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

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

***Abstract*— Analyzing land cover changes with multitemporal remote sensing (RS) images is crucial for environmental protection and land planning. In this article, we explore RS image change captioning (RSICC), a new task aiming at generating human-like language descriptions for the land cover changes in multitemporal RS images. We propose a novel Transformer-based RSICC (RSICCformer) model. 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 (DTE) to improve the feature discrimination capacity for the changes; and 3) a caption decoder to generate sentences describing the differences. The DTE consists of a hierarchy of processing stages to capture and recognize multiple changes of interest. Concretely, we use the bitemporal feature differences as keys to enhance image features (queries) from each temporal image in the dual-branch Transformer encoder (DTE). To explore the RSICC task, we build a large-scale dataset named LEVIR-CC, which contains 10 077 pairs of bitemporal RS images and 50 385 sentences describing the differences between images. We benchmark existing state-of-the-art synthetic image change captioning methods on the LEVIR Change Captioning dataset (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*— Change captioning (CC), change detection (CD), image captioning, remote sensing (RS) images, Transformer.**

I. Introduction

THE development of human society and the evolution of the natural environment are accelerating global surface changes. The availability of multitemporal remote sensing (RS) images provides an opportunity to study surface changes [1], [2], [3]. The RS image change detection (RSICD) determining pixel-level changed regions has become an emerging RS image interpretation task [3], [4], [5], [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 article, we explore a new task in the RS field, namely RS 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], [8], [9], 3-D scene CC [10], [11], and synthetic image CC [12], [13], [14], [15], [16], [17]. For instance, Oluwasanmi et al. [8] use 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], the CLEVR-Multi-Change [17],

Manuscript received 27 April 2022; revised 17 September 2022; accepted 29 October 2022. Date of publication 1 November 2022; date of current version 14 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62125102 and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhenwei Shi.)*

Chenyang Liu, 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 (e-mail: shizhenwei@buaa.edu.cn).

Rui Zhao is with the Fuxi AI Laboratory, NetEase, Hangzhou, Zhejiang 310052, China.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China.

Digital Object Identifier 10.1109/TGRS.2022.3218921

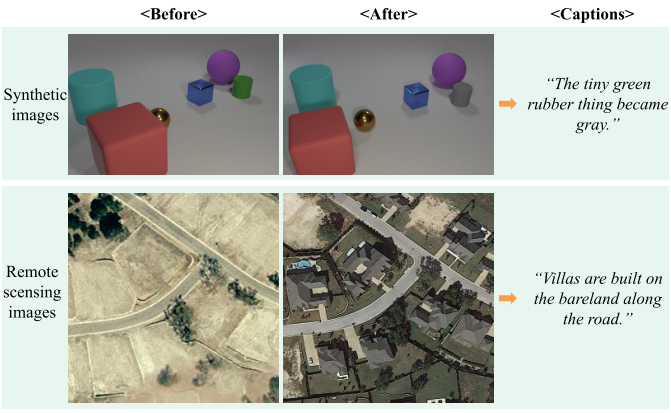

Fig. 1. Illustration of CC. The synthetic images of the CLEVR-Change dataset are automatically generated by the CLEVR engine [20], and the object categories are monotonous. However, RS images of our dataset have a broader scale range, richer object categories, and more complex ground details. Besides,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

and the CLEVR-DC dataset [16]) have provided fundamentals and testbeds for designing and evaluating CC methods. However, compared to the synthetic scenes, the real RS scenes have a broader scale range, richer object categories, and more complex ground details. That increases the difficulty of the CC with RS images. Besides, the RS community currently lacks a publicly available large dataset for the RSICC task. Therefore, we build a large-scale dataset with real RS 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 Change Captioning dataset (LEVIR-CC dataset) contains 10 077 pairs of bitemporal RS images and 50 385 sentences describing the differences between images. The novel dataset provides an opportunity to explore models that align visual changes and language. In this article, we refer to relatively small and uninteresting changes as irrelevant changes, such as light changes, which bring significant challenges to CC. 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; and 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 bitemporal images. Some fusion strategies are designed to fuse the bitemporal 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 article, 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 DTE, and a caption decoder. The CNN-based feature extractor generates bitemporal 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 (DE) module utilizing bitemporal 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 multistage bitemporal 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 DTE 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 multihead attention computing patch-level similarity. Specifically, MCCFormers-S concatenates bitemporal 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 coattention mechanism [24] to capture relevance between local regions of two images. For the coattention 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 bitemporal 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.

1) We conduct an in-depth study of an emerging task named RSICC that aims at generating human-like language descriptions for the ground feature changes in multitemporal 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.
2) We propose a novel RSICCformer model. 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.
3) We build a publicly available large-scale dataset named LEVIR-CC to advance the CC task in the RS community. We benchmark existing state-of-the-art synthetic image CC 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 (CD) and image captioning. Unlike CD requiring pixel-level change localization, CC 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, CC requires describing the visual changes of interest in image pairs. Here, we discuss previous work on RSICD and RS image captioning.

A. Remote Sensing Change Detection

The RS CD task aims at locating and recognizing pixel-level changes in the bitemporal 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 CD 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 multitemporal 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 [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 multitemporal 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 multitemporal images for recognizing change areas. Nielsen [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], [35], [36] or direct classification (i.e., performing classification on the data stack consisting of multitemporal images) [37], [38]. For example, Bruzzone et al. [36] proposed a land-cover CD system composed of multiple classifiers and a compound classification decision strategy. Tewkesbury et al. [39] provided a comprehensive review of traditional CD methods.

Due to superior feature representation capability compared to traditional methods, deep learning (DL) has recently achieved great success in CD. Most methods use DL-based extractors to extract high-level semantic features from raw multitemporal images and use DL-based classifiers to generate change maps. Commonly used networks of extractors and classifiers include CNNs [40], [41], [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 CD 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 multiscale feature extraction and a multiple side-output fusion strategy for deep supervision. Daudt et al. [42] proposed three fully convolutional networks (FCNs) for CD 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], [50], [51], [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 coattention module captures the correlation between bitemporal 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 LSTM (Conv-LSTM) [40] layers into UNet [53] to fuse spatial and temporal features for CD. 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 bitemporal image Transformer (BIT), in which a semantic tokenizer represents the images as a few visual words, and the Transformer utilizes them to refine the bitemporal features extracted by CNN. BIT can efficiently and effectively identify the change of interest. Zhang et al. [46] proposed a pure Transformer CD 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 [54]. 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 bitemporal images with building-related changes and corresponding masks. Zhao et al. [48] proposed an attention gates' generative adversarial adaptation network (AG-GAAN) for CD. 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 CD methods can be found in [56] and [57].

