A Divided Spatial and Temporal Context Network for Remote Sensing Change Detection

Nian Shi, Keming Chen, and Guangyao Zhou

***Abstract*—In recent days, change detection has become one of the central tasks for remote sensing image analyses. Due to the powerful discriminative abilities, various convolutional-based approaches have been applied and shown favorable performance in change detection. However, these approaches either require numerous parameters to obtain refined features or cannot make full use of the global context information, which is crucial for change detection. Motivated by the recently proposed visual transformers, we introduce a divided spatial and temporal context modeling network to tackle such shortcomings, which is tokens-based and passes the global context by well-modeled tokens. Specifically, to model the spatial context, we first use a spatial self-attention to make each token implicitly incorporate the spatial information of the corresponding image. Then, a followed temporal self-attention is used to model the temporal context. Together with the spatial self-attention, it makes the learned tokens contain the global context and become more representational and suitable for change detection. Finally, a prediction head is used to output change detection results over the token space without additional transformer decoder or skip connections between features and tokens, thus reducing the model parameters and computational costs. Thanks to the superior global context modeling capabilities of the proposed method, we further develop a simplified variant with much smaller parameters but only a slight drop in F1 and IoU scores. Our proposed method has shown competitive performance and surpasses several state-of-the-art methods according to our experiments.**

***Index Terms*—Change detection, convolutional neural networks, self-attention, spatial-temporal transformer.**

I. Introduction

CHANGE detection is defined as the process of labeling the pixel- or region-level differences of an object or phenomenon in remote sensing images that are acquired on the same location but different times [1]. As one of the most significant tasks for remote sensing image analyses, change detection has contributed significantly to Earth observation, such as urbanization investigation [2], resource exploration [3], [4], and disaster assessment [5].

In recent years, due to the powerful discriminative abilities and strong feature representation capabilities, various convolutional-based approaches have been successfully explored and shown promising performance in change detection [6], [7]. These convolutional-based methods are designed to extract refined features while the global context, consisting of spatial context and temporal context, is not fully exploited, which is also critically important for change detection. We summarize existing challenges and corresponding research into three aspects, feature design, spatial, and temporal modeling.

First, many existing change detection methods are derived from well-performing segmentation networks, such as U-Net [8]. They mainly apply a Siamese backbone to extract features [7], [9] from multitemporal images, and adopt some mathematical operations to fuse features, such as concatenation and difference [10], often coming with attention mechanisms over channels, space, and time to improve the quality of extracted features [6]. These methods are highly feature-dependent and require a well-designed structure to obtain refined features, which introduces much more parameters. However, unlike segmentation tasks, change detection is mainly concerned with whether the pixels in input multitemporal images have changed or what kind of change has occurred, rather than which specific category they each belong to. Therefore, highly refined features may be redundant, and what we demand are the representations with more semantic relevance to changes. The encoder–decoder structure is mostly adopted in recently proposed supervised approaches to obtain advanced semantically relevant features [7], [11], but it always comes with some down-sampling operations and thus reduces the resolution of features, which is crucial in change detection [12]. To fuse the high-resolution features, the model needs to add some skip connections or complex decoders, which in turn increases the number of parameters. For more parameters mean more difficulties in practical applications, efficient and effective representations are still active demands for change detection. The global context has been proved one of the most suitable representations for change detection in recent studies [9], [13], [14]. And global context modeling is also what we are investigating in this work.

Second, although the changes refer to temporal changes, the spatial context can still greatly help to improve the results [15]. As shown in Fig. 1, the brightness, contrast, and spatial textures of the same objects are always different among input image pairs due to different imaging conditions, and thus make it difficult to compare the patches to obtain the change map directly. Although existing methods can extract unified features from different images with deeper networks and additional constraint designs, with this comes an increase in the number of parameters. While

Manuscript received January 26, 2022; revised March 23, 2022 and May 4, 2022; accepted May 13, 2022. Date of publication May 23, 2022; date of current version June 24, 2022. *(Corresponding author: Keming Chen.)*

The authors are with the Key Laboratory of Network Information System Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100194, China and with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: shinian.work@gmail.com; kmchen.ie@gmail.com; zhouguangyao@aircas.ac.cn).

Our model and code will be available at [Online]. Available: https://github.com/shinianzhihou/ChangeDetection.

