization results also suggest that our model can capture multiple changes of interest and ignore irrelevant changes.
- We build a publicly available large-scale dataset named LEVIR-CC to advance the change captioning task in the remote sensing community. We benchmark existing state-of-the-art synthetic image change captioning methods on the LEVIR-CC dataset, and our RSICCformer outperforms these methods with a significant margin (+4.98% on BLEU-4 and +9.86% on CIDEr-D).

II. Related work

The RSICC task can be considered a combination of change detection and image captioning. Unlike change detection requiring pixel-level change localization, change captioning requires understanding changes at the semantic level, including the attributes of the changed object and the positional relationship between objects. Unlike image captioning describing the visual content of interest in a single image, change captioning requires describing the visual changes of interest in image pairs. Here we discuss previous work on remote sensing image change detection and remote sensing image captioning.

A. Remote Sensing Change Detection

The remote sensing change detection (CD) task aims at locating and recognizing pixel-level changes in the bi-temporal RS images. The output of the CD system is a binary change map indicating change regions or a semantic change map indicating the change type of each pixel.

The early traditional change detection methods can be divided into three main categories: 1) algebra-based method, 2) transformation-based method, and 3) classification-based method. The algebra-based methods perform the algebraic operation on multi-temporal RS images to obtain the change maps, such as image difference [25], image ratio [26], and change vector analysis (CVA) [27], [28]. For example, Malila et al. [27] employed the CVA method to calculate the change intensity and change direction between pixels, and then the change regions are discriminated by setting reasonable thresholds. The transformation-based methods employ data reduction methods to suppress correlated information and extract effective features from the original multi-temporal RS images for change discrimination. Commonly used methods mainly include principal component analysis (PCA) [29], [30], multivariate alteration detection (MAD) [31], [32], and Gramm–Schmidt [33]. For example, Deng *et al.* [29] used PCA to suppress correlated information and highlight variance in multi-temporal images for recognizing change areas. Nielsen *et al.* [32] improved MAD [31] by iterating the weights of different observations to highlight changes. The classification-based methods obtain the change map by post-classification (i.e., comparing multiple classification maps) [34]–[36] or direct classification (i.e., performing classification on the data stack consisting of multi-temporal images) [37], [38]. For example, Bruzzone *et al.* [36] proposed a land-cover change detection system composed of multiple classifiers and a compound classification decision strategy. Tewkesbury *et al.* provided a comprehensive review of traditional change detection methods in [39].

Due to superior feature representation capability compared to traditional methods, deep learning (DL) has recently achieved great success in change detection. Most methods use DL-based extractors to extract high-level semantic features from raw multi-temporal images and use DL-based classifiers to generate change maps. Commonly used networks of extractors and classifiers include CNNs [40]–[42], RNNs [43], [44], transformers [45], [46], and generative adversarial networks (GANs) [47], [48]. For the CNN-based methods, Zhang *et al.* [40] explored a CNN-based siamese network for change detection and used the weighted contrastive loss to optimize the network. Peng *et al.* [41] proposed an improved UNet++ model with dense skip connections for the multi-scale feature extraction and a multiple side-output fusion strategy for deep supervision. Daudt *et al.* [42] proposed three fully convolutional networks (FCNs) for change detection and tried to modify the FCN encoder-decoder into a siamese architecture with skip connections. Many recent works have introduced attention mechanisms to improve feature representation and discrimination ability [5], [49]–[52]. For instance, Huang *et al.* [52] proposed a multiple attention Siamese network (MASNet), in which the attention feature fusion module (AFFM) is utilized to fuse features from different layers and branches in the Siamese network. Jiang *et al.* [50] proposed a pyramid feature-based attention-guided Siamese network (PGA-SiamNet), in which a co-attention module captures the correlation between bi-temporal images and a context fusion strategy is designed to fuse low-level and high-level features. For the RNN-based methods, Sun *et al.* [43] introduced the conventional long short-term memory (Conv-LSTM) [53] layers into UNet [54] to fuse spatial and temporal features for change detection. Chen *et al.* [44] used deep siamese CNNs to extract spatial-spectral features from both homogeneous and heterogeneous RS images, and used multiple-layer LSTM to map the extracted features into a new feature space and fully excavate the change information. For the transformer-based methods, Chen *et al.* [45] proposed a bi-temporal image Transformer (BIT), in which a semantic tokenizer represents the images as a few visual words, and the Transformer utilizes them to refine the bi-temporal features extracted by CNN. BIT can efficiently and effectively identify the change of interest. Zhang *et al.* [46] proposed a pure Transformer change detection network, named SwinSUNet, with a siamese U-shaped structure. Different from the convolution unit of previous CNN-based methods, the basic unit of SwinSUNet is the Swin transformer block [55]. For the GAN-based methods, Chen *et al.* [47] proposed an instance-level change augmentation method based on GAN to synthesize new training samples containing generated bi-temporal images with building-related changes and corresponding masks. Zhao *et al.* [48] proposed an attention gates generative adversarial adaptation network (AG-GAAN) for change detection. In the AG-GAAN model, image pairs are input into the generator to generate predictive change maps, and the discriminator distinguishes the change maps and ground truth labels. The generator employs AGs, a spatial constraint mechanism, to locate change regions and suppress irrelevant interference. Two comprehensive surveys on the DL-based change detection methods can be found in [56], [57].

B. Remote Sensing Image Captioning

Remote sensing image captioning is an active research topic at the intersection of remote sensing image processing and natural language processing (NLP). Unlike object detection and image segmentation, the image captioning task aims to understand the high-level semantic information of the visual

images and describe the scenes and ground objects in language as a human.

The early captioning techniques used template-based methods [58]–[60] and retrieval-based methods [61]–[66]. The template-based methods detect ground objects and fill the corresponding words into a human-designed template. For example, Shi *et al.* [58] employed a fully convolutional network (FCN) to extract ground elements of three levels, which helps recognize objects of different sizes. The retrieval-based methods require a large database to retrieve the most similar image to a given image, and then output its corresponding annotation sentence. For example, Wang *et al.* [66] embedded the image and corresponding sentence into a common semantic space and minimized the distance metric measuring similarity between them during the training phase. The model retrieves a candidate sentence whose representation is the closest to the given image representation during the inference phase. However, the template-based and retrieval-based methods cannot generate novel sentences, and the sentences are relatively limited and rigid.

Current image captioning methods are mainly based on deep learning-based generative models. The most widely used framework is the encoder-decoder framework [67]–[77], in which the visual encoders extract image features and the language models as decoders use the features to perform cross-modal generation for captioning. In the visual encoding stage, CNNs and Vision Transformers are usually used as backbone networks. In the language generation stage, many language models are used to utilize image features for captioning, such as RNNs in [70], [78], [79], SVMs in [80], and Transformers in [74], [81]–[83].

Qu *et al.* [67] first explored CNNs as encoders and RNNs as decoders to generate sentences describing RS images. Lu *et al.* [84] provided a large dataset named RSICD for RS image captioning and compared the effects of hand-crafted and CNN-based features based on the encoder-decoder framework. RS images have many objects of different sizes. Most current methods aim to enhance the visual representation ability of the model. Designing different attention modules is a commonly used approach. To capture multi-scale object information, Ma *et al.* [71] used multi-head attention to obtain the context feature from different layers and chose object detection as an auxiliary task to obtain target-level features. Li *et al.* [72] proposed a three-level attention model, which contains the attention to the image regions, the attention to words, and the attention to vision and semantics. Unlike many attention-based methods that only establish the relationship between the local features, Zhang *et al.* [85] proposed a global visual feature-guided attention (GVFGA) module. The module performs a mean pooling operation to obtain global features, and and an attention gate and feature fusion strategy are designed to fuse global and local features. To exploit structured spatial relations of semantic contents, Zhao *et al.* [70] proposed the structured attention to utilize pixel-level image segmentation region proposals extracted by the selective search method. Besides, the method can deal with image captioning and segmentation under a unified framework. Recently, some transformer-based captioning methods have been proposed, inspired by the effectiveness of transformers in the field of CV and NLP. For example, Liu *et al.* [74] used a Transformer [86] encoder to process grid-based visual features extracted by CNN and used LSTM to aggregate the features from different Transformer encoding layers. Chen *et al.* [81] proposed TypeFormer, in which a multi-scale vision transformer is used to capture multi-scale object information, and a user-oriented controller is designed to control the types of generated sentences. A comprehensive survey on image captioning can be found in [87], [88].

III. Proposed LEVIR-CC Dataset

Our LEVIR-CC dataset (LEVIR Change Captioning dataset) contains bi-temporal images and corresponding sentences describing differences[1]. We will make the LEVIR-CC dataset publicly available, and we believe the dataset will promote the research of RSICC. Our dataset and code will be publicly available at: *https://github.com/Chen-Yang-Liu/RSICC*

A. Image Pairs Collection

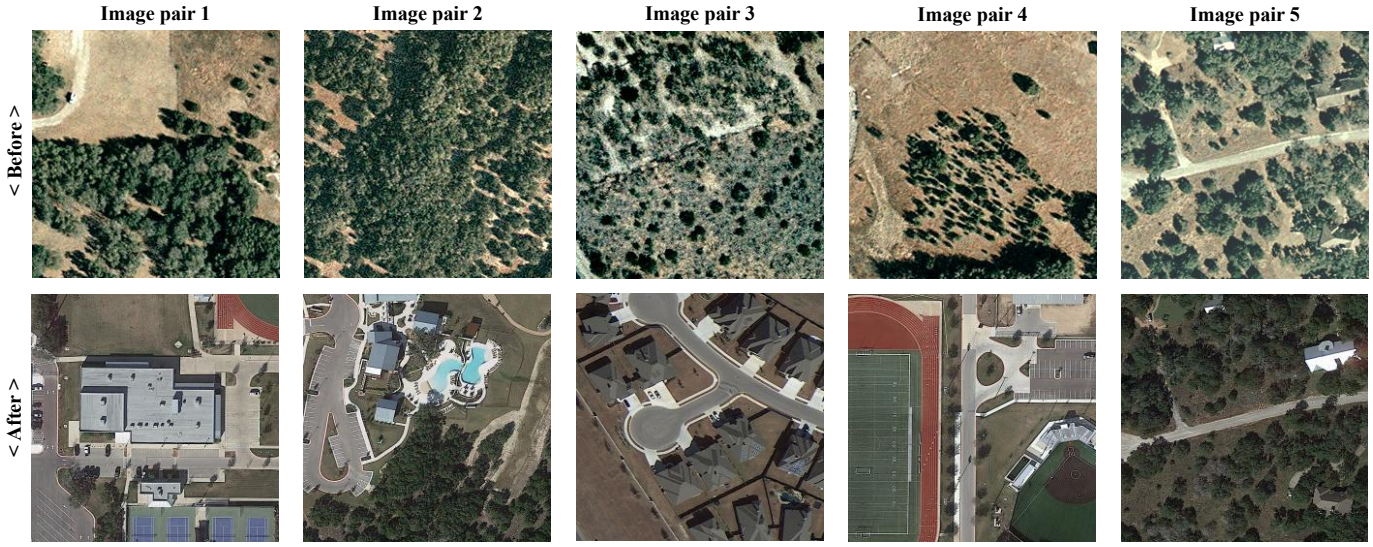
The images of the LEVIR-CC dataset are mainly from the change detection dataset LEVIR-CD [5], where each image has a spatial size of 1024×1024 pixels with a high resolution of 0.5 m/pixel. These bi-temporal images are from 20 regions in Texas, USA, and have a time span of 5∼15 years. Since each image pair in the LEVIR-CD dataset contains very dense ground objects and changes, it is difficult to describe the changes accurately and adequately in a few sentences. Therefore, we crop the bi-temporal images to 256×256 pixels in our LEVIR-CC dataset. Besides, the bi-temporal RS images are well registered pixel-by-pixel, so our dataset has no viewpoint change.