B. Remote Sensing Image Captioning

RS image captioning is an active research topic at the intersection of RS 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 [57], [58], [59] and retrieval-based methods [60], [61], [62], [63], [64], [65]. The template-based methods detect ground objects and fill the corresponding words into a human-designed template. For example, Shi and Zou [57] employed an 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. [65] 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- and retrieval-based methods cannot generate novel sentences, and the sentences are relatively limited and rigid.

Current image captioning methods are mainly based on DL-based generative models. The most widely used framework is the encoder–decoder framework [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], 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], and [79], SVMs in [79], and Transformers in [74], [81], [82], and [83].

Qu et al. [66] first explored CNNs as encoders and RNNs as decoders to generate sentences describing RS images. Lu et al. [67] provided a large dataset named RSICD for RS image captioning and compared the effects of handcrafted 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 multiscale object information, Ma et al. [70] used multihead 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. [71] 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. [82] proposed a global visual feature-guided attention (GVFGA) module. The module performs a mean pooling operation to obtain global features, 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. [69] 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. [73] used a Transformer [83] 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. [80] proposed TypeFormer, in which a multiscale vision Transformer is used to capture multiscale 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] and [88].

III. Proposed LEVIR-CC Dataset

Our LEVIR-CC dataset contains bitemporal images and corresponding sentences describing differences.[1] We will make the LEVIR-CC dataset publicly available, and we believe that 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 CD 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 bitemporal 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 bitemporal images to 256 × 256 pixels in our LEVIR-CC dataset. Besides, the bitemporal 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 bitemporal 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.

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

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

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, and so on.

"no change has occurred," and "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, and so on.

C. Dataset Analysis

1) Image Pairs: The LEVIR-CC dataset contains 5038 image pairs with changes and 5039 image pairs without changes. Table 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 and villa); 2) parking lot changes; 3) road-related changes (e.g., crossroad and path); 4) vegetation-related changes (e.g., tree and bush); and 5) water-related changes (e.g., river and lake). Table 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
NUMBER OF BITEMPORAL 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 50 385 sentences describing the differences between images. We report

TABLE II

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 sentence quantity, the word quantity, and the average sentence length in the LEVIR-CC dataset, as shown in Table III. The 25 195 sentences describing image pairs without changes are relatively short, and the average sentence length is five words. The average length of 25 190 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 five and 15 words in length.

The annotated sentences of our dataset contain a total of 402 512 words, as shown in Table 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, and 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], [86] 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 [87], ROUGE-L [88], and METEOR [89]

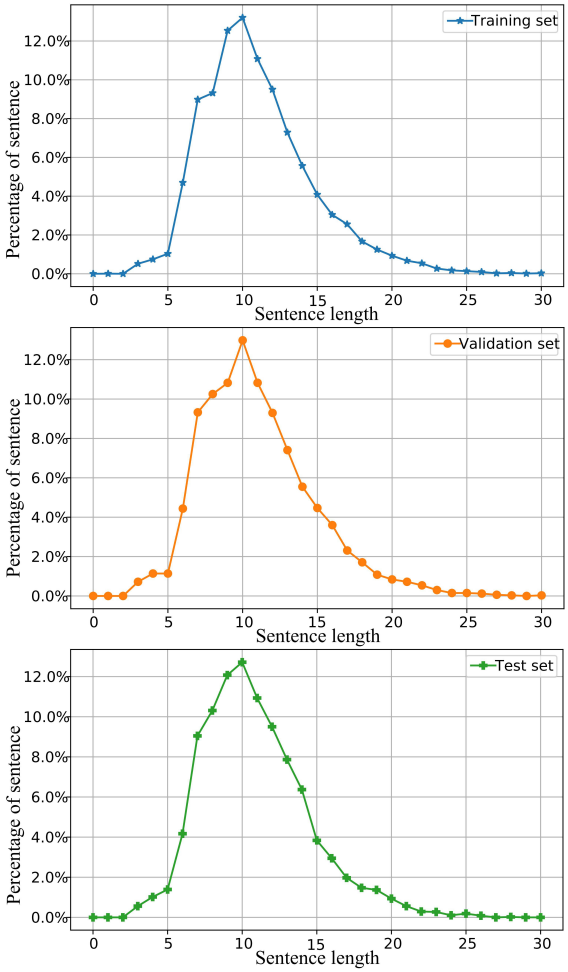

Fig. 3. Percentages of sentences of different lengths for 25 190 sentences describing image pairs with changes in three sets. The sentence length distribution in the three sets is similar, and most of these sentences are between five and 15 words in length.

TABLE IV

HUMAN AGREEMENT COMPARISON OF LEVIR-CC AND MS-COCO C5 DATASETS. 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

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 [86], which is a large manually annotated image captioning dataset with five annotated sentences per image. Table IV shows the comparison results. We can see that our dataset is reasonable and comparable to MS-COCO C5 in terms of human agreement.

4) Dataset Comparison: Table V compares our dataset with three existing CC 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

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, and 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
COMPARISON OF OUR LEVIR-CC WITH THREE EXISTING CC 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

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 bitemporal images, so 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, RS 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 RS image changes and human language. This dataset will advance the CC task in the RS 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 DTE to improve the feature discrimination capacity for the changes; and 3) a caption decoder to generate sentences describing the differences. The DTE consists of a hierarchy of processing stages to capture and recognize multiple changes of interest. Concretely, we use the bitemporal 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.