Digital Object Identifier 10.1109/JSTARS.2022.3176858

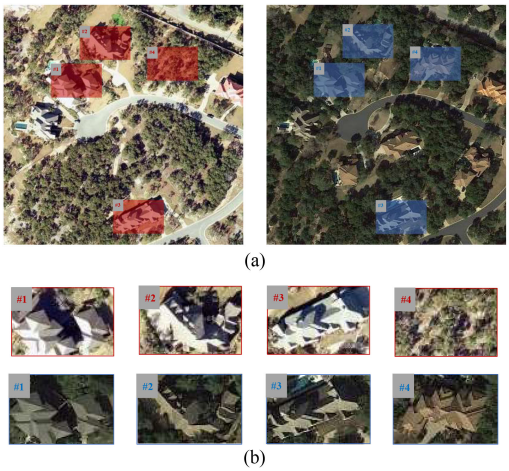

Fig. 1. Illustration of the spatial and temporal context. (a) is a paired input images used for change detection and we select four pairs of patches for illustration. These selected pairs of patches are zoomed in for better view in (b). As shown in (b), the patches with the same color in the same row mean that they are from the same image but at different positions. The patches with the same number in the same column mean that they are at the same position but from different images. For patch #1, its spatial context are patches #2, #3, and #4. Its temporal context is #1.

the patches in the same image are similar in properties and easy to compare, then the idea comes as no surprise that using the spatial context to help to model the global context. To fully leverage the spatial context, Mou *et al.* [16] tried to get more representational features by integrating the spatial information into a recurrent neural network. Sun *et al.* [15] first proved that a single patch could be approximately rebuilt by some other related patches in the same image, and then introduced the patch similarity graph matrix to obtain change maps. Chen *et al.* [13] used spatial attention to help to transfer extracted feature maps into a collection of tokens, which are vectors embedded from image patches, and generated the weighed feature maps for change detection with a transformer structure. In addition, some multilevel approaches were also designed to exploit the spatial context [9], [17]. If we can make full use of the spatial context, then we can reduce the number of parameters by eliminating much of the design effort spent on extracting comparable features. Therefore, it is pretty essential to further explore the utilization of spatial context in the change detection task.

Third, the input images for change detection are generally time-series dependent, so how to model the temporal context to achieve better results is undoubtedly essential. The most common approach is to fuse features of different images to obtain final change maps [10], [18], and there are also some attention mechanisms applied to reweight the fused temporal features/images [19]. These deep learning-based methods either directly concatenate/add/differ the obtained temporal features or perform attention mechanisms with much more parameters [7]. These methods can achieve good change detection results as they have been well trained to extract refined features. However, when the spatial context modeling is done, as we have mentioned before, the performance of these methods will be degraded since the modeled tokens cannot be compared in these ways. In addition, most methods are not easily adjusted when there are more than two images as input. Although the transformer-based methods have been introduced to change detection [13], the tokens from multitemporal images are also simply concatenated and then fed to the encoder, while the temporal context remains under-explored. Therefore, how to model temporal context for better change detection results is still a challenge worthy of further investigation.

To address the problems mentioned above, a new global context modeling network for change detection is introduce in this work. Motivated by the successful application of visual transformers, we adopt a modified transformer to model global context over the embedded token space. To further improve the representational capabilities of tokens and better model the temporal context, we separate the temporal attention from spatial-temporal attention and use two cascaded multihead self-attention (MSA) blocks to model spatial and temporal context one after the other. The spatial self-attention block will first make each token implicitly contain the information of the corresponding image. Then, the temporal self-attention block is used to model the temporal context and make the tokens more semantically relevant to changes. The visualization analysis also indicates that these well modeled tokens are representational enough for change detection. Besides, the high resolutions of features/tokens will be maintained in our proposed method since we have removed some down-sampling operations in the feature extractor. Therefore, we do not need additional transformer decoder or skip connections between tokens and features. Change detection results are directly generated over the token space in an efficient and effective manner. In addition, considering the limitations of parameters for practical applications, we propose a variant model with much smaller parameters but only a slight drop in F1 and IoU scores.

In this work, we do not pay much attention on extracting refined features but mainly explore how to model global context effectively and efficiently from multitemporal remote sensing images. As shown in Fig. 2, we first simply apply several convolutional layers as the feature extractor and embed the extracted features into tokens spatially and temporally. Then, these tokens will be fed to the modified encoder to model global context with cascaded MSA blocks. Finally, a prediction head is used to directly output change detection results from tokens. The main contributions of our work are as follows.