B. Captions Collection

The LEVIR-CC dataset contains 10,077 pairs of images with the size of 256×256 pixels. For each image pair, we collected five annotated sentences provided by five different annotators to describe the differences between images. Since the bi-temporal RS images in our dataset have a long time span and irrelevant changes are common in RS images, a pair of RS images cannot be exactly the same. We consider the image pairs with only irrelevant changes as no-change image pairs. We formulated the following annotation guidelines for annotators:

- Describe significant changes (e.g., ground object changes) and ignore irrelevant changes and unimportant distractors (e.g., light changes).
- Describe the changed objects and the change types, such as appearing and disappearing.
- When describing changes of interest, avoid using meaningless phrases, such as "there is".
- The following five sentences are used as no-change annotations: "The two scenes seem identical", "The scene is

[1]LEVIR is our laboratory name: the Learning, Vision, and Remote Sensing Laboratory.



(1) The mall with a parking lot is built next to a playground in place of the trees
(2) Trees and bareland are removed and a building with a parking lot and a playing field appears
(3) The trees are removed and a large building with a parking lot and a playground is built
(4) A large building replaces the woods
(5) A large building with parking lots and playgrounds replace the plants

(1) Some villas with a swimming pool and a parking lot appear and many trees are removed
(2) The woods are removed and some buildings with a parking lot are built
(3) Many trees are replaced by several houses with swimming pools and a large parking lot
(4) Villas replace the forest and only a small part of the forest is left
(5) Most of the woods are replaced by a group of buildings with parking lots and swimming pools

(1) All the trees disappear and a t-shaped road runs across a residential area
(2) Trees are removed and a road with a branch is built with villas built along
(3) A road has been built across the forest and many buildings have been constructed on both sides of the road
(4) Trees disappear and houses are built around the corner of the road
(5) Many houses along the road appear in the desert

(1) The bareland with some plants have been replaced by a road with a playground a parking lot and a golf course around
(2) Trees and bareland are replaced by a playing field with a playground on the left and a parking lot on the right
(3) A road is built and a playground and a parking lot are built on both sides of the road
(4) Many sports venues are built on open spaces
(5) Two playgrounds with buildings nearby and a parking lot appears in the desert

(1) The two scenes seem identical
(2) The scene is the same as before
(3) No change has occurred
(4) Almost nothing has changed
(5) There is no difference

Fig. 2. Some examples in the LEVIR-CC dataset, where each image has a spatial size of 256×256 pixels with a high resolution of 0.5 m/pixel. Each image pair has five annotated sentences describing the differences between images. Regarding changes of interest, unlike LEVIR-CD, which only focuses on building-related changes, our dataset focuses on multiple change types, such as buildings, roads, and rivers. Besides, we consider the image pairs with only irrelevant changes as no-change image pairs. For example, in the fifth image pair, we ignore the light change, shadow change, small tree-related changes, etc.

the same as before", "There is no difference", "No change has occurred", "Almost nothing has changed".

After collecting all the annotations, we checked and fixed spelling and grammatical errors in all annotation sentences. Finally, our dataset contains a total of 50,385 sentences. Fig. 2 shows some examples in our dataset. Regarding changes of interest, unlike LEVIR-CD, which only focuses on building-related changes, our dataset focuses on multiple changed scenes and objects, such as buildings, roads, playgrounds, and rivers. Besides, we show a no-change example as the fifth image pair in Fig. 2. We ignore the light change, shadow change, small tree-related changes, etc.

C. Dataset Analysis

1) Image pairs: The LEVIR-CC dataset contains 5038 image pairs with changes and 5039 image pairs without changes. Tab. I reports the number of image pairs in training, validation, and test sets. The proportion of image pairs with and without change is almost the same in the three sets. This demonstrates that the distribution of image pairs in the three sets is similar. Besides, we artificially counted the number of image pairs from the perspective of changed object categories: 1) building-related changes (e.g., residential area, villa), 2) parking lot changes, 3) road-related changes (e.g., crossroad, path), 4) vegetation-related changes (e.g., tree, bush), 5) water-related changes (e.g., river, lake). Tab. II reports the result. The mean number of changed object categories per image pair with changes is 1.71. We can observe many building-related changes, which are related to human activity and urban sprawl.

TABLE I
THE NUMBER OF BI-TEMPORAL IMAGE PAIRS IN THREE SETS OF THE LEVIR-CC DATASET.

Change or not	Training	Validation	Test	Total
Change	3407	667	964	5038
No change	3408	666	965	5039
Total	6815	1333	1929	10077

2) Sentences and words: Our dataset contains 50385 sentences describing the differences between images. We report the sentence quantity, the word quantity, and the average sentence length in the LEVIR-CC dataset, as shown in Tab. III.

TABLE II
THE NUMBER OF IMAGE PAIRS FOR EACH CHANGED OBJECT CATEGORY. M.C. REFERS TO THE MEAN NUMBER OF CHANGES PER IMAGE PAIR WITH CHANGES.

Changed object category	Training	Validation	Test	Total	Proportion
Building	2919	558	844	4322	85.8%
Road	1683	347	515	2545	50.5%
Parking lot	197	33	56	286	5.7%
Vegetation	796	294	273	1363	27.1%
Water	72	8	18	98	1.9%
M.C.	1.67	1.86	1.80	1.71	–

TABLE III
STATISTICS ON THE SENTENCE QUANTITY, THE WORD QUANTITY, AND THE AVERAGE SENTENCE LENGTH IN THE LEVIR-CC DATASET.

Change or not	Sentence quantity	Word quantity	Average length
Change	25190	276537	10.98
No change	25195	125975	5.00
Total	50385	402512	7.99

The 25195 sentences describing image pairs without changes are relatively short, and the average sentence length is 5 words. The average length of 25190 sentences describing image pairs with changes is about 11 words. Besides, we count the percentages of sentences of different lengths, as shown in Fig. 3. We can see that the sentence length distribution in three sets of our dataset is similar and most of these sentences are between 5 and 15 words in length.

The annotated sentences of our dataset contain a total of 402512 words, as shown in Tab. III. To more intuitively show the frequency of each word, we build a word cloud map based on the word frequency, as shown in Fig. 4. We have removed stop words (e.g., the, a, is) to focus more on informative words. The larger the word size, the more frequently the word appears in the annotated sentences. We can observe that some words, such as "nothing" and "no", appear very frequently. This is because the words are often used in sentences describing image pairs without changes, which make up half of the dataset. Besides, as we know, building-related changes often occur at the ground surface due to the influence of human activities. Therefore, we can see that building-related words appear more frequently, as shown in Fig. 4.

3) Human agreement: For captioning-related tasks, humans have many equivalent language expressions to describe the same thing. Human agreement [7], [89] means that these different expressions reflect the same essence and meaning. To quantify the human agreement between five annotated sentences for the same image pairs in the LEVIR-CC dataset, we report BLEU-N [90], ROUGE-L [91], and METEOR [92] by conducting the following experiments: one sentence is chosen as a hypothesis sentence while the remaining four are references. We compare the human agreement of our dataset with that of MS-COCO C5 [89], which is a large manually annotated image captioning dataset with five annotated sentences per image. Tab. IV shows the comparison results. We can see that our dataset is reasonable and comparable to MS-COCO C5 in terms of human agreement.

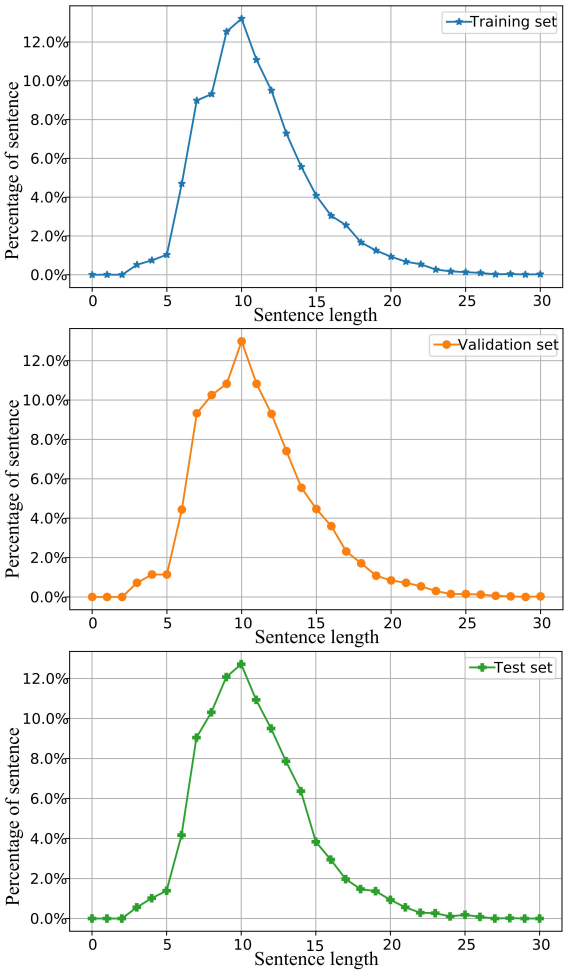


Fig. 3. The percentages of sentences of different lengths for 25190 sentences describing image pairs with changes in three set. The sentence length distribution in three sets is similar and most of these sentences are between 5 and 15 words in length.

TABLE IV
HUMAN AGREEMENT COMPARISON OF LEVIR-CC AND MS-COCO C5 DATASET. WE REPORT BLEU-N, ROUGE-L, AND METEOR WHEN ONE SENTENCE IS CHOSEN AS A HYPOTHESIS SENTENCE WHILE THE REMAINING FOUR ARE REFERENCES.

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
MS-COCO C5	66.3	46.9	32.1	21.7	25.2	48.4
LEVIR-CC	65.0	46.7	32.5	22.6	23.3	45.6

4) Dataset comparison: Tab. V compares our dataset with three existing change captioning datasets. For the CLEVR-Change [12] and CLEVR-Multi-Change [17] datasets, images are synthetic, and object categories are monotonous. The viewpoint change is a common distractor in their image pairs. Besides, annotation sentences are template-based, so they are relatively rigid. For the Spot-the-Diff [7] dataset, image pairs are frames at different moments extracted from surveillance video in the real world. The Spot-the-Diff dataset and our LEVIR-CC dataset have well-aligned bi-temporal images, so