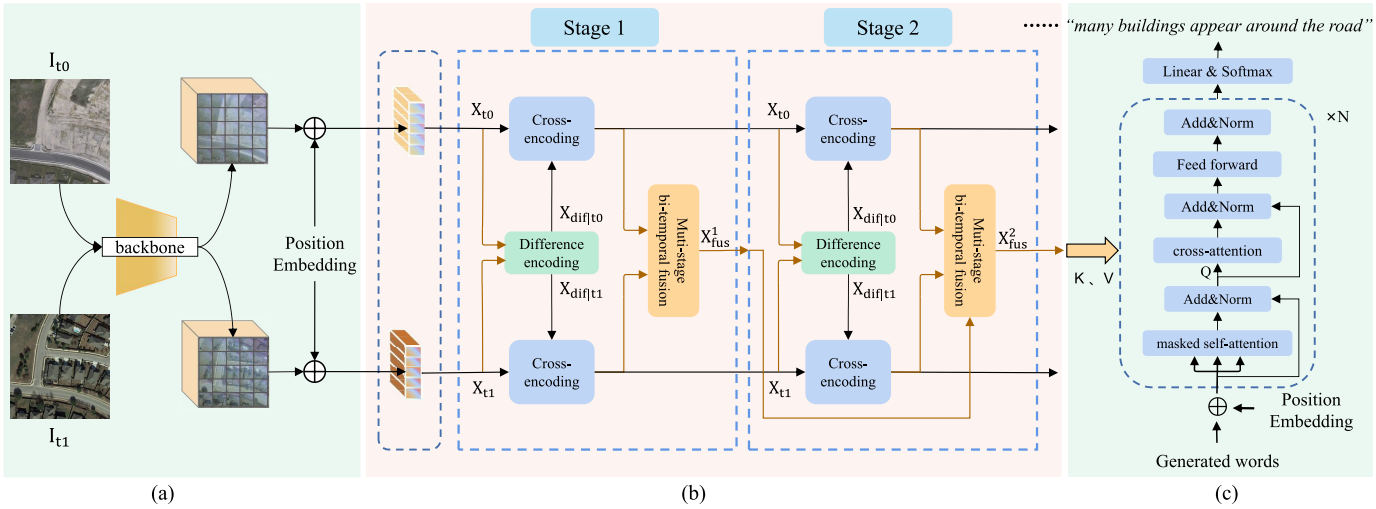

Fig. 5. Overall structure of our RSICCformer-based model. The proposed model consists of three main components: a CNN-based feature extractor, a DTE, and a caption decoder. The CNN-based feature extractor generates bitemporal image features. We flatten the bitemporal feature maps into two sequences, respectively, add the learnable position embedding, and then feed them into the 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 DE module, a Siamese CE module, and an MBF module. (a) Feature extraction. (b) DTE. (c) Caption decoder.

Algorithm 1 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 multistage 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 $Caption$

The DE module utilizes bitemporal features to obtain the high-level semantic features revealing the differences between two images. As shown in Fig. 5, the output of the DE module contains $X_{\text{dif}|t_0}$ and $X_{\text{dif}|t_1}$. We define $X_{\text{dif}|i}$ $(i = t_0, t_1)$ as the difference representation, which will be fed into the Siamese CE modules along with 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 DE strategies to generate $X_{\text{dif}|i}$ $(i = t_0, t_1)$. These strategies can be formulated as follows:

$$\text{DE}_1 : X_{\text{dif}|t_0} = X_{\text{dif}|t_1} = X_{t_1} - X_{t_0} \tag{1}$$

$$\text{DE}_2 : X_{\text{dif}|t_0} = X_{\text{dif}|t_1} = \text{Abs}(X_{t_1} - X_{t_0}) \tag{2}$$

$$\text{DE}_3 : X_{\text{dif}|i} = \text{RELU}(\text{conv}([X_{t_1} - X_{t_0}; X_i])) \tag{3}$$

where Abs denotes the elementwise operation of absolute value. conv denotes 2-D convolution for feature transformation. Rectified linear unit (RELU) is an activation function.

B. Siamese Cross-Encoding Modules

Our intuition is that the Siamese CE modules utilize a difference representation from the DE module to capture changes. Fig. 6 illustrates the structure of the CE module. Unlike the Transformer encoding layer [83], we replace the self-attention mechanism with a cross-attention mechanism. Each CE module contains a residual unit with the multihead cross-attention (MCA) mechanism and a residual unit with the feed-forward network. Layer normalization (LN) [90] 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 $X_{\text{dif}|i}$ $(i = t_0, t_1)$ output by the DE module. Assume that $X^l_i \in R^{HW \times C}$ and $X_{\text{dif}|i} \in R^{HW \times C}(i = t_0, t_1)$ are input to the CE module of the lth DTE stage. H, W, and C, respectively, are the height, width, and channel dimension of the feature map. The cross-attention mechanism of the CE module at the lth DTE stage can be expressed as follows:

$$Q_i = X^l_i W^Q \tag{4}$$

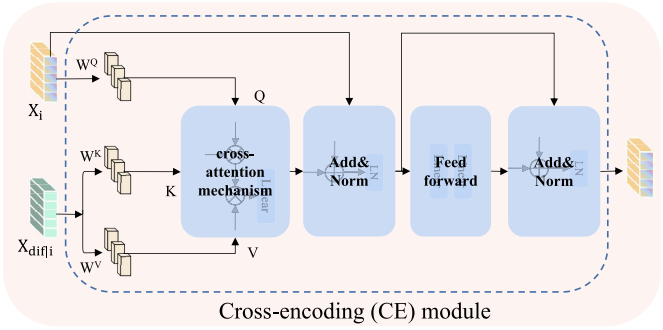

Fig. 6. Structure of the CE module. The module contains two residual units, and LN [90] is performed after each residual unit.

$$K_i = X^l_{\text{dif}|i} W^K \tag{5}$$
$$V_i = X^l_{\text{dif}|i} W^V \tag{6}$$

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, and d is the dimension of Q_i, K_i and V_i. Single-head cross-attention can be formulated as follows:

$$\text{CrossAtt}(Q_i, K_i, V_i) = \sigma\left(\frac{Q_i {K_i}^T}{\sqrt{d}}\right) V_i \tag{7}$$

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 proven that the MCA mechanism pays attention to input features from different aspects [83]. 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 lth DTE stage can be formulated as follows:

$$\text{MultiHead}\left(X^l_i, X^l_{\text{dif}|i}\right) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^o \tag{8}$$

where

$$\text{head}_j = \text{CrossAtt}\left(X^l_i W^Q_j, X^l_{\text{dif}|i} W^K_j, X^l_{\text{dif}|i} W^V_j\right) \tag{9}$$

where $W^Q_j, W^K_j, W^V_j \in R^{C\times d}$ are trainable matrices of the jth head, $W^o \in R^{hd\times C}$ is a trainable linear projection matrix to transform the feature channel dimension, and h is the head number of the MCA.