1) Our proposed method is mainly token-based, and no complex decoder or skip connection is applied between features and tokens. The change detection results are directly obtained from tokens, which is more effective and efficient.
2) By introducing the transformer into modeling the global context, a novel framework for change detection is proposed in this work, which can fully leverage the spatial and temporal context of multitemporal images. After the global context is modeled, the tokens will become more representational and suitable for change detection.
3) The proposed network is highly scalable and easy to make a tradeoff between accuracy and efficiency. For example,

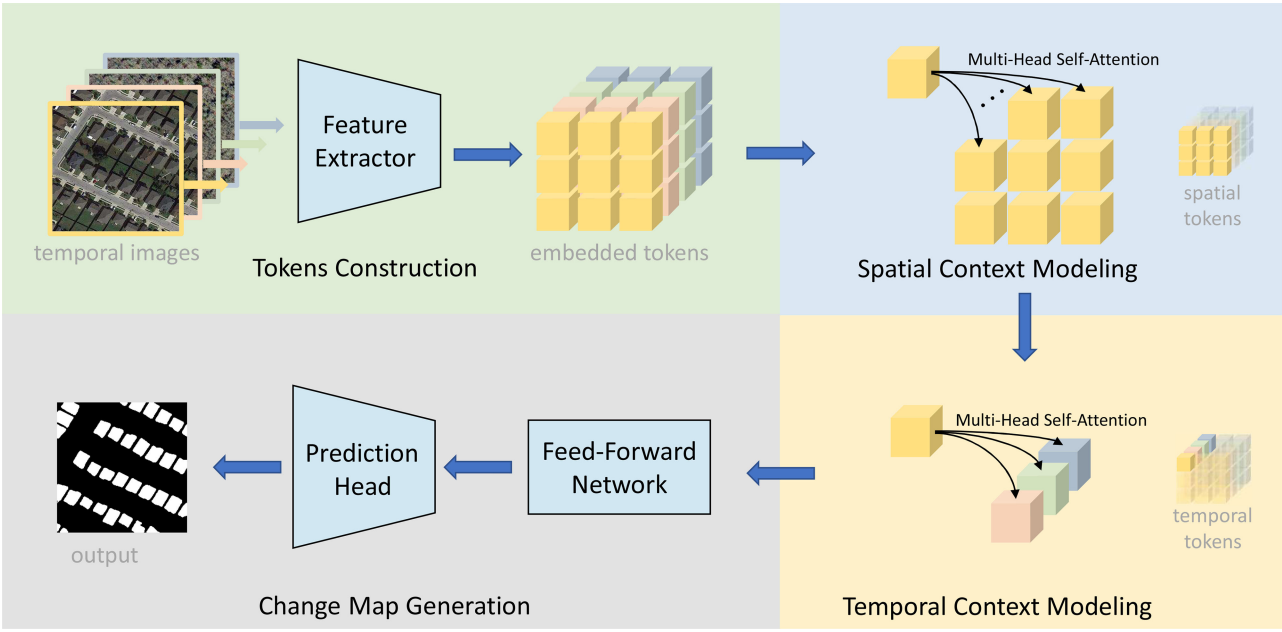

Fig. 2. Illustration of the proposed CDViT for change detection. We first use a feature extractor, consisting of several convolutional layers, to extract features $\mathbf{X}$ and then embed them into tokens $\mathbf{z}$. Here, we use the squares with different colors to represent the tokens from different images. The context modeling is mainly based on the MSA modules. The spatial context modeling is performed over the spatial tokens and the temporal context modeling is performed over the temporal tokens. The spatial MSA, temporal MSA, and a feed-forward network consist of a layer of the modified transformer encoder. Finally, a prediction head is used to output the change detection result.

after reducing the token size and encoder depths, there will be a slight drop (1.0% and 1.4%) in terms of F1 and IoU scores but a considerable decrease in parameters (20.9 M), which also demonstrates the robustness of the network. Experiments on two change detection datasets also demonstrate the efficiency and effectiveness.

The rest of this article is organized as follows. Section II describes some related works of change detection and recent visual transformer techniques. Section III introduces the proposed method in detail. The quantitative comparisons and analyses are shown in Section IV. Finally, Section V concludes this article.