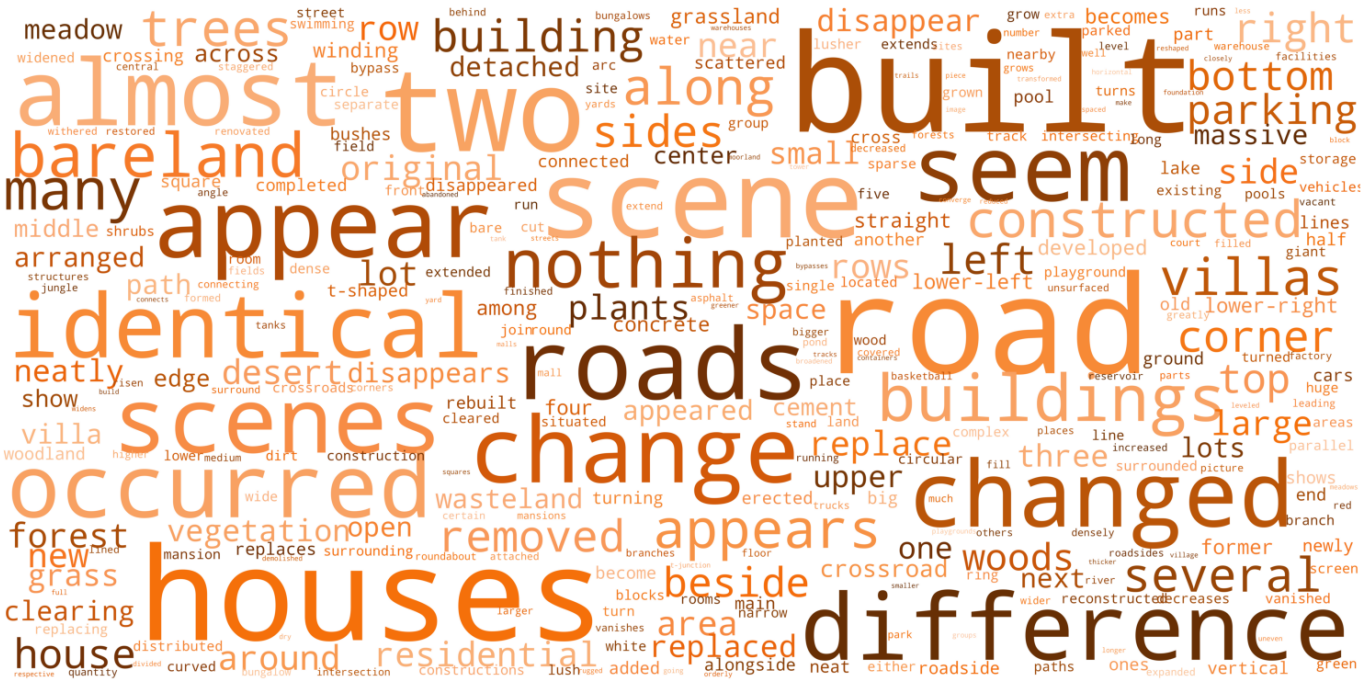


Fig. 4. Word cloud map based on the word frequency in the LEVIR-CC dataset. The larger the word size, the more frequently it appears in the annotated sentences. Note that we removed stop words (e.g., the, a, is) to focus more on informative words. Some words, such as "nothing" and "identical", appear very frequently. This is because the words are often used in sentences describing image pairs without changes, which make up half of the dataset. Besides, as we know, building-related changes often occur at the ground surface due to the influence of human activities. Therefore, building-related words appear more frequently.

TABLE V
THE COMPARISON OF OUR LEVIR-CC WITH THREE EXISTING CHANGE CAPTIONING DATASETS.

Dataset	Real world	Viewpoint change	Flexible sentences	Time span	Number of image pairs
CLEVR-Change [12]	×	✓	×	–	79,606
CLEVR-Multi-Change [17]	×	✓	×	–	60,000
Spot-the-Diff [7]	✓	×	✓	0 ∼ 8.5 hours	13,192
LEVIR-CC	✓	×	✓	5 ∼ 15 years	10,077

there is no viewpoint change. All sentences of the two datasets are generated by humans, so they are more flexible than the other two datasets. Besides, remote sensing images of our LEVIR-CC dataset are taken from the God perspective and have a longer time span. The images have a broad scale range and complex ground information. Our novel dataset provides the opportunity to align remote sensing image changes and human language. This dataset will advance the change captioning task in the remote sensing community.

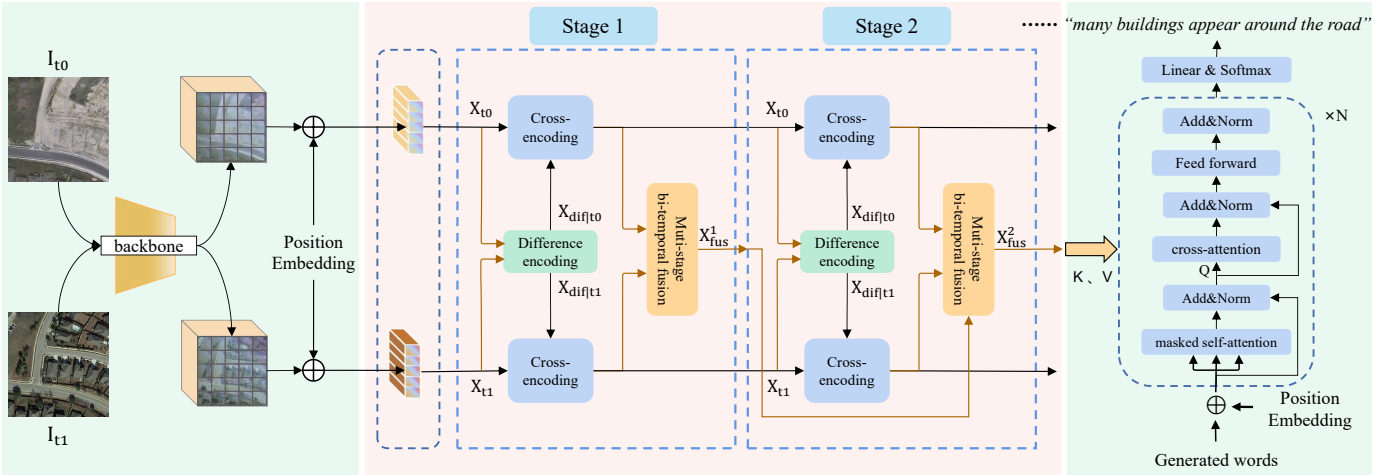
IV. METHODOLOGY

Fig. 5 shows an overview of our RSICCformer-based model. The proposed model consists of three main components: 1) a CNN-based feature extractor to generate high-level features of RS image pairs, 2) a dual-branch Transformer encoder (DTE) to improve the feature discrimination capacity for the changes, and 3) a caption decoder to generate sentences describing the differences. The dual-branch Transformer encoder consists of a hierarchy of processing stages to capture and recognize multiple changes of interest. Concretely, we use the bi-temporal feature differences as keys to enhance image features (queries) from each temporal image in each stage. With the above design, our method can capture the relationship between two images, recognize changes of interest and generate correct language descriptions.

The procedure of our RSICCformer-based model is shown in Algorithm 1.

A. Difference Encoding Module

The difference information can reflect the change regions and the degree of change, which is valuable for the model to capture changes of interest and ignore irrelevant changes. The difference encoding module utilizes bi-temporal features to obtain the high-level semantic features revealing the differences between two images. As shown in Fig. 5, the output of the difference encoding module contains $X_{dif|t_0}$ and $X_{dif|t_1}$. We define $X_{dif|i}$ $(i = t_0, t_1)$ as the difference representation, which will be fed into the siamese CE modules along with



(a) Feature Extraction (b) Dual-branch Transformer Encoder (c) Caption Decoder

Fig. 5. The overall structure of our RSICCformer-based model. The proposed model consists of three main components: a CNN-based feature extractor, a dual-branch Transformer encoder, and a caption decoder. The CNN-based feature extractor generates bi-temporal image features. We flatten the bi-temporal feature maps into two sequences respectively, add the learnable position embedding and then feed them into the dual-branch Transformer encoder (DTE). Finally, a Transformer decoder as the caption decoder utilizes the features from the DTE to generate sentences. Concretely, the DTE consists of a hierarchy of processing stages where in each stage, there are three modules: a difference encoding module, a siamese cross-encoding module, and a multi-stage bi-temporal fusion module.

Algorithm 1 The procedure of our RSICCformer-based Model.

Input: $I = (I_{t_0}, I_{t_1})$ (a pair of bi-temporal images)
Output: $Caption$
Define: $DE \leftarrow$ the difference encoding module
$CE \leftarrow$ the cross-encoding module
$MBF \leftarrow$ the multi-stage bi-temporal fusion module
$CD \leftarrow$ the caption decoder
$EM \leftarrow$ the word embedding
// step1: Feature Extraction
for i in (t_0, t_1) **do**
  $X_i = backbone(I_i)$
end for
// step2: Dual-branch Transformer Encoder
$X^0_{fus} = 0$
for l in $(1 \sim N)$ **do**
  $X^l_{dif|t_0}, X^l_{dif|t_1} = DE(X^l_{t_1}, X^l_{t_0})$
  $X^l_{t_0} = CE(X^l_{t_0}, X^l_{dif|t_0})$
  $X^l_{t_1} = CE(X^l_{t_1}, X^l_{dif|t_1})$
  $X^l_{fus} = MBF(X^l_{t_0}, X^l_{t_1}, X^{l-1}_{fus})$
end for
// step3: Caption Decoder
$Caption = EM(\text{"}start\text{"})$
while $w \neq EM(\text{"}end\text{"})$ **do**
  $w = CD(X^N_{fus}, Caption)$
  $Caption = [Caption; w]$
end while
return $Caption$

X_i $(i = t_0, t_1)$ to capture and recognize multiple changes in the image at the time i $(i = t_0, t_1)$. We have explored three difference encoding (DE) strategies to generate $X_{dif|i}$ $(i = t_0, t_1)$. These strategies can be formulated as follows:

$$DE_1 : X_{dif|t_0} = X_{dif|t_1} = X_{t_1} - X_{t_0} \tag{1}$$

$$DE_2 : X_{dif|t_0} = X_{dif|t_1} = Abs(X_{t_1} - X_{t_0}) \tag{2}$$

$$DE_3 : X_{dif|i} = RELU(conv([X_{t_1} - X_{t_0}; X_i])) \tag{3}$$

where Abs denotes the element-wise operation of absolute value. $conv$ denotes 2D convolution for feature transformation. $RELU$ is a activation function.

B. Siamese Cross-encoding Modules

Our intuition is that the siamese cross-encoding (CE) modules utilize the difference representation from the difference encoding module to capture changes. Fig. 6 illustrates the structure of the CE module. Unlike the Transformer encoding layer [86], we replace the self-attention mechanism with a cross-attention mechanism. Each CE module contains a residual unit with the multi-head cross-attention mechanism and a residual unit with the feed-forward network. Layer normalization (LN) [93] is performed after each residual unit.

For the cross-attention mechanism of the CE module, the query (Q) is the linear transformation of single temporal image features, while key (K) and value (V) are from the linear transformation of the $X_{dif|i}$ $(i = t_0, t_1)$ output by the difference encoding module. Assume that $X^l_i \in R^{HW \times C}$ and $X_{dif|i} \in R^{HW \times C} (i = t_0, t_1)$ are input to the CE module of the $l - th$ DTE stage. H, W, and C, respectively, are the height, width, and channel dimension of the feature map. The cross-

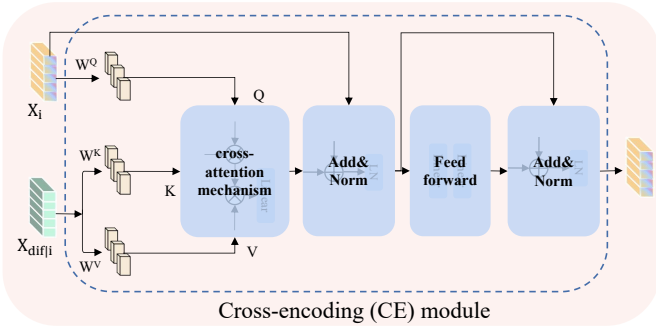


Fig. 6. The structure of the cross-encoding module. The module contains two residual units, and layer normalization (LN) [93] is performed after each residual unit.