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

$$\text{FF}(X) = \text{RELU}(X W_1) W_2 \tag{10}$$

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 lth DTE stage can be expressed as follows:

$$\text{CE}\left(X^l_i\right) = \text{LN}(X + \text{FF}(X)) \tag{11}$$
$$X = \text{LN}\left(X^l_i + \text{MultiHead}\left(X^l_i, X^l_{\text{dif}|i}\right)\right) \tag{12}$$

where LN denotes the layer normalization [90].

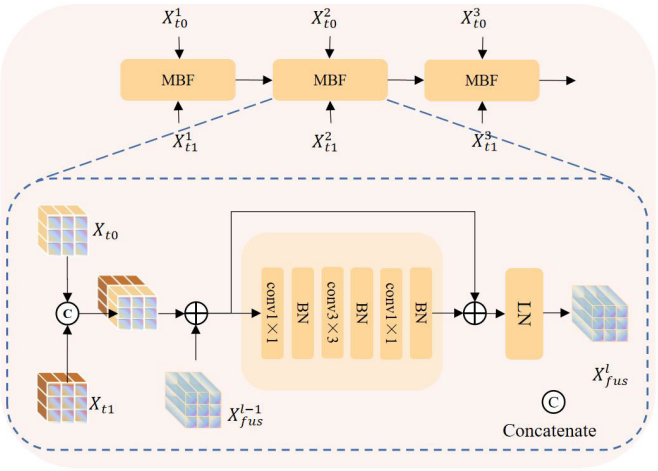

Fig. 7. Structure of the MBF module. MBF utilizes the bitemporal features from the Siamese CE modules and features from the previous MBF module to obtain high-level semantic features revealing multiple changes of interest.

C. Multistage Bitemporal 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 bitemporal RS images, we designed an 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 bitemporal 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 bitemporal feature fusion. The LN is then performed, which is proven to be valuable in experiments. The procedure of MBF at the lth DTE stage can be formulated as follows:

$$X^l = \left[X^l_{t_0}; X^l_{t_1}\right] \tag{13}$$
$$X^l_{\text{fus}} = X^l + X^{l-1}_{\text{fus}} \tag{14}$$
$$X^l_{\text{res}} = \text{conv}_3\left(\text{RELU}\left(\text{conv}_2\left(\text{RELU}\left(\text{conv}_1\left(X^l_{\text{fus}}\right)\right)\right)\right)\right) \tag{15}$$
$$X^l_{\text{fus}} = \text{LN}\left(X^l_{\text{fus}} + X^l_{\text{res}}\right) \tag{16}$$

where $X^l_i \in R^{H\times W\times d} (i = t_0, t_1)$ are the reshaped features from the Siamese CE modules at the lth DTE stage. [;] denotes the concatenation operation. conv_1, conv_2, and conv_3 are 2-D convolutions with 1×1, 3×3, and 1×1 kernels. 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 sentences describing the differences between images. For the caption decoder, we use a standard Transformer decoder [83], which is a stack of multiple decoding layers. Each decoding layer contains three sublayers, including multihead masked self-attention,

TABLE VI
EFFECTS OF THREE DE 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**

encoder–decoder attention (EDAtt), and the feed-forward network. Besides, each sublayer contains a residual connection and is followed by the LN operation.

Unlike multihead self-attention, the multihead masked self-attention masks out the values at position t and subsequent positions when predicting the tth word in the training stage. It ensures that each word prediction only depends on the generated words [83]. For the EDAtt of the caption decoder, it is similar to the MCA of the Siamese CE modules in the DTE. However, Q is from the previous masked self-attention sublayer, and K and V are from the transformation of X_{fus}. The EDAtt of the lth Transformer decoding layer can be formulated as follows:

$$\text{EDAtt}\left(S^l, X_{\text{fus}}\right) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^o \tag{17}$$

where

$$\text{head}_j = \text{CrossAtt}\left(S^l W_j^Q, X_{\text{fus}} W_j^K, X_{\text{fus}} W_j^V\right) \tag{18}$$

where $W_j^Q, W_j^K, W_j^V \in R^{d\times d}$ are trainable matrices of the jth 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 sublayer, L is the length of the sentence, and h is the head number of the MCA.

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

$$\boldsymbol{p_t} = \text{Softmax}(\boldsymbol{y_t}) = \frac{\exp(\boldsymbol{y_t})}{\sum_{i=1}^{K}\exp\left(y_t^{(i)}\right)} \tag{19}$$

where $\boldsymbol{y}_t = [y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(K)}]$ 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 [91] pretrained on the ImageNet dataset [92] as the backbone to extract image features from bitemporal 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 DTE and one decoding layer in the caption decoder. The head number of all multihead 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 DTE and the caption decoder. We use the learnable 2-D position embedding [93] to incorporate spatial position information in our mode before bitemporal feature sequences are fed to the DTE. 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 [83].

3) Loss Function: The simple and effective cross-entropy loss has been widely used in captioning-related tasks [67], [69], [70], [73]. 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:

$$\text{loss} = -\sum_{t=1}^{L}\log\left(\sum_{k=1}^{K}\widetilde{y}_t^{(k)}p_t^{(k)}\right) \tag{20}$$

where L is the sentence length and K is the vocabulary size. $\widetilde{\boldsymbol{y}}_l = [\widetilde{y}_t^{(1)}, \widetilde{y}_t^{(2)}, \ldots, \widetilde{y}_t^{(K)}]$ is the vector representation of the word at position t in the reference sentence. $\boldsymbol{p}_t = [p_t^{(1)}, p_t^{(2)}, \ldots, p_t^{(K)}]$ 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 DL framework. All models are trained and evaluated on the NVIDIA GTX 1080Ti GPU. In the training stage, we use the Adam optimizer [94]. We set the maximum epoch to 40. We set 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 bitemporal images. The automatic evaluation metrics can automatically measure the accuracy of the generated sentences based on the annotated reference sentences. In this

TABLE VII

ABLATION STUDIES ON THE SIAMESE CE MODULES AND THE 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**

work, we use four different metrics to evaluate the captioning accuracy, including BLEU-N ($N = 1, 2, 3, 4$) [87], ROUGE-L [88], METEOR [89], and CIDEr-D [95], which are widely used in the RS image captioning tasks [67], [69], [70], [73] and CC tasks [12], [13], [14], [15], [16], [17], [96]. Vedantam et al. [95] 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 DE module utilizes bitemporal features to obtain the feature representation that reveals the differences between bitemporal images. Siamese CE modules then utilize the feature difference to recognize multiple changes. Table VI reports the effects of the three DE strategies, which have been introduced in Section IV-A. The baseline does not use the difference information of bitemporal images. It employs Siamese Transformer encoders without the DE module to process bitemporal features separately. We can see that the three DE 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 the 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; and 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.