II. Related Work

A. Deep Learning-Based Change Detection

Due to the extraordinary performance in classification, segmentation, and other vision tasks, deep learning technology has become a research hotspot in designing advanced change detection approaches [20], [21]. Beyond the remote sensing, there are also some deep learning-based methods, which have made a great contribution to change detection [22]–[24]. For example, Alcantarilla *et al.* [23] proposed a deconvolutional network architecture, consisting of a contraction network and a expansion network for street change detection. For remote sensing change detection, Daudt *et al.* [10] proposed to use Siamese structure for feature extraction and introduced three fully convolutional neural network into change detection. They simply early concatenated the images to the encoder or passed the concatenation/difference features to the decoder. To improve the quality of extracted features, some attention-based mechanisms are introduced to the network. For example, in each block of the deeply supervised image fusion network proposed in [9], the channel and spatial attention are first performed over the concatenated features and then a deeply supervised difference discrimination network is used to generate final change maps. Fang *et al.* [7] combined Siamese network and Nested U-Net [25] to propose a new structure for change detection. They first extracted features at different semantic levels from input paired images using a Siamese encoder, and then fused two branches of features with plenty of skip connections and convolutions. Finally, for different semantic levels of features, they introduced the ensemble channel attention module to refine features and generate better change detection results. To further enhance the feature representations, Liu *et al.* [26] used the spatial and channel attention to construct a dual attention module, and then implemented it to build the dual-task constrained network consisting of two segmentation and a change detection networks.

These purely convolutional and attention-based methods are mainly designed for obtaining more refined features and such networks tend to have more parameters, which is unsuitable to practical applications. Compared to refined features, global context modeling may be more important for change detection. It is because that the actual scene across time and space and associated changes are always complex, and the global context can significantly help us focus on the changes of interest in multitemporal images. Once the global context is well modeled, we can eliminate many extra designs on feature extraction and use the well-modeled tokens for change detection, which thus reduces the number of parameters.

The global context in change detection mainly consists of spatial and temporal context, as shown in Fig. 1. To better

model spatial context, several enhanced approaches have been employed, such as spatial attention [26], deeper backbone [6], and dilated convolution [27]. To model the temporal context, the existing methods include using channel attention to reweight features [17], [26], utilizing self-attention to explore time-related context [6], adopting nonlocal technics to exploit global relationships among pixels in space-time [6], etc. However, these methods either treated the attention mechanics as enhancing feature modules or simply used attention to update the features in different dimensions [13]. They are still struggling to explore more sophisticated features despite the high computational complexity and growing parameters.

In our proposed method, to reduce the parameters and obtain more appropriate representations, we do not focus on the features but the tokens which attend to be more semantically relevant to changes. We integrate the global context into the semantic tokens with a modified transformer encoder and make them contain high-level semantic information while not reducing the spatial resolution. This means that we do not need additional complex decoder or skip connections to update the features, and the change detection results can be directly generated from token space.

B. Self-Attention in Change Detection

Self-attention, sometimes called intraattention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence [28]. It is the core component of the widely used structure, transformer, in natural language processing (NLP) and has shown great potential for extensive use in AI applications [29]. Since it is greatly exploit in transformer [28], many variants such as bidirectional encoder representations from transformers (BERT) [30], robustly optimized BERT pretraining [31], and generative pretrained transformer v1–3 [32] have been exploited and demonstrated exemplary performance on a broad range of language tasks. Undoubtedly, such a great breakthrough in NLP has prompted many researchers' interest in different visual tasks.

Specifically, for change detection task, many self-attention-based methods have also been proposed. For example, Chen *et al.* [14] introduced the self-attention mechanism into change detection task and designed a Siamese structure with spatial-temporal attention to fully exploit the corresponding spatial-temporal relationships. Besides, the self-attention mechanism is also used to help extract more refined and discriminative features from input multitemporal images with a multiple scales structure in [14]. Furthermore, the transformer structure, which is entirely based on self-attention, has also been explored in change detection. For example, Zheng *et al.* [42] proposed a deep multitask encoder–transformer–decoder architecture (ChangeMask) for Semantic change detection, which is trained with not only the binary changed labels but also the semantically changed labels. Although the spatial and temporal context are well modeled in ChangeMask, this requires additional semantic change information as guidance, which is not available in most change detection datasets. Besides, Chen *et al.* [43] used bitemporal image transformer (BiT), a transformer-based method to model the context within the bitemporal input images. They first embedded the extracted features of different images into tokens and then used a transformer encoder to model the global context within the concatenated tokens. After that, a transformer decoder was applied to refine the extracted features using the modeled tokens [13]. Bandara *et al.* [33] proposed the ChangeFormer, which replaced the commonly used convolutional-based backbone with a hierarchical transformer encoder to render multiscale long-range details for better change detection results. Although the transformer decoder was removed in ChangeFormer, the multiscale features of bitemporal images were also simply fused with several difference modules and the temporal context was not fully exploited. However, these methods still tried to extract more sophisticated features with the self-attention mechanisms but ignored the temporal context modeling, which is very important for change detection.