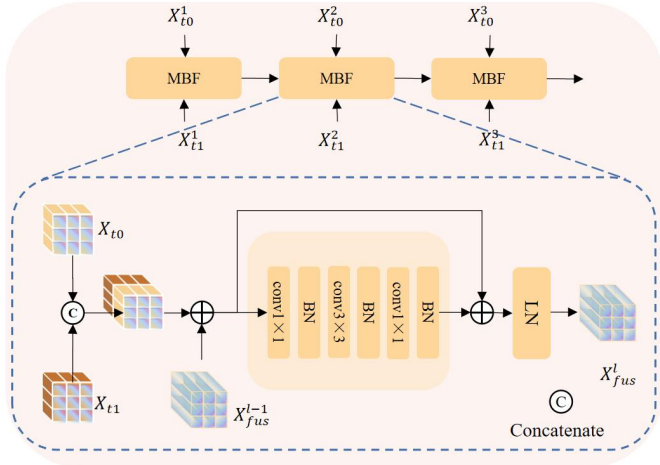


Fig. 7. The structure of the multi-stage bi-temporal fusion module (MBF). MBF utilizes the bi-temporal features from the siamese CE modules and features from the previous MBF module to obtain high-level semantic features revealing multiple changes of interest.

attention mechanism of the CE module at the $l-th$ DTE stage can be expressed as follows:

$$Q_i = X_i^l W^Q \tag{4}$$
$$K_i = X_{dif|i}^l W^K \tag{5}$$
$$V_i = X_{dif|i}^l W^V \tag{6}$$
$$\tag{7}$$

where $W^Q, W^K, W^V \in R^{C\times d}$ are trainable weight matrices, C is the dimension of a single image feature X_i, d is the dimension of the Q_i, K_i and V_i. Single-head cross-attention can be formulated as follows:

$$CrossAtt(Q_i, K_i, V_i) = \sigma(\frac{Q_i K_i^T}{\sqrt{d}})V_i \tag{8}$$

where σ is the softmax function.

A single attention head can project input features into one representation subspace. Multiple attention heads can obtain multiple representation subspaces because of multiple (W^Q, W^K, W^V). It has been proved that the multi-head cross-attention (MCA) mechanism pays attention to input features from different aspects [86]. MCA contains multiple parallel single-head cross-attention. The resulting features from multiple heads are concatenated and then sent into a linear layer to make the output features have the same dimension as the input features. The procedure of MCA at the $l-th$ DTE stage can be formulated as follows:

$$MultiHead(X_i^l, X_{dif|i}^l) = Concat(head_1, ...head_h)W^o \tag{9}$$
$$where \quad head_j = CrossAtt(X_i^l W_j^Q, X_{dif|i}^l W_j^K, X_{dif|i}^l W_j^V) \tag{10}$$

where $W_j^Q, W_j^K, W_j^V \in R^{C\times d}$ are trainable matrices of the $j-th$ head, $W^o \in R^{hd\times C}$ is a trainable linear projection matrix to transform the feature channel dimension, h is the head number of the multi-head cross-attention.

The feed-forward neural network of the CE module consists of two linear layers with a RELU activation function. It can be formulated as follows:

$$FF(X) = RELU(XW_1)W_2 \tag{11}$$

where $X \in R^{HW\times d}$ is the input. $W_1 \in R^{d\times 4d}$ and $W_2 \in R^{4d\times d}$ are trainable matrices.

The procedure of the CE module at the $l-th$ DTE stage can be expressed as follows:

$$CE(X_i^l) = LN(X + FF(X)) \tag{12}$$
$$X = LN(X_i^l + MultiHead(X_i^l, X_{dif|i}^l)) \tag{13}$$

where LN denotes the layer normalization [93].

C. Multi-stage Bi-temporal Fusion Module

The DTE consists of a hierarchy of processing stages. The siamese CE modules at different stages can focus on different changes. To focus on multiple changes of interest and ignore irrelevant changes in the bi-temporal RS images, we designed a multi-stage bi-temporal fusion (MBF) module to utilize the features from different stage siamese CE modules to obtain better high-level semantic feature representations revealing multiple changes of interest.

Fig. 7 illustrates the structure of the MBF module. In the MBF module, we concatenate the bi-temporal features (X_{t_0}, X_{t_1}) from the siamese CE modules together in the channel dimension. We then add them together with the features from the previous MBF module. The resulting features are fed into a residual unit with three convolutional layers to perform bi-temporal feature fusion. The layer normalization is then performed, which is proved to be valuable in experiments. The procedure of MBF at the $l-th$ DTE stage can be formulated as follows:

$$X^l = [X_{t_0}^l; X_{t_1}^l] \tag{14}$$
$$X_{fus}^l = X^l + X_{fus}^{l-1} \tag{15}$$
$$X_{res}^l = conv_3(RELU(conv_2(RELU(conv_1(X_{fus}^l))))) \tag{16}$$
$$X_{fus}^l = LN(X_{fus}^l + X_{res}^l) \tag{17}$$

where $X_i^l \in R^{H\times W\times d}(i = t_0, t_1)$ are the reshaped features from the siamese CE modules at the $l-th$ DTE stage. [;] denotes

the concatenation operation. $conv_1, conv_2$, and $conv_3$ are 2D convolutions with a 1×1, 3×3, and 1×1 kernel. *LN* represents the layer normalization.

D. Caption Decoder

The caption decoder utilizes X_{fus} from the MBF module at the last stage of the DTE to generate the sentences describing the differences between images. For the caption decoder, we use a standard Transformer decoder [86], which is a stack of multiple decoding layers. Each decoding layer contains three sub-layers, including multi-head masked self-attention, encoder-decoder attention, and feed-forward network. Besides, each sub-layer contains a residual connection and is followed by the layer normalization operation.

Unlike multi-head self-attention, the multi-head masked self-attention masks out the values at position t and subsequent positions when predicting the $t-th$ word in the training stage. It ensures that each word prediction only depends on the generated words [86]. For the encoder-decoder attention of the caption decoder, it is similar to the multi-head cross-attention of the siamese CE modules in the DTE. However, the Q is from the previous masked self-attention sub-layer, and the K and V are from the transformation of X_{fus}. The encoder-decoder attention ($EDAtt$) of the $l-th$ Transformer decoding layer can be formulated as follows:

$$EDAtt(S^l, X_{fus}) = Concat(head_1, ...head_h)W^o \quad (18)$$

$$where \quad head_j = CrossAtt(S^l W_j^Q, X_{fus} W_j^K, X_{fus} W_j^V) \quad (19)$$

where $W_j^Q, W_j^K, W_j^V \in R^{d\times d}$ are trainable matrices of the $j-th$ head, $W^o \in R^{hd\times d}$ is a trainable matrix to perform linear projection changing the feature channel dimension, $S_l^i \in R^{L\times d}$ is the sentence embedding from the masked self-attention sub-layer, L is the length of the sentence, h is the head number of the multi-head cross-attention.

The output of the Transformer decoder is sent to a linear layer with a softmax activation function to generate word probabilities $P = [\boldsymbol{p_1}, \boldsymbol{p_2}, ...\boldsymbol{p_L}] \in R^{L\times K}$. K is the vocabulary size. The word probability vector $\boldsymbol{p_t} = [\ p_t^{(1)}, p_t^{(2)}, ..., p_t^{(K)}]$ is used to predict the word at position t in the generated sentence. It can be formulated as follows:

$$\boldsymbol{p_t} = Softmax(\boldsymbol{y_t}) = \frac{exp(\boldsymbol{y_t})}{\sum_{i=1}^{K} exp(y_t^{(i)})} \quad (20)$$

where $\boldsymbol{y}_t = [\ y_t^{(1)}, y_t^{(2)}, ...y_t^{(K)}]$ is a vector from the linear layer. $Softmax$ is a softmax activation function.

E. Network Details

1) RSICCformer-based model: We use a modified ResNet-101 [94] pretrained on the ImageNet dataset [95] as backbone to extract image features from bi-temporal images. We replace the conv5_x and subsequent layers of ResNet-101 with a 1×1 convolution to reduce the feature channel dimension from 1024 to 512. We use three processing stages in the dual-branch Transformer encoder and one decoding layer in the caption decoder. The head number of all multi-head attention is set to 8. We train word embeddings from scratch and set the dimension of the word embedding to 1024. Besides, we used the teacher forcing strategy in the training stage.

2) Position Embedding: Two position embedding methods are used in the dual-branch Transformer encoder and the caption decoder. We use the learnable 2-D position embedding [96] to incorporate spatial position information in our mode before bi-temporal feature sequences are fed to the dual-branch Transformer encoder. Besides, the word order of the sentence is valuable for captioning. In the caption decoder, we use sine and cosine functions to perform the position embedding for the sequence, as introduced in [86].

3) Loss function: The simple and effective CrossEntropy loss has been widely used in captioning-related tasks [68], [70], [71], [74]. We use it as the loss function and minimize it to optimize the model in the training stage. The loss function can be formulated as follows:

$$loss = -\sum_{t=1}^{L} log(\sum_{k=1}^{K} \widetilde{y}_t^{(k)} p_t^{(k)}) \quad (21)$$

where L is the sentence length, K is the vocabulary size. $\widetilde{\boldsymbol{y}}_l = [\ \widetilde{y}_t^{(1)}, \widetilde{y}_t^{(2)}, ...\widetilde{y}_t^{(K)}]$ is the vector representation of the word at position t in the reference sentence. $\boldsymbol{p}_t = [\ p_t^{(1)}, p_t^{(2)}, ..., p_t^{(K)}]$ is the word probability vector which can be used to predict the word at position t in the generated sentence.

V. EXPERIMENTS

A. Experimental Setup

1) Implementation details: We implemented all models based on the PyTorch deep learning framework. All models are trained and evaluated on the NVIDIA GTX 1080Ti GPU. In the training stage, we use the Adam optimizer [97]. We set the maximum epoch to 40. The dimension of the word embedding to 512. The initial learning rate is 0.0001. We evaluate the model on the validation set after each training epoch. The learning rate will decay by a weight of 0.7 when the BLEU-4 metric on the validation set decreases three epochs. The best model on the validation set is selected for evaluation on the test set.

2) Evaluation metrics: The performance evaluation of the captioning model depends on whether the generated descriptive sentences conform to human judgments about differences between bi-temporal images. The automatic evaluation metrics can automatically measure the accuracy of the generated sentences based on the annotated reference sentences. In this work, we use four different metrics to evaluate the captioning accuracy, including BLEU-N (N=1,2,3,4) [90], ROUGE-L [91], METEOR [92], and CIDEr-D [98], which are widely used in the remote sensing image captioning tasks [68], [70], [71], [74] and change captioning tasks [12]–[17], [99]. Vedantam *et al.* [98] have demonstrated that these evaluation metrics can highly correlate with human judgment. The higher the metric scores, the higher the similarity between generated sentences and reference sentences, and therefore the higher the captioning accuracy.

B. Effects of Different Difference Encoding Strategies

The difference encoding module utilizes bi-temporal features to obtain the feature representation that reveals the differences between bi-temporal images. Siamese CE modules then

TABLE VI
THE EFFECTS OF THREE DIFFERENCE ENCODING STRATEGIES, WHICH HAVE BEEN INTRODUCED IN SECTION IV-A. WE CAN SEE THAT THE THREE STRATEGIES ARE EFFECTIVE.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Baseline	78.04	69.17	61.95	56.35	35.92	68.77	118.25
DE_1	83.09	74.32	66.66	60.44	**38.76**	72.63	130.00
DE_2	79.48	69.65	61.10	53.86	36.45	71.47	124.59
DE_3	**83.19**	**74.53**	**67.37**	**61.75**	38.51	**72.67**	**131.24**

TABLE VII
ABLATION STUDIES ON THE SIAMESE CE MODULES AND MBF MODULE. THE MODEL PERFORMANCE IS EVALUATED THE MODEL PERFORMANCE ON THE TEST SET FROM THREE ASPECTS. WE DID NOT REPORT THE CIDER-D METRIC FOR THE FIRST TEST METHOD AS THIS METRIC WOULD PENALIZE THE SCORES FOR FREQUENTLY OCCURRING WORDS. SINCE THE WORDS EXPRESSING NO CHANGE ARE RELATIVELY MONOTONOUS, CIDER-D WILL APPROACH 0. THEREFORE, CIDER-D CANNOT MEASURE SENTENCE ACCURACY IN THIS CASE.