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

1) Siamese Cross-Encoding Module: Table VII shows that RSICCformer* outperforms baseline in three test ways. This demonstrates that the Siamese CE modules are effective and can significantly improve the CC performance of the model. It is attributed to the difference representation being utilized by the Siamese CE module. Since bitemporal 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 CC methods have attempted to correlate identical regions of two images with viewpoint change in order to capture and recognize changes [12], [16], [17], [96]. 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 encoders, which make their respective salient regions to be paid attention to. For our method, the attention maps for the bitemporal images seem

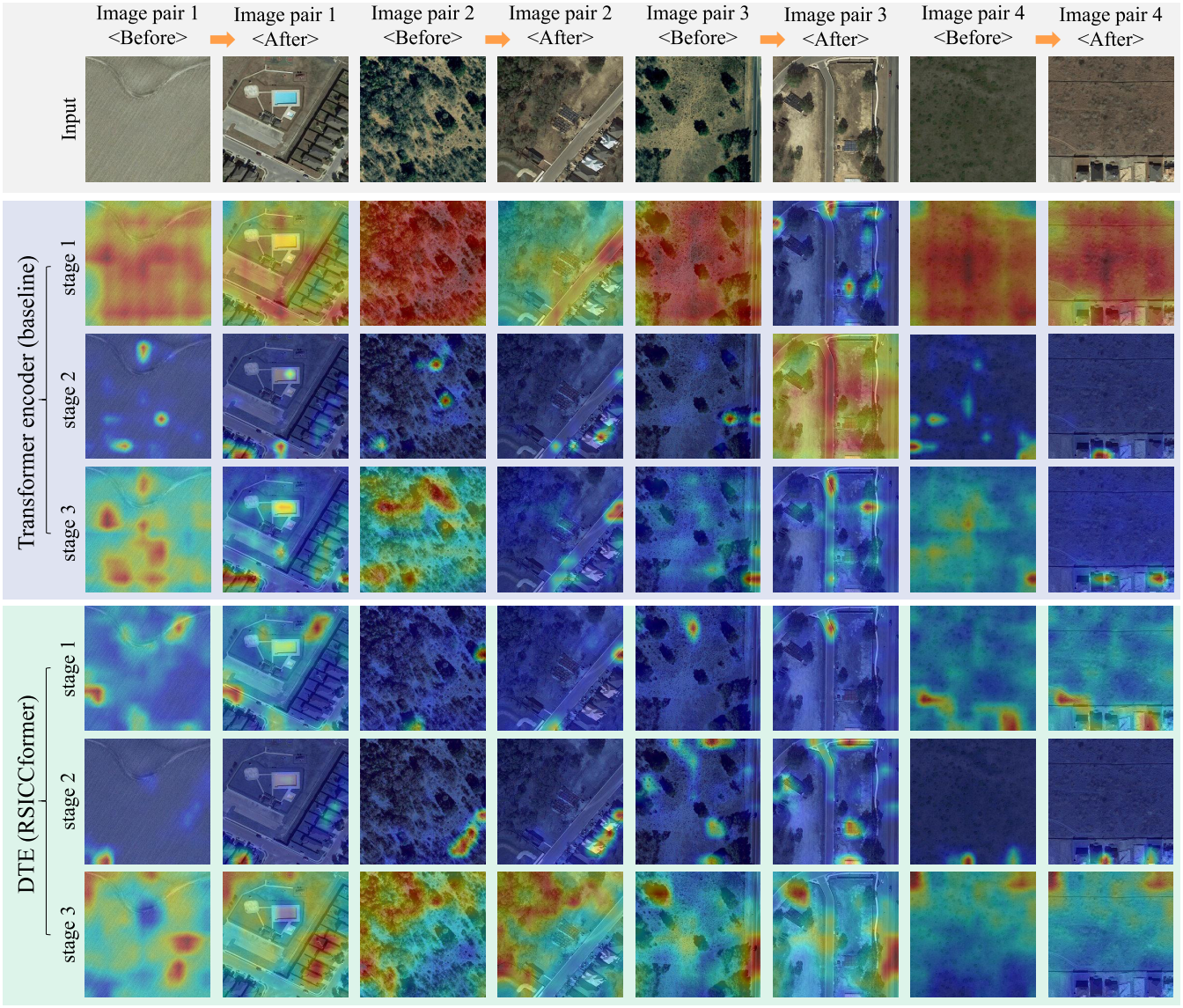

Fig. 8. Visualization of the attention states of each stage in the Transformer encoder of baseline and DTE of RSICCformer. Specifically, for the baseline, the attention maps are obtained from the self-attention of the Transformer encoding layer. For the DTE, the attention maps are obtained from the cross-attention 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.

to be the same. It illustrates that the Siamese CE modules can pay attention to the positions of changed regions in the bitemporal 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 focuses on the building change, and the third stage focuses on the vegetation change.

2) Multistage Bitemporal Fusion Module: Fig. 8 shows that the Siamese CE modules can focus on different changes at different stages. Besides, bitemporal 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 bitemporal RS images, we designed the MBF to utilize the bitemporal features from different stage CE modules to obtain high-level semantic feature representation, which reveals multiple changes of interest and excludes irrelevant changes. Table VII demonstrates that MBF can significantly improve the model performance, which is attributed to the ability of MBF to incorporate multiple changes of interest and exclude irrelevant changes.

TABLE VIII
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**%

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. Table 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*, Table 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, Tables VII and VIII 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 CC methods on our LEVIR-CC dataset, including four LSTM-based methods (Capt-Rep-Diff [12], Capt-Att [12], Capt-Dual-Att [12], and DUDA [12]) and two Transformer-based methods (MCCFormers-S [17] and MCCFormers-D [17]). These methods are introduced as follows.