In this article, we do not focus on using the self-attention to extract more refined features or modify the extracted features from input images, but mainly explore how to effectively and efficiently apply the self-attention into modeling global context to obtain more semantically relevant tokens for the change detection task.

III. Methodology

The overall architecture of the proposed method, namely CDViT, is shown in Fig. 2. First, a shallow feature extractor is used to extract features from a multitemporal image. These features will be divided into nonoverlapped patches and then embedded into tokens. After that, these tokens will be passed to two cascaded MSA modules for spatial and temporal context modeling successively, followed by a feed-forward network. Finally, these tokens will be reshaped to features and passed to a prediction head to generate the final change detection result.

In the following sections, we first describe how the tokens are embedded from input multitemporal images. Then, we give our motivation starting with the basic MSA. After that, we introduce how the spatial and temporal context are modeled. Finally, we present some details of the proposed CDViT for better reproduction.

A. Tokens Construction

The input multitemporal images $\mathbf{I}$ are expected to be of size $T \times C \times H \times W$, where T is the number of images. C, H, and W are the number of channels, height, and width of each image, respectively. As illustrated in Fig. 2, we use several shallow convolutional blocks to extract features $\mathbf{X} = \{\mathbf{X}^t | 1 \leq t \leq T\}$ from multi-temporal images where $\mathbf{X}^t \in \mathbb{R}^{C' \times H' \times W'}$. C', H', and W' are the number of channels, height, and width of feature maps, respectively. In addition, considering that the attention modules in transformer encoders are always highly computational, we reduce the channels of input features by a convolutional block with kernel size 1×1 from C' to C''. Thus, the parameters of the model can get reduced. Following the operation in ViT [34], we decompose features of each image into N nonoverlapping patches with a size of $C'' \times P \times P$ where $N = H'W'/P^2$. Then, these patches will be flattened and linearly projected into

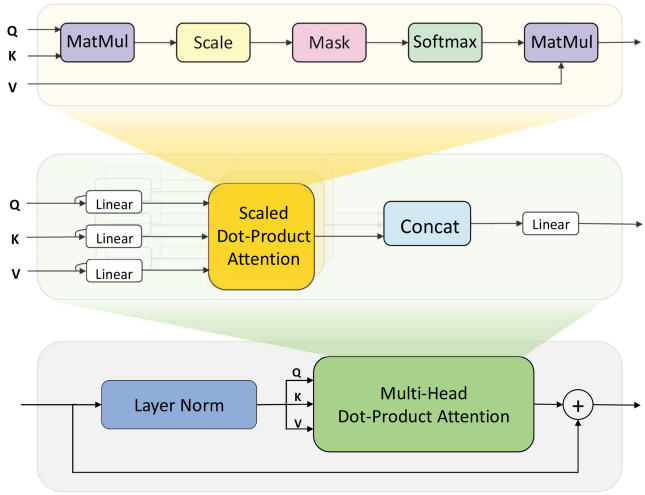

Fig. 3. Architecture of MSA.

embedding vectors $\mathbf{z} = \{\mathbf{z}_{(t,n)} | 1 \le t \le T, 1 \le n \le N\}$, where $\mathbf{z}_{(t,n)} \in \mathbb{R}^D$, and D is the length of flattened patches $\mathbf{x}_{(t,n)}$. For a flatten patch vector $\mathbf{x}_{(t,n)} \in \mathbb{R}^{C''P^2}$, the corresponding embedding $\mathbf{z}_{(t,n)}$ can be calculated as follows:

$$\mathbf{z}_{(t,n)} = \mathbf{E}\mathbf{x}_{(t,n)} + \mathbf{p}_{(t,n)} \tag{1}$$

where $\mathbf{p}_{(t,n)}$ is the learnable positional embedding and the linear projection operation $\mathbf{E} \in \mathbb{R}^{D \times C''P^2}$ is also learnable. Besides, similar to ViT, a learnable vector $\mathbf{z}_{(0,0)}$ is added to serve as the image representation [34]. After embedding input images into tokens $\mathbf{z} \in \mathbb{R}^{T \times N \times D}$, we can obtain T tokens at each position and N tokens at each time.