Test Range	Method	CE	MBF	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Test set (only no-change)	Baseline	×	×	91.07	89.77	89.12	88.69	66.72	91.88	-
	RSICCformer*	✓	×	91.02	89.62	88.81	88.23	66.67	92.24	-
	RSICCformer	✓	✓	**95.05**	**94.24**	**93.76**	**93.42**	**72.20**	**95.68**	-
Test set (only change)	Baseline	×	×	67.31	52.21	39.12	29.01	21.85	45.62	40.92
	RSICCformer*	✓	×	75.94	61.25	47.85	37.08	**25.88**	**52.75**	**60.59**
	RSICCformer	✓	✓	**76.43**	**61.92**	**48.81**	**38.14**	25.72	52.53	60.56
Test set (entire set)	Baseline	×	×	78.04	69.17	61.95	56.35	35.92	68.77	118.25
	RSICCformer*	✓	×	83.09	74.32	66.66	60.44	38.76	72.63	130.00
	RSICCformer	✓	✓	**84.72**	**76.27**	**68.87**	**62.77**	**39.61**	**74.12**	**134.12**

utilize the feature difference to recognize multiple changes. Tab. VI reports the effects of the three difference encoding strategies, which have been introduced in Section IV-A. The baseline does not use the difference information of bi-temporal images. It employs siamese Transformer encoders without the difference encoding module to process bi-temporal features separately. We can see that the three difference encoding strategies are effective for improving model performance.

For the second strategy, the operation of taking the absolute value will lead to performance degradation, which may be because the timing information is valuable for the RSICC task when describing the change types, such as appearing and disappearing. Besides, the first strategy has low computational complexity and is comparable to the third strategy. Therefore, we employed the first straightforward and effective strategy in subsequent experiments.

C. Ablation Studies

We conducted ablation studies to validate the effectiveness of the siamese CE modules and MBF module. We trained the model on the same training set and then evaluated the model performance on the test set from three aspects: 1) only testing the image pairs with changes, 2) only testing the image pairs without changes, 3) testing on the entire test set. Note that we did not report the CIDEr-D metric for the first test method as this metric would penalize the scores for frequently occurring words. Since the words used to express no change are relatively monotonous, CIDEr-D will approach 0. Therefore, CIDEr-D cannot measure sentence accuracy in this case.

Tab. VII reports the quantitative ablation results. Baseline and RSICCformer* have the same feature extraction module and caption decoder as RSICCformer. However, their encoders are different from RSICCformer. The bi-temporal features from the dual-branch encoder are concatenated on the channel dimension and then sent to the caption decoder.

1) Siamese cross-encoding module: Tab. VII shows that RSICCformer* outperforms baseline in three test ways. This demonstrates that the siamese CE modules are effective and can significantly improve the change captioning performance of the model. It is attributed to the difference representation being utilized by the siamese CE module. Since bi-temporal images do not have the viewpoint change, the approach of exploiting the difference between the corresponding positions of two images as a prior is valuable for improving the feature representation ability of the model and can be viewed as introducing an inductive bias to the model.

Many previous synthetic image change captioning methods have attempted to correlate identical regions of two images with viewpoint change in order to capture and recognize changes [12], [16], [17], [99]. Although our dataset has no viewpoint change, it is still valuable to pay attention to the corresponding regions before and after the change in the two images. That helps the model recognize the change types and the ground appearance before and after the change. We visualize the attention states of each stage in the DTE of RSICCformer and Transformer encoder of baseline, as shown in Fig. 8. For the baseline, the attention regions for the two images are different at each stage. It is attributed to the fact that two images are processed separately by two Transformer

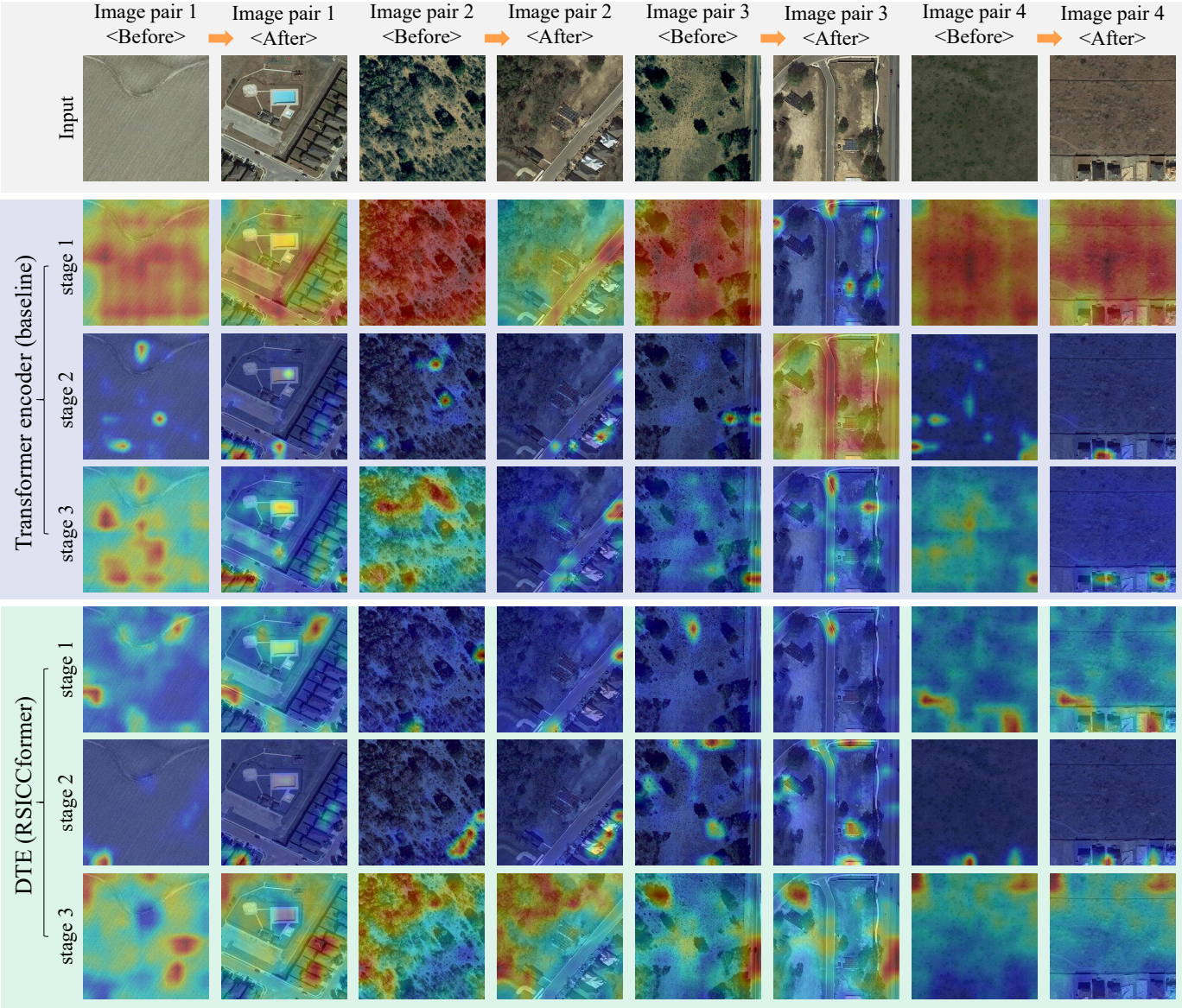


Fig. 8. Visualization of the attention states of each stage in the Transformer encoder of baseline and DTE of RSICCformer. Specifically, for the baseline, the attention maps are obtained from the self-attention of the Transformer encoding layer. For the DTE, the attention maps are obtained from the cross-attention of the CE module. The warmer the color, the higher the attention value. Red areas are paid more attention, while blue areas are paid less attention. We can observe that our method could focus on multiple changes at different stages. Regarding image pair 1, the first stage focuses on the changed winding path, swimming pool, and road. The second stage focuses on the small building change area, and the third stage focuses on the large building change area. Regarding image pair 2, the first stage focuses on road change, the second stage focuses on building change, and the third stage focuses on vegetation change. Regarding image pair 3, the first stage focuses on the road change, the second stage focuses on building and road changes, and the third stage focuses on vegetation change.

encoders, which make their respective salient regions to be paid attention to. For our method, the attention maps for the bi-temporal images seem to be the same. It illustrates that the siamese CE modules can pay attention to the positions of changed regions in the bi-temporal images at each stage. That is attributed to the ability of the siamese CE modules to utilize the different features revealing change positions. Besides, we can observe that the siamese CE modules at different stages can focus on multiple different changed objects, some large and some small. For example, regarding image pair 2, the first stage focuses on the road change, the second stage on the building change, and the third stage on the vegetation change.

2) Multi-stage bi-temporal fusion module: Fig. 8, shows that the siamese CE modules can focus on different changes at different stages. Besides, bi-temporal images contain some irrelevant changes that we are not interested in. For example, regarding image pair 2 in Fig. 8, we are interested in road change and building change, not vegetation change and light change. To capture multiple changes of interest and ignore irrelevant changes in the bi-temporal RS images, we designed the MBF to utilize the bi-temporal features from different stage CE modules to obtain high-level semantic feature representation, which reveals multiple changes of interest and excludes irrelevant changes. Tab. VII demonstrates that MBF can signif-

TABLE VIII
THE CHANGE DISCRIMINATION ABILITY OF THE MODELS. WE CLASSIFIED THE GENERATED SENTENCES INTO TWO CATEGORIES: SENTENCES INDICATING CHANGE AND THOSE INDICATING NO CHANGE. WE THEN CALCULATED THE DISCRIMINATION ACCURACY. OUR METHOD EFFECTIVELY IMPROVES THE CHANGE DISCRIMINATION ABILITY.

Method	Change accuracy	No-change accuracy	Total accuracy
Baseline	83.52%	89.61%	86.57%
RSICCformer*	**94.02**%	88.97%	91.50%
RSICCformer	90.91%	**94.48**%	**92.70**%

icantly improve the model performance, which is attributed to the ability of MBF to incorporate multiple changes of interest and exclude irrelevant changes.

The ability of the model to discriminate whether changes of interest have occurred is significant. To validate the effectiveness of the siamese CE modules and MBF module in improving the change discrimination ability of the model, we conducted experiments to evaluate the ability qualitatively. We classified the generated sentences into two categories: sentences indicating change and those indicating no change. We then calculated the accuracy of change discrimination. Tab. VIII demonstrates that our method can effectively improve the change discrimination ability. Specifically, RSICCformer performs better than RSICCformer* on the no-change accuracy, which may be attributed to the ability of MBF to exclude the interference of irrelevant changes. Although the change accuracy of RSICCformer is lower than that of RSICCformer*, Tab. VII shows that RSICCformer performs better than RSICCformer* in generating sentences describing these changes, which is attributed to the ability of MBF to obtain better feature representations revealing multiple changes of interest. In conclusion, Tab. VIII and Tab. VII demonstrate that our method effectively improves the change discrimination accuracy and sentence accuracy.

D. Comparison to state-of-the-art

We benchmark several state-of-the-art synthetic image change captioning methods on our LEVIR-CC dataset, including four LSTM-based methods (Capt-Rep-Diff [12], Capt-Att [12], Capt-Dual-Att [12], DUDA [12]) and two Transformer-based methods (MCCFormers-S [17], MCCFormers-D [17]). These methods are introduced as follows:

- Capt-Rep-Diff [12]: ResNet-101 extracts the "before" feature map and "after" feature map from the bi-temporal images. The element-wise subtraction of two feature maps is defined as the "difference" feature map. The three feature maps are concatenated on the channel dimension and then sent to an LSTM for captioning.
- Capt-Att [12]: Capt-Att introduces the spatial attention mechanism. The feature extraction of Capt-Att is the same as that of Capt-Rep-Diff. Two convolutional layers utilize the "difference" feature map to generate a single spatial attention weight map to attend both "before" and "after" feature maps. The resulting attended feature maps are then subtracted and input into an LSTM for captioning.
- Capt-Dual-Att [12]: There is no viewpoint change in the synthetic image CC dataset CLEVR-Change. Single attention may not establish the relationship between the objects in two images. Unlike Capt-Att, Capt-Dual-Att utilizes two convolutional layers to generate two spatial attention weight maps to attend "before" and "after" feature maps.
- DUDA [12]: Different from Capt-Dual-Att, DUDA employs a dynamic speaker to adaptively focus on "before", "after" or "difference" visual representations. The dynamic speaker consists of two LSTMs. An LSTM utilizes "before", "after", "difference" feature representations, and already generated words to generate three attention weights for three representations. The three feature representations are input into the LSTM to generate sentences after weighted summation.
- MCCFormers-S [17]: ResNet-101 extracts the "before" feature map and "after" feature map from two images. The two feature maps are flattened into two feature sequences. They are concatenated together after the linear transformation and position encoding. The resulting features are sent into a Transformer encoder and a Transformer decoder for captioning.
- MCCFormers-D [17]: The feature extraction is the same as MCCFormers-S. The extracted features are then sent into the MCCFormers-D, a siamese Transformer-based network. MCCFormers-D employs the co-attention mechanism [24] to capture relationships between two images and localize changed regions. Finally, a Transformer decoder utilizes the features from MCCFormers-D to generate sentences describing the differences between two images.