1) *Capt-Rep-Diff [12]:* ResNet-101 extracts the "before" feature map and "after" feature map from the bitemporal images. The elementwise 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.
2) *Capt-Att [12]:* It 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.
3) *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.
4) *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," and "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.
5) *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.
6) *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 coattention 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.

Table IX shows the comparison results of our RSICCformer and existing state-of-the-art synthetic image CC 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 bitemporal feature difference of corresponding positions. Since bitemporal 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. Table X shows the comparison results of our method and other methods on the Spot-the-Diff dataset. Though our method is proposed for RS CC, it still

TABLE IX

COMPARISONS WITH OTHER EXISTING STATE-OF-THE-ART SYNTHETIC IMAGE CC 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**

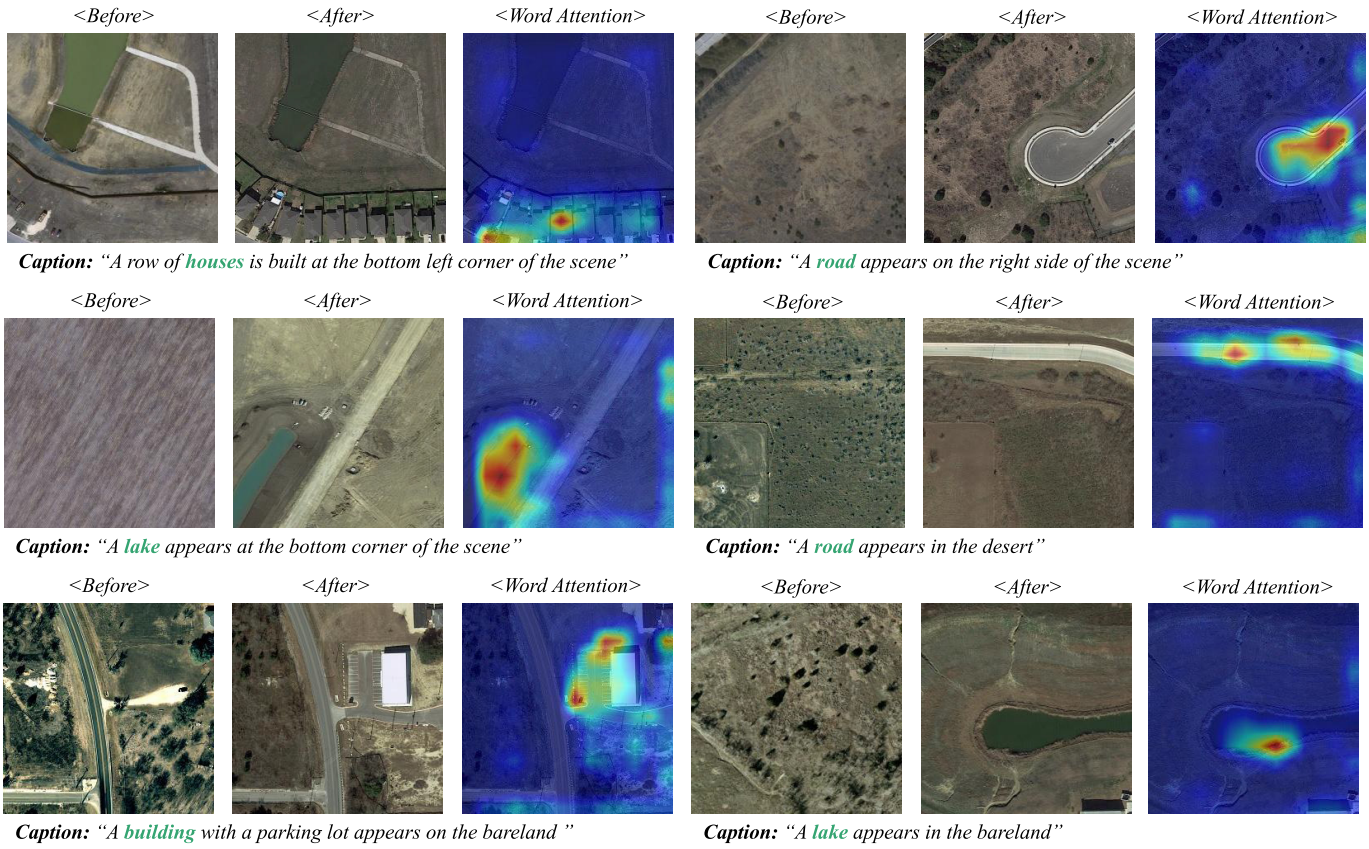

Fig. 9. Visualization of corresponding attention states of the caption decoder when our model generates an object-related word. We mark that word in green in the generated sentence. We superimpose the corresponding attention states on the "after" image to observe the attended image regions better. We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

obtains scores comparable to the 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 the baseline. For example, for the first image pair, our model can not only describe objects such as "houses" and "roads" but also scene concepts