B. Multihead Self-Attention

MSA [28] plays a significant role in our proposed CDViT and is also an essential component in recently proposed transformer-based methods. In a transformer encoder with L encoding blocks, the output of the ℓth block can be calculated as follows:

$$\mathbf{z}'^{(\ell)} = \mathbf{z}^{(\ell-1)} + \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{z}^{(\ell-1)}\right)\right) \tag{2}$$

$$\mathbf{z}^{\ell} = \mathbf{z}'^{(\ell)} + \mathrm{FFN}\left(\mathrm{LN}\left(\mathbf{z}'^{(\ell)}\right)\right) \tag{3}$$

where LN denotes LayerNorm [35]. Feed-forward network (FFN) will be described later. Here, we start with the basic MSA and then introduce how to perform spatial and temporal context modeling step by step.

The basic architecture of an MSA module is shown in Fig. 3. Taking the embedded tokens of a single image as input, MSA enables the model to pay attention to the information at different positions. However, for temporal input images, the tokens are simultaneously in both temporal and spatial dimensions so that they cannot be directly fed to the MSA module. A simple way is to flatten the tokens $\mathbf{z} \in \mathbb{R}^{T \times N \times D}$ into $\mathbf{y} \in \mathbb{R}^{TN \times D}$ then perform spatial-temporal MSA over the flattened tokens $\mathbf{y}$. If so, at each head of the ℓth encoder block, the query, key, and value can be calculated as follows:

$$\mathbf{Q}^{(\ell)} = \mathrm{LN}\left(\mathbf{y}^{(\ell-1)}\right)\mathbf{W}_Q^{\ell} \tag{4}$$

$$\mathbf{K}^{(\ell)} = \mathrm{LN}\left(\mathbf{y}^{(\ell-1)}\right)\mathbf{W}_K^{\ell} \tag{5}$$

$$\mathbf{V}^{(\ell)} = \mathrm{LN}\left(\mathbf{y}^{(\ell-1)}\right)\mathbf{W}_V^{\ell} \tag{6}$$

where $\mathbf{W}_Q^{\ell}$, $\mathbf{W}_K^{\ell}$, and $\mathbf{W}_V^{\ell} \in \mathbb{R}^{D \times D_h}$ are the learnable parameter matrices. D_h is the dimension of each attention head and it is set to $D_h = D/N_h$, where N_h is the number of heads in MSA. However, this approach is computationally intensive and unsuitable for change detection task that mainly pay attention to temporal changes. Different from this way, in a BiT [13], the tokens from different temporal images are first concatenated then fed to a transformer encoder to model global semantic information. However, the concatenation operation cannot make full use of the temporal information of tokens and an additional transformer decoder is also needed to refine the features from different temporal images in BiT. For this reason, as shown in Fig. 2, we highlight and separate the temporal attention and perform the spatial MSA and temporal MSA one after the other in the proposed CDViT.

C. Spatial Context Modeling

As shown in Fig. 2, we use different colors to represent the tokens embedded from different temporal images. The spatial MSA will be performed over the tokens of each input image, i.e., each token will be compared with all the tokens in the same image but different positions. In the ℓth layer, the query, key, and value of each head can be generated from the tokens $\mathbf{z}_t^{(\ell-1)} \in \mathbb{R}^{N \times D}$ in the time t as follows:

$$\mathbf{Q}_t^{(\ell)} = \mathrm{LN}\left(\mathbf{z}_{(t)}^{(\ell-1)}\right)\mathbf{W}_{Q_t}^{\ell} \tag{7}$$

$$\mathbf{K}_t^{(\ell)} = \mathrm{LN}\left(\mathbf{z}_{(t)}^{(\ell-1)}\right)\mathbf{W}_{K_t}^{\ell} \tag{8}$$

$$\mathbf{V}_t^{(\ell)} = \mathrm{LN}\left(\mathbf{z}_{(t)}^{(\ell-1)}\right)\mathbf{W}_{V_t}^{\ell} \tag{9}$$

where $\mathbf{W}_{Q_t}^{\ell}$, $\mathbf{W}_{K_t}^{\ell}$, and $\mathbf{W}_{V_t}^{\ell} \in \mathbb{R}^{D \times D_h}$ are learnable parameter matrices. Then, the spatial attention $\boldsymbol{\alpha}_t^{\ell}$ for each head in spatial MSA can be calculated as follows:

$$\boldsymbol{\alpha}_t^{\ell} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_t^{(\ell)}\mathbf{K}_t^{(\ell)\top}}{\sqrt{D_h}}\right)\mathbf{V}_t^{(\ell)}. \tag{10}$$