Tab. IX shows the comparison results of our RSICCformer and existing state-of-the-art synthetic image change captioning methods. We can see that our method outperforms the other methods on all metrics by a significant margin. For example, our RSICCformer exceeds the recent MCCFormers-D by 6.39% on BLEU-4 and 9.86% on CIDEr-D. It may attribute to the ability of RSICCformer to capture relationships between two images and localize change regions, which is valuable to recognize multiple changes of interest and ignore irrelevant changes.

Besides, Qiu *et al.* [17] demonstrate that MCCCFormer-D outperforms DUDA [12] on the CLEVER dataset due to the ability of the model to locate multiple changes under viewpoint change. However, DUDA outperforms MCCCFormer-D on our LEVIR-CC dataset without viewpoint change. It may be because DUDA utilizes the bi-temporal feature difference of corresponding positions. Since bi-temporal RS images of our dataset are well registered in a pixel-by-pixel manner, the comparison results suggest that exploiting the difference between the corresponding positions of two images as a prior is valuable for improving the feature representation and discriminative ability of the model.

The Spot-the-Diff [7] dataset is also a real-world CC dataset, which describes the changes in surveillance scenarios, as introduced in Section III-C4. Tab. X shows the comparison results of our method and other methods on the Spot-the-Diff dataset. Though our method is proposed for remote sensing change captioning, it still obtains scores comparable to the

TABLE IX
COMPARISONS WITH OTHER EXISTING STATE-OF-THE-ART SYNTHETIC IMAGE CHANGE CAPTIONING METHODS ON THE LEVIR-CC DATASET. THE HIGHER THE SCORE, THE HIGHER THE ACCURACY OF THE SENTENCE OUTPUT BY THE MODEL. OUR RSICCFORMER OUTPERFORMS THE PREVIOUS THOSE METHODS ON ALL METRICS BY A SIGNIFICANT MARGIN.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Capt-Rep-Diff [12]	72.90	61.98	53.62	47.41	34.47	65.64	110.57
Capt-Att [12]	77.64	67.40	59.24	53.15	36.58	69.73	121.22
Capt-Dual-Att [12]	79.51	70.57	63.23	57.46	36.56	70.69	124.42
DUDA [12]	81.44	72.22	64.24	57.79	37.15	71.04	124.32
MCCFormers-S [17]	79.90	70.26	62.68	56.68	36.17	69.46	120.39
MCCFormers-D [17]	80.42	70.87	62.86	56.38	37.29	70.32	124.44
Baseline	78.04	69.17	61.95	56.35	35.92	68.77	118.25
Ours	**84.72**	**76.27**	**68.87**	**62.77**	**39.61**	**74.12**	**134.12**

TABLE X
COMPARISONS WITH OTHER EXISTING STATE-OF-THE-ART METHODS ON THE SPOT-THE-DIFF DATASET.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Capt-Rep-Diff [12]	69.68	14.59	27.20	29.63
Capt-Att [12]	54.78	13.28	25.74	23.43
Capt-Dual-Att [12]	71.18	14.45	28.12	31.38
DUDA [12]	74.71	14.23	27.38	30.86
MCCFormers-S [17]	63.37	13.38	26.33	25.06
MCCFormers-D [17]	71.84	**15.31**	**28.92**	31.26
Ours	**75.78**	14.95	28.02	**35.36**

state-of-the-art methods in this dataset.

E. Qualitative Results

Fig. 9 shows the visualization of corresponding attention states of the caption decoder when our model generates an object-related word. We mark that word in green in the generated sentence. We superimpose the corresponding attention states on the "after" image to observe the attended image region better. We observe that our model can focus well on corresponding image change regions when generating object-related words.

Fig. 10 shows some captioning results on the LEVIR-CC dataset. For each image pair, we provide one of the five ground-truth sentences, the sentences generated by the baseline and our RSICCformer. More accurate and detailed words are marked in red. Green words are not accurate. We can observe that our RSICCformer can generate more accurate and detailed descriptions than baseline. For example, for the first image pair, our model can not only describe objects such as "houses" and "roads" but also scene concepts such as "A residential area". For the second image pair, our model can focus on the changed road, whereas the baseline cannot. That is because RSICCformer can significantly exploit the difference information to recognize multiple changes of interest accurately.

F. Effects on different change categories

Tab. XI reports the performance of RSICCformer on each changed object category in our LEVIR-CC dataset. Our model performs well in describing building-related changes but poorly in describing water-related changes and parking lot changes. The relatively poor performance on specific changes is most likely due to the lacking of training data.

G. Adaptation to Different Backbones

We benchmark our RSICCformer with different backbones on the LEVIR-CC dataset. We employ five classical backbones to extract bi-temporal features, including three CNN-based backbones (VGG-16 [100], ResNet-50, ResNet-101 [94]) and two Transformer-based backbones (ViT-B/16, ViT-L/16 [96]). These backbones have been pretrained on the ImageNet dataset [95]. Tab. XII reports captioning metrics, training time per epoch measured in minutes (min) and inference speed measured in frames per second (fps). We can observe that our method has good adaptability to different backbones. Besides, the model performs better when the model employs a deeper CNN-based backbone or a deeper Transformer-based backbone. Although ViT-L/16 performs better than other backbones, it will lead to longer training time and slower inference speed. To balance the performance and speed, our model uses the ResNet-101 as the backbone to extract image features from bi-temporal images in our RSICCformer model.

H. Parameter Analysis

1) Depth of the network.: The dual-branch Transformer encoder and the caption decoder are stacked by multiple stages, and the depth of the network is an important hyperparameter. We compare the performance of models of different depths, as shown in Tab. XIII. E.D. denotes the depth of the dual-branch Transformer encoder, and D.D. denotes the depth of the caption decoder. We observe that the model performs best when E.D. and D.D. are 3. Fewer layers will reduce the ability of the model to learn to align visual changes of interest and language. More layers reduce performance, which may be because the increased complexity makes it difficult for the network to learn the optimal parameters.

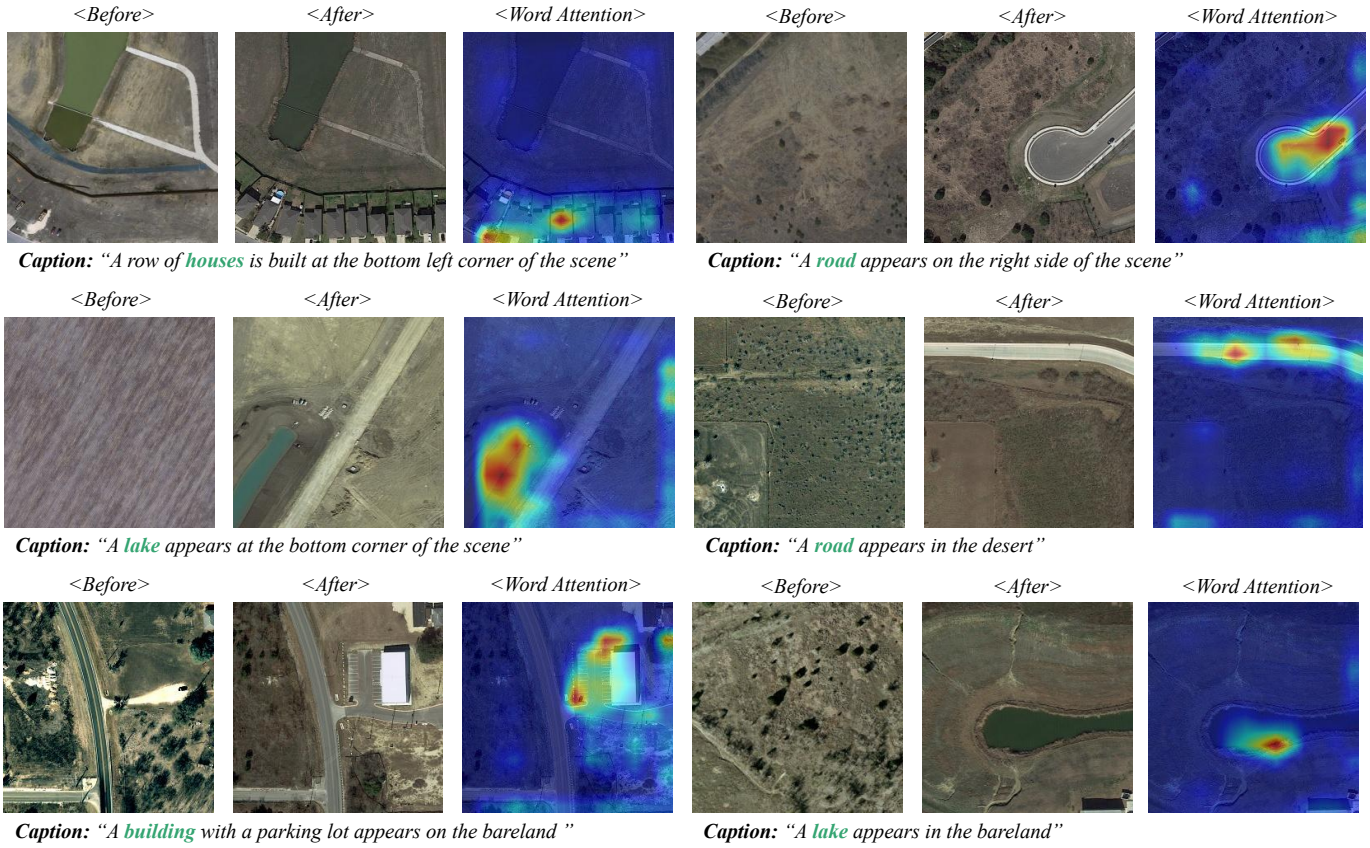


Fig. 9. Visualization of corresponding attention states of the caption decoder when our model generates an object-related word. We mark that word in green in the generated sentence. We superimpose the corresponding attention states on the "after" image to observe the attended image regions better. We observe that our model can focus well on corresponding image change regions when generating object-related words.

TABLE XI
THE PERFORMANCE OF RSICCFORMER ON EACH CHANGED OBJECT CATEGORY IN OUR LEVIR-CC DATASET.

Changed object category	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Building	77.48	63.06	49.94	39.24	26.17	53.32	61.83
Road	76.97	61.36	47.23	36.17	25.90	52.03	54.95
Parking lot	74.68	58.17	43.94	31.94	23.44	47.88	43.82
Vegetation	75.91	59.51	45.19	34.49	24.50	49.33	50.13
Water	69.39	54.78	42.04	31.97	24.20	48.53	46.86

TABLE XII
THE PERFORMANCE OF RSICCFORMER-BASED MODEL EMPLOYING DIFFERENT BACKBONES, INCLUDING THREE CNN-BASED BACKBONES (VGG-16 [100], RESNET-50, RESNET-101 [94]) AND TWO TRANSFORMER-BASED BACKBONES (VIT-B/16, VIT-L/16 [96]). THE TRAINING TIME PER EPOCH IS MEASURED IN MINUTES (MIN) AND THE INFERENCE SPEED IS MEASURED IN FRAMES PER SECOND (FPS).