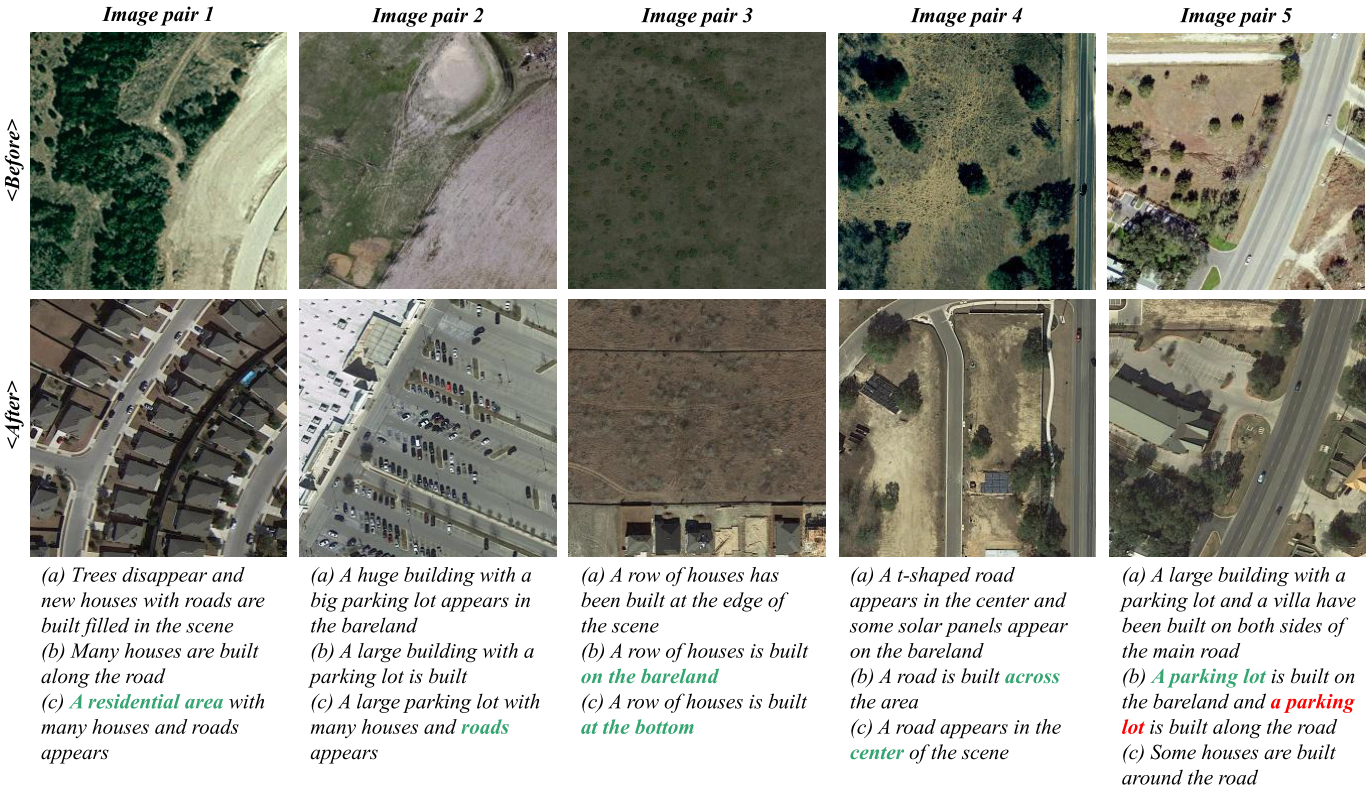

Fig. 10. Captioning results on the LEVIR-CC dataset. Sentence (a) is one of the five ground-truth sentences. The baseline and our RSICCformer generate sentences (b) and (c). More accurate and detailed words are marked in green. Red words are not accurate. We can observe that our RSICCformer can generate more accurate and detailed descriptions than the 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 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**

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

Table 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 bitemporal features, including three CNN-based backbones (VGG-16 [97], ResNet-50, and ResNet-101 [91]) and two Transformer-based backbones (ViT-B/16 and ViT-L/16 [93]). These backbones have been pretrained on the ImageNet dataset [92]. Table XII reports captioning metrics, training time per epoch measured in minutes (min), and inference speed measured in frames/s (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 bitemporal images in our RSICCformer model.

H. Parameter Analysis

1) Depth of the Network: The DTE and the caption decoder are stacked by multiple stages, and the depth of

TABLE XI
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
PERFORMANCE OF RSICCFORMER-BASED MODEL EMPLOYING DIFFERENT BACKBONES, INCLUDING THREE CNN-BASED BACKBONES (VGG-16 [97], RESNET-50, AND RESNET-101 [91]) AND TWO TRANSFORMER-BASED BACKBONES (VIT-B/16 AND VIT-L/16 [93]). THE TRAINING TIME PER EPOCH IS MEASURED IN MINUTES (min), AND THE INFERENCE SPEED IS MEASURED IN 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

TABLE XIII
PERFORMANCE OF THE MODEL IN DIFFERENT DEPTHS ON THE LEVIR-CC DATASET. E.D. DENOTES THE DEPTH OF THE DTE, AND D.D. DENOTES 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

the network is an important hyperparameter. We compare the performance of models of different depths, as shown in Table XIII. E.D. denotes the depth of the DTE, 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.

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: $(H \times W, d) \rightarrow (H, W, d)$; and 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 [93] in those three positions, as shown in Table XIV. We can observe that the position embedding is effective. Besides, the model performs best when we add position embedding at position 1.

3) Beam Search Strategy for Inference: The beam search strategy [98] has been widely used to improve model performance in machine translation [99], [100], [101] and captioning tasks [69], [73]. 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. Table 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

Table 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

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

NUMBER OF MODEL PARAMETERS AND 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 Table VII, which we consider acceptable.

VI. DISCUSSION

To explore the RSICC task, we provide a large-scale RSICC 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.

1) Our method utilizes the differences between bitemporal 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.
2) 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.
3) 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 article, we explore the RSICC task and provide a large-scale real-world dataset for this task. The dataset contains 10 077 pairs of bitemporal RS images and 50 385 sentences. We also propose a novel method named RSICCformer for the RSICC task. Specifically, RSICCformer contains a DTE 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 CC accuracy. Besides, we benchmark existing state-of-the-art synthetic image CC 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, "Building damage detection in satellite imagery using convolutional neural networks," 2019, *arXiv:1910.06444*.
[2] K. Sakurada and T. Okatani, "Change detection from a street image pair using CNN features and superpixel segmentation," in *Proc. BMVC*, vol. 61, 2015, pp. 1–12.
[3] P. de Bem, O. de Carvalho Junior, R. Fontes Guimarães, and R. Trancoso Gomes, "Change detection of deforestation in the Brazilian Amazon using landsat data and convolutional neural networks," *Remote Sens.*, vol. 12, no. 6, p. 901, Mar. 2020.
[4] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5407–5423, Sep. 2017.

[5] H. Chen and Z. Shi, "A spatial–temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, 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," *Int. J. Remote Sens.*, vol. 23, no. 15, pp. 3057–3078, Jan. 2002.

[7] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1–11.

[8] A. Oluwasanmi, "CaptionNet: Automatic end-to-end Siamese difference captioning model with attention," *IEEE Access*, vol. 7, pp. 106773–106783, 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. 175929–175939, 2019.

[10] Y. Qiu, Y. Satoh, R. Suzuki, K. Iwata, and H. Kataoka, "3D-aware scene change captioning from multiview images," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4743–4750, Jul. 2020.

[11] Z. Liao, Q. Huang, Y. Liang, M. Fu, Y. Cai, and Q. Li, "Scene graph with 3D information for change captioning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5074–5082.

[12] D. H. Park, T. Darrell, and A. Rohrbach, "Robust change captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4623–4632.