Backbone	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	Training Time	Inference Speed
VGG-16	82.57	73.91	66.67	60.76	37.83	71.66	127.35	11 min	16.0 fps
ResNet-50	84.50	75.92	68.53	62.39	39.57	**74.18**	**134.68**	10 min	15.2 fps
ResNet-101	**84.72**	**76.27**	**68.87**	**62.77**	**39.61**	74.12	134.12	11 min	12.4 fps
ViT-B/16	84.27	76.06	68.93	63.05	39.25	73.80	133.47	13 min	15.3 fps
ViT-L/16	**84.81**	**76.39**	**69.14**	**63.07**	**39.30**	**74.01**	**134.01**	27 min	10.1 fps

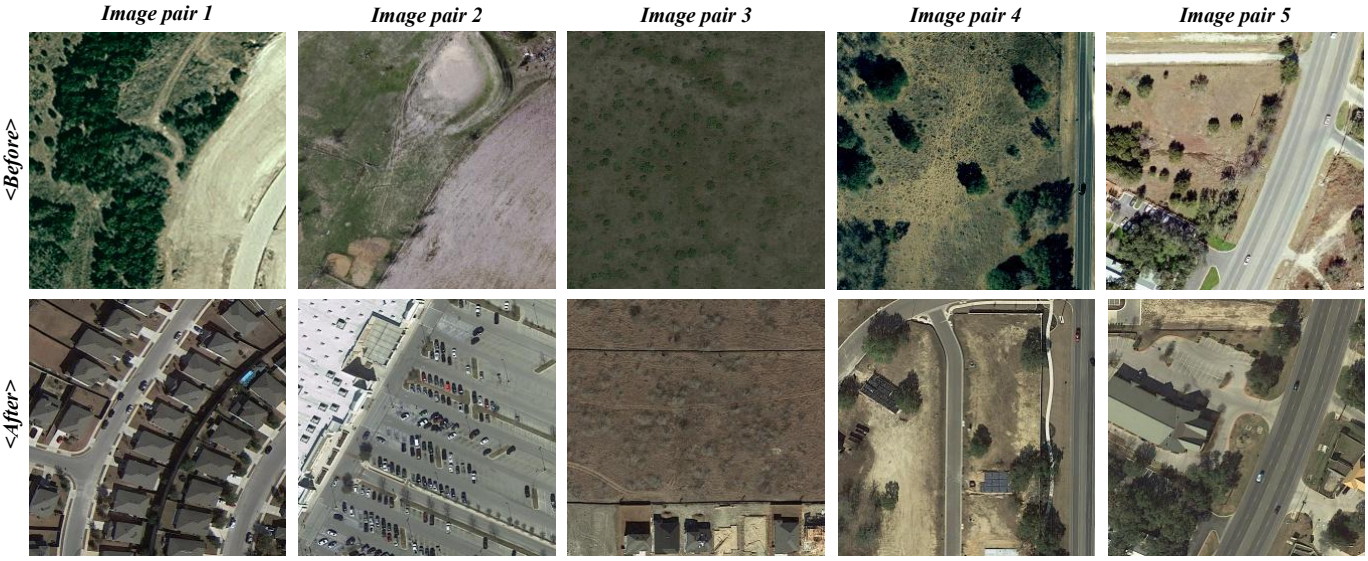


Fig. 10. Captioning results on the LEVIR-CC dataset. Sentence (a) is one of the five ground-truth sentences. The baseline and our RSICCformer generate sentences (b) and (c). More accurate and detailed words are marked in green. Red words are not accurate. We can observe that our RSICCformer can generate more accurate and detailed descriptions than baseline. For example, for the first image pair, our model can not only describe objects such as "houses" and "roads" but also scene concepts such as "A residential area".

TABLE XIII
THE PERFORMANCE OF THE MODEL IN DIFFERENT DEPTHS ON THE LEVIR-CC DATASET. E.D. DENOTES TO THE DEPTH OF THE DUAL-BRANCH TRANSFORMER ENCODER, AND D.D. DENOTES TO THE DEPTH OF THE CAPTION DECODER.

E.D.	D.D.	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
1	1	83.55	75.31	68.23	62.36	38.45	72.89	129.78
2	1	84.11	75.71	68.41	62.38	38.82	73.24	131.35
3	1	**84.72**	**76.27**	**68.87**	**62.77**	**39.61**	**74.12**	**134.12**
4	1	83.65	75.09	67.75	61.72	38.99	73.11	131.00
1	2	83.17	74.69	67.62	61.94	38.14	72.64	129.40
1	3	83.05	74.27	67.00	61.21	38.29	72.64	129.88
1	4	81.85	73.38	66.46	61.10	37.94	72.41	129.04
2	2	83.40	74.67	67.26	61.31	38.40	72.76	129.93
3	3	83.79	74.85	67.38	61.33	38.78	73.36	132.10
4	4	82.59	73.75	66.31	60.31	38.88	73.10	130.95
2	3	82.51	73.51	66.05	60.16	38.18	72.41	128.80
4	3	83.33	74.41	66.87	60.74	38.83	73.33	131.38
3	2	83.69	74.74	67.41	61.47	39.09	73.45	132.38
3	4	83.39	74.13	66.53	60.43	38.85	72.99	131.27

2) Position embedding: For our model in Fig. 5, we have reshaped the features in three positions: 1) The feature map from the CNN backbone is rearranged into the feature sequence before being input to the DTE: $(H, W, d) \rightarrow (H \times W, d)$. 2) The feature sequence X_i $(i = t_0, t_1)$ from the CE module is rearranged into the feature map before being input into MBF: $(H \times W, d) \rightarrow (H, W, d)$. 3) The feature map X_{fus} from MBF is rearranged into the feature sequence before being input to the Transformer decoder: $(H, W, d) \rightarrow (H \times W, d)$. We have tried to add the learnable position embeddings [96] in those three positions, as shown in Tab. XIV. We can observe that the position embedding is effective. Besides, the model performs best when we add position embedding at position 1.

3) The Beam Search Strategy for Inference: The beam search strategy [101] has been widely used to improve model performance in machine translation [102]–[104] and captioning tasks [70], [74]. The beam size will affect the sentence accuracy and inference speed in the inference stage. When the beam size is 1, the strategy is a greedy search strategy. We test our RSICCformer with different beam sizes on the LEVIR-CC dataset. Tab. XV demonstrates that the beam search strategy effectively improves sentence inference accuracy. Besides, we can observe that increasing beam size will lead to a decrease in inference speed due to a large amount of computation and the evaluation metrics tend to saturate when beam size increases to 3. Model inference effect when choosing beams of different sizes.

I. Complexity Analysis

Tab. XVI reports the number of model parameters and floating-point operations (FLOPs). The siamese CE modules significantly improve model performance without increasing parameters and computational complexity. The MBF module

TABLE XIV
THE PERFORMANCE OF OUR MODEL WITH DIFFERENT POSITION EMBEDDINGS ON THE LEVIR-CC DATASET. WE CAN OBSERVE THAT THE POSITION EMBEDDING IS EFFECTIVE, AND THE MODEL PERFORMS BEST WHEN WE ONLY ADD POSITION EMBEDDING AT POSITION 1.

Position 1	Position 2	Position 3	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
×	×	×	83.59	74.92	67.27	60.92	38.85	72.80	130.23
✓	×	×	**84.72**	**76.27**	**68.87**	**62.77**	**39.61**	**74.12**	**134.12**
×	✓	×	84.17	75.53	68.42	62.35	38.91	72.92	130.66
×	×	✓	84.08	75.53	68.20	62.15	39.07	73.28	131.55
✓	✓	×	83.40	74.61	67.18	61.11	39.25	73.40	131.17
✓	×	✓	83.19	74.47	67.23	61.34	38.74	72.84	130.28
✓	✓	✓	83.04	74.17	66.75	60.78	38.89	72.40	129.59

TABLE XV
MODEL INFERENCE EFFECT WHEN CHOOSING DIFFERENT BEAM SIZES ON THE LEVIR-CC DATASET. THE BEAM SEARCH STRATEGY EFFECTIVELY IMPROVES SENTENCE INFERENCE ACCURACY. BESIDES, THE EVALUATION METRICS TEND TO SATURATE WHEN BEAM SIZE INCREASES TO 3.

Beam size	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	Inference Speed
1	84.72	76.27	68.87	62.77	39.61	74.12	134.12	12.4 fps
2	85.07	77.27	70.39	64.74	39.94	74.52	135.36	11.3 fps
3	**85.24**	**77.38**	70.68	65.24	40.06	74.61	**135.57**	10.9 fps
4	85.12	77.34	70.74	65.43	**40.07**	**74.62**	135.54	10.4 fps
5	84.97	77.22	**70.69**	**65.45**	**40.07**	74.60	135.45	10.0 fps

TABLE XVI
THE NUMBER OF MODEL PARAMETERS AND FLOATING-POINT OPERATIONS (FLOPS).

Method	DE	CE	MBF	Parameters	FLOPs
Baseline	×	×	×	71.27M	21.48G
RSICCformer*	✓	✓	×	71.27M	21.48G
RSICCformer	✓	✓	✓	81.51M	23.48G

slightly increases parameters and computational complexity but significantly improves performance, as shown in Tab. VII, which we consider acceptable.

VI. DISCUSSION

To explore the RSICC task, we provide a large-scale remote sensing image change captioning dataset. We also propose an effective method to capture multiple changes of interest and ignore irrelevant changes. Extensive experiments have demonstrated the effectiveness of our method. However, there are still some problems that need to be addressed. With our dataset, researchers can focus on the following problems in the future:

- Our method utilizes the differences between bi-temporal features to determine the change regions of interest and recognize multiple changes. Exploring other methods to capture relationships between two images, localize change regions, and recognize the changes is significant.
- Recognizing changed objects of different sizes, their attributes, and object relationships is challenging. Although our method can ease the problem, the problem still needs to be further addressed in the future.
- Improving the ability of the model to distinguish whether changes of interest have occurred in the images is significant for the captioning accuracy.

VII. CONCLUSION

In this paper, we explore the Remote Sensing Image Change Captioning (RSICC) task and provide a large-scale real-world dataset for this task. The dataset contains 10077 pairs of bi-temporal remote sensing images and 50385 sentences. We also propose a novel method named RSICCformer for the RSICC task. Specifically, RSICCformer contains a dual-branch Transformer encoder consisting of a hierarchy of processing stages to capture and recognize multiple changes of interest. Extensive experiments have validated the effectiveness of our method, and also suggest that exploiting the difference between the corresponding positions of two images is crucial for improving the change captioning accuracy. Besides, we benchmark existing state-of-the-art synthetic image change captioning methods on the LEVIR-CC dataset, and our RSICCformer outperforms previous methods with a significant margin (+4.98% on BLEU-4 and +9.86% on CIDEr-D).

REFERENCES

[1] J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, *arXiv preprint arXiv:1910.06444*, 2019.

[2] K. Sakurada and T. Okatani, "Change detection from a street image pair using cnn features and superpixel segmentation." in *BMVC*, vol. 61, 2015, pp. 1–12.

[3] P. P. De Bem, O. A. de Carvalho Junior, R. Fontes Guimarães, and R. A. Trancoso Gomes, "Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks," *Remote Sensing*, vol. 12, no. 6, p. 901, 2020.

[4] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5407–5423, 2017.

[5] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[6] Q. Zhang, J. Wang, X. Peng, P. Gong, and P. Shi, "Urban built-up land change detection with road density and spectral information from multi-temporal landsat tm data," *International journal of remote sensing*, vol. 23, no. 15, pp. 3057–3078, 2002.

[7] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[8] A. Oluwasanmi, M. U. Aftab, E. Alabdulkreem, B. Kumeda, E. Y. Baagyere, and Z. Qin, "Captionnet: Automatic end-to-end siamese difference captioning model with attention," *IEEE Access*, vol. 7, pp. 106 773–106 783, 2019.

[9] A. Oluwasanmi, E. Frimpong, M. U. Aftab, E. Y. Baagyere, Z. Qin, and K. Ullah, "Fully convolutional captionnet: Siamese difference captioning attention model," *IEEE Access*, vol. 7, pp. 175 929–175 939, 2019.

[10] Y. Qiu, Y. Satoh, R. Suzuki, K. Iwata, and H. Kataoka, "3d-aware scene change captioning from multiview images," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4743–4750, 2020.

[11] Z. Liao, Q. Huang, Y. Liang, M. Fu, Y. Cai, and Q. Li, "Scene graph with 3d information for change captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5074–5082.

[12] D. H. Park, T. Darrell, and A. Rohrbach, "Robust change captioning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4623–4632.

[13] Y. Tu, T. Yao, L. Li, J. Lou, S. Gao, Z. Yu, and C. Yan, "Semantic relation-aware difference representation learning for change captioning," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 63–73.

[14] M. Hosseinzadeh and Y. Wang, "Image change captioning by learning from an auxiliary task," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2724–2733.

[15] K. E. Ak, Y. Sun, and J. H. Lim, "Learning by imagination: A joint framework for text-based image manipulation and change captioning," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[16] H. Kim, J. Kim, H. Lee, H. Park, and G. Kim, "Viewpoint-agnostic change captioning with cycle consistency," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2075–2084.

[17] Y. Qiu, S. Yamamoto, K. Nakashima, R. Suzuki, K. Iwata, H. Kataoka, and Y. Satoh, "Describing and localizing multiple changes with transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1951–1960.

[18] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1988–1997.

[21] S. Chouaf, G. Hoxha, Y. Smara, and F. Melgani, "Captioning changes in bi-temporal remote sensing images," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 2891–2894.

[22] G. Hoxha, S. Chouaf, F. Melgani, and Y. Smara, "Change captioning: A new paradigm for multitemporal remote sensing image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2022.

[23] V. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[24] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[25] J. Gao, C. Xu, L. Zhang, S. Liu, W. Feng, S. Xiong, and S. Tan, "Infrared image change detection of substation equipment in power system using markov random field," in *2017 International Conference on Computing Intelligence and Information System (CIIS)*. IEEE, 2017, pp. 332–337.

[26] M. Sumaiya and R. S. S. Kumari, "Logarithmic mean-based thresholding for sar image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 11, pp. 1726–1728, 2016.

[27] W. A. Malila, "Change vector analysis: An approach for detecting forest changes with landsat," in *LARS symposia*, 1980, p. 385.

[28] O. A. Carvalho Júnior, R. F. Guimarães, A. R. Gillespie, N. C. Silva, and R. A. Gomes, "A new approach to change vector analysis using distance and similarity measures," *Remote Sensing*, vol. 3, no. 11, pp. 2473–2493, 2011.

[29] J. Deng, K. Wang, Y. Deng, and G. Qi, "Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data," *International Journal of Remote Sensing*, vol. 29, no. 16, pp. 4823–4838, 2008.

[30] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering," *IEEE geoscience and remote sensing letters*, vol. 6, no. 4, pp. 772–776, 2009.

[31] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (mad) and maf postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sensing of Environment*, vol. 64, no. 1, pp. 1–19, 1998.

[32] A. A. Nielsen, "The regularized iteratively reweighted mad method for change detection in multi-and hyperspectral data," *IEEE Transactions on Image processing*, vol. 16, no. 2, pp. 463–478, 2007.

[33] J. B. Collins and C. E. Woodcock, "An assessment of several linear change detection techniques for mapping forest mortality using multitemporal landsat tm data," *Remote sensing of Environment*, vol. 56, no. 1, pp. 66–77, 1996.

[34] O. Abd El-Kawy, J. Rød, H. Ismail, and A. Suliman, "Land use and land cover change detection in the western nile delta of egypt using remote sensing data," *Applied geography*, vol. 31, no. 2, pp. 483–494, 2011.

[35] T. Chou, T. Lei, S. Wan, and L. Yang, "Spatial knowledge databases as applied to the detection of changes in urban land use," *International Journal of Remote Sensing*, vol. 26, no. 14, pp. 3047–3068, 2005.

[36] L. Bruzzone, R. Cossu, and G. Vernazza, "Detection of land-cover transitions by combining multidate classifiers," *Pattern Recognition Letters*, vol. 25, no. 13, pp. 1491–1500, 2004.

[37] T. Hame, I. Heiler, and J. San Miguel-Ayanz, "An unsupervised change detection and recognition system for forestry," *International journal of remote sensing*, vol. 19, no. 6, pp. 1079–1099, 1998.

[38] F. Gao, E. B. De Colstoun, R. Ma, Q. Weng, J. G. Masek, J. Chen, Y. Pan, and C. Song, "Mapping impervious surface expansion using medium-resolution satellite image time series: a case study in the yangtze river delta, china," *International Journal of Remote Sensing*, vol. 33, no. 24, pp. 7609–7628, 2012.

[39] A. P. Tewkesbury, A. J. Comber, N. J. Tate, A. Lamb, and P. F. Fisher, "A critical synthesis of remotely sensed optical image change detection techniques," *Remote Sensing of Environment*, vol. 160, pp. 1–14, 2015.

[40] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

[41] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved unet++," *Remote Sensing*, vol. 11, no. 11, p. 1382, 2019.

[42] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[43] S. Sun, L. Mu, L. Wang, and P. Liu, "L-unet: An lstm network for remote sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[44] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource vhr images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2848–2864, 2019.

[45] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[46] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[47] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[48] W. Zhao, X. Chen, X. Ge, and J. Chen, "Using adversarial network for multiple change detection in bitemporal remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[49] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2020.

[50] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection," *Remote Sensing*, vol. 12, no. 3, p. 484, 2020.

[51] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.

[52] J. Huang, Q. Shen, M. Wang, and M. Yang, "Multiple attention siamese network for high-resolution image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[53] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

[54] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[55] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002.

[56] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sensing*, vol. 12, no. 10, p. 1688, 2020.

[57] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *IEEE Access*, vol. 8, pp. 126 385–126 400, 2020.

[58] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, June 2017.

[59] A. Gupta and P. Mannem, "From image annotation to image description," in *International conference on neural information processing*. Springer, 2012, pp. 196–204.

[60] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[61] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*. Springer, 2010, pp. 15–29.

[62] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, vol. 24, 2011.

[63] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.

[64] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," *Advances in neural information processing systems*, vol. 27, 2014.

[65] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 256–270, 2020.

[66] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1274–1278, 2019.

[67] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International conference on computer, information and telecommunication systems (Cits)*. IEEE, 2016, pp. 1–5.

[68] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2018.

[69] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word-sentence framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[70] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[71] X. Ma, R. Zhao, and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[72] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, "A multi-level attention model for remote sensing image captions," *Remote Sensing*, vol. 12, no. 6, p. 939, 2020.

[73] S. Wu, X. Zhang, X. Wang, C. Li, and L. Jiao, "Scene attention mechanism for remote sensing image caption generation," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[74] C. Liu, R. Zhao, and Z. Shi, "Remote sensing image captioning based on multi-layer aggregated transformer," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–1, 2022.

[75] S. Zhuang, P. Wang, G. Wang, D. Wang, J. Chen, and F. Gao, "Improving remote sensing image captioning by combining grid features and transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[76] Y. Li, X. Zhang, J. Gu, C. Li, X. Wang, X. Tang, and L. Jiao, "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[77] Y. Wang, W. Zhang, Z. Zhang, X. Gao, and X. Sun, "Multiscale multiinteraction network for remote sensing image captioning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2154–2165, 2022.

[78] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1985–2000, 2019.

[79] R. Chavhan, B. Banerjee, X. X. Zhu, and S. Chaudhuri, "A novel actor dual-critic model for remote sensing image captioning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4918–4925.

[80] G. Hoxha and F. Melgani, "A novel svm-based decoder for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[81] Z. Chen, J. Wang, A. Ma, and Y. Zhong, "Typeformer: Multiscale transformer with type controller for remote sensing image caption," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[82] S. Zhuang, P. Wang, G. Wang, D. Wang, J. Chen, and F. Gao, "Improving remote sensing image captioning by combining grid features and transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[83] H. Kandala, S. Saha, B. Banerjee, and X. X. Zhu, "Exploring transformer and multi-label classification for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–1, 2022.

[84] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, April 2018.

[85] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[86] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[87] B. Zhao, "A systematic survey of remote sensing image captioning," *IEEE Access*, vol. 9, pp. 154 086–154 111, 2021.

[88] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: a survey on deep learning-based image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[89] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[90] Papineni, Kishore, Roukos, Salim, Ward, Todd, Zhu, and Wei-Jing, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[91] Lin and C. Yew, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.

[92] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, ser. StatMT '07. USA: Association for Computational Linguistics, 2007, p. 228–231.

[93] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[94] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[95] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[96] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[97] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[98] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[99] X. Shi, X. Yang, J. Gu, S. Joty, and J. Cai, "Finding it at another side: A viewpoint-adapted matching encoder for change captioning," in *European Conference on Computer Vision*. Springer, 2020, pp. 574–590.

[100] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[101] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," *arXiv preprint arXiv:1702.01806*, 2017.

[102] J. Lee, J.-H. Shin, and J.-S. Kim, "Interactive visualization and manipulation of attention-based neural machine translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 121–126.

[103] F. Stahlberg, "Neural machine translation: A review," *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020.

[104] X. Zhang, W. Chen, F. Wang, S. Xu, and B. Xu, "Towards compact and fast neural machine translation using a combined method," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1475–1481.

Chenyang Liu received his B.S. degree from the Image Processing Center, School of Astronautics, Beihang University in 2021. He is currently working towards the M.S. degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include machine learning, computer vision and multimodal learning.

Rui Zhao received the B.S. and M.S. degrees from the Image Processing Center, School of Astronautics, Beihang University, in 2019 and 2022, respectively. He is currently a Researcher with NetEase's Fuxi AI Laboratory. His research interests include computer vision, deep learning, and related problems in remote sensing and video games.

Hao Chen received his B.S. degree from the Image Processing Center School of Astronautics, Beihang University in 2017. He is currently pursuing his doctorate degree in the Image Processing Center, School of Astronautics, Beihang University.His research interests include machine learning, deep learning and semantic segmentation.

Zhengxia Zou received his BS degree and his Ph.D. degree from Beihang University in 2013 and 2018. He is currently an Associate Professor at the School of Astronautics, Beihang University. During 2018-2021, he was a postdoc research fellow at the University of Michigan, Ann Arbor. His research interests include computer vision and related problems in remote sensing. He has published more than 20 peer-reviewed papers in top-tier journals and conferences, including TPAMI, TIP, TGRS, CVPR, ICCV, AAAI. His research was featured in more than 30 global tech media and was adopted by a number of application platforms with over 50 million users worldwide. His personal website is https://zhengxiazou.github.io/.

Zhenwei Shi (Member, IEEE) received the Ph.D. degree in mathematics from the Dalian University of Technology, Dalian, China, in 2005.

He was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, from 2013 to 2014. He is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing. He has authored or coauthored over 200 scientific papers in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters, the IEEE Conference on Computer Vision and Pattern Recognition, and the IEEE International Conference on Computer Vision. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi also serves as an Editor for the Pattern Recognition, the ISPRS Journal of Photogrammetry and Remote Sensing, the Infrared Physics and Technology, and so on. His personal website is http://levir.buaa.edu.cn/.