[13] Y. Tu et al., "Semantic relation-aware difference representation learning for change captioning," in *Proc. Findings Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2021, pp. 63–73.

[14] M. Hosseinzadeh and Y. Wang, "Image change captioning by learning from an auxiliary task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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 Trans. Multimedia*, early access, Feb. 24, 2022, doi: 10.1109/TMM.2022.3154154.

[16] H. Kim, J. Kim, H. Lee, H. Park, and G. Kim, "Viewpoint-agnostic change captioning with cycle consistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2075–2084.

[17] Y. Qiu et al., "Describing and localizing multiple changes with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1951–1960.

[18] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1988–1997.

[21] S. Chouaf, G. Hoxha, Y. Smara, and F. Melgani, "Captioning changes in bi-temporal remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[23] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[24] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[25] J. Gao et al., "Infrared image change detection of substation equipment in power system using Markov random field," in *Proc. Int. Conf. Comput. Intell. Inf. Syst. (CIIS)*, Apr. 2017, pp. 332–337.

[26] M. N. Sumaiya and R. S. S. Kumari, "Logarithmic mean-based thresholding for SAR image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1726–1728, Nov. 2016.

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

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

[29] J. S. Deng, K. Wang, Y. H. Deng, and G. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, 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 Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 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 Sens. Environ.*, 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 Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 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 Sens. Environ.*, vol. 56, no. 1, pp. 66–77, 1996.

[34] O. A. 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," *Appl. Geography*, vol. 31, no. 2, pp. 483–494, Apr. 2011.

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

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

[37] T. Hame, I. Heiler, and J. S. Miguel-Ayanz, "An unsupervised change detection and recognition system for forestry," *Int. J. Remote Sens.*, vol. 19, no. 6, pp. 1079–1099, 1998.

[38] F. Gao et al., "Mapping impervious surface expansion using medium-resolution satellite image time series: A case study," *Int. J. Remote Sens.*, 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 Sens. Environ.*, vol. 160, pp. 1–14, Apr. 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 Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.

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

[42] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 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 Geosci. Remote Sens. Lett.*, 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 Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.

[45] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[46] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, 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 Trans. Geosci. Remote Sens.*, 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 Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[49] J. Chen et al., "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 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 Sens.*, vol. 12, no. 3, p. 484, Feb. 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 Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[52] J. Huang, Q. Shen, M. Wang, and M. Yang, "Multiple attention Siamese network for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[53] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[54] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

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

[56] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *IEEE Access*, vol. 8, pp. 126385–126400, 2020.

[57] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[58] A. Gupta and P. Mannem, "From image annotation to image description," in *Proc. Int. Conf. Neural Inf. Process.* Berlin, Germany: Springer, 2012, pp. 196–204.

[59] G. Kulkarni et al., "BabyTalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[60] A. Farhadi et al., "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 15–29.

[61] V. Ordonez, G. Kulkarni, and T. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1–9.

[62] A. Frome et al., "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.

[63] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[64] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 256–270, 2020.

[65] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

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

[67] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[68] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10532–10543, Dec. 2021.

[69] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[70] X. Ma, R. Zhao, and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 2001–2005, Nov. 2021.

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

[72] S. Wu, X. Zhang, X. Wang, C. Li, and L. Jiao, "Scene attention mechanism for remote sensing image caption generation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[73] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6506605.

[74] 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 Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[75] Y. Li et al., "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608816.

[76] Y. Wang, W. Zhang, Z. Zhang, X. Gao, and X. Sun, "Multiscale multi-interaction network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2154–2165, 2022.

[77] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1985–2000, Mar. 2020.

[78] R. Chavhan, B. Banerjee, X. X. Zhu, and S. Chaudhuri, "A novel actor dual-critic model for remote sensing image captioning," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 4918–4925.

[79] G. Hoxha and F. Melgani, "A novel SVM-based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[80] Z. Chen, J. Wang, A. Ma, and Y. Zhong, "TypeFormer: Multiscale transformer with type controller for remote sensing image caption," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[81] H. Kandala, S. Saha, B. Banerjee, and X. X. Zhu, "Exploring transformer and multilabel classification for remote sensing image captioning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[82] 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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[83] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[84] B. Zhao, "A systematic survey of remote sensing image captioning," *IEEE Access*, vol. 9, pp. 154086–154111, 2021.

[85] 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 Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 7, 2022, doi: 10.1109/TPAMI.2022.3148210.

[86] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[87] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318, doi: 10.3115/1073083.1073135.

[88] C. Yew, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out (WAS)*, 2004.

[89] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl.*, New York, NY, USA, 2007, pp. 228–231.

[90] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

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

[92] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[93] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

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

[95] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[96] 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 *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 574–590.

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

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

[99] J. Lee, J.-H. Shin, and J.-S. Kim, "Interactive visualization and manipulation of attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2017, pp. 121–126.

[100] F. Stahlberg, "Neural machine translation: A review," *J. Artif. Intell. Res.*, vol. 69, pp. 343–418, Oct. 2020.

[101] X. Zhang, W. Chen, F. Wang, S. Xu, and B. Xu, "Towards compact and fast neural machine translation using a combined method," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1475–1481.

Chenyang Liu received the B.S. degree from the Image Processing Center, School of Astronautics, Beihang University, Beijing, China, in 2021, where he is currently pursuing the M.S. degree.

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, Beijing, China, in 2019 and 2022, respectively.

He is currently a Researcher with the Fuxi AI Laboratory, NetEase, Hangzhou, China. His research interests include computer vision, deep learning, and related problems in remote sensing and video games.

Hao Chen (Graduate Student Member, IEEE) received the B.S. degree from the Image Processing Center, School of Astronautics, Beihang University, Beijing, China, in 2017, where he is currently pursuing the Ph.D. degree.

His research interests include machine learning, deep learning, and semantic segmentation.

Zhengxia Zou received the B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 2013 and 2018, respectively.

From 2018 to 2021, he was a Post-Doctoral Research Fellow with the University of Michigan, Ann Arbor, MI, USA. He is currently an Associate Professor with the School of Astronautics, Beihang University. He has published more than 20 peer-reviewed papers in top-tier journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), TGRS, the IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), International Conference on Computer Vision (ICCV), and AAAI Conference on Artificial Intelligence (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 research interests include computer vision and related problems in remote sensing. 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 a 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 (TPAMI), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and the IEEE International Conference on Computer Vision (ICCV). 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/.