After the spatial attention of each head are calculated, they will be first concatenated and the final output of spatial MSA will be generated with a learnable parameter matrix $\mathbf{W}_t^O$ as follows:

$$\mathrm{MSA}_t^{\ell}\left(\mathbf{Q}_t^{(\ell)}, \mathbf{K}_t^{(\ell)}, \mathbf{V}_t^{(\ell)}\right) = \mathrm{Concat}\left(\boldsymbol{\alpha}_{t,1}^{\ell}, \ldots, \boldsymbol{\alpha}_{t,h}^{\ell}\right)\mathbf{W}_t^O. \tag{11}$$

According to the spatial MSA, the token at each position will implicitly contain the information of the whole image and is more representational for making comparisons over time.

D. Temporal Context Modeling

After the tokens processed by the spatial MSA, the temporal MSA will be used to model the temporal information. Similarly, in the ℓth layer, we can get $\mathbf{Q}_n^{(\ell)}$, $\mathbf{K}_n^{(\ell)}$, and $\mathbf{V}_n^{(\ell)} \in \mathbb{R}^{D\times D}$ from the tokens $\mathbf{z}_n^{(\ell)} \in \mathbb{R}^{T\times D}$ in the position n where the tokens $\mathbf{z}_n^{(\ell)}$ are reshaped from the output of spatial MSA in the position n. Then, the temporal attention $\boldsymbol{\beta}_n^{\ell}$ for each head and the temporal MSA can be calculated as follows:

$$\boldsymbol{\beta}_n^{\ell} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_n^{(\ell)}\mathbf{K}_n^{(\ell)\top}}{\sqrt{D_h}}\right)\mathbf{V}_n^{(\ell)} \tag{12}$$

$$\mathrm{MSA}_n^{\ell}\left(\mathbf{Q}_n^{(\ell)},\mathbf{K}_n^{(\ell)},\mathbf{V}_n^{(\ell)}\right) = \mathrm{Concat}\left(\boldsymbol{\beta}_{n,1}^{\ell},\ldots,\boldsymbol{\beta}_{n,h}^{\ell}\right)\mathbf{W}_n^{O}. \tag{13}$$

The temporal MSA can fully utilize the temporal information and model the global context together with spatial MSA, making the tokens more semantic suitable for change detection. Although an MSA layer has more parameters than a convolutional layer, we can extract comparable tokens with two spatial and temporal MSA layers, whereas for convolutional-based methods, many convolutional layers are needed to achieve the same purpose. This also means that we can achieve competitive results with smaller parameters.

E. Details in Network

1) Feature Extractor: We use the first three layers of ResNet18 [36] as the feature extractor in the proposed network. Different from the original network structure of ResNet18, we remove the pooling operation in the first layer to ensure a better resolution of feature maps. In other words, when the input images are 256×256 pixels, the extracted feature maps will be 64×64 pixels after two max pooling operations. In addition, a convolutional layer with a kernel size of 1×1 is also added to reduce the number of channels, thereby reducing the parameters in the model.

2) FFN: A feed-forward network is mainly composed of two linear projection layers, between which is a GELU [37] activation layer. Taking the token $z_{(t,n)}^{\prime(\ell)}$ processed by the preceding MSA module as input, the output of FFN is

$$\mathrm{FFN}(\mathbf{z}_{(t,n)}^{\prime(\ell)}) = \mathbf{W}_2 \cdot \sigma\left(\mathbf{W}_1\mathbf{z}_{(t,n)}^{\prime(\ell)} + \mathbf{b}_1\right) + \mathbf{b}_2 \tag{14}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D'\times D}$ in the first linear layer and $\mathbf{W}_2 \in \mathbb{R}^{D\times D'}$ in the second linear layer are the learnable weights, $\mathbf{b}_1 \in \mathbb{R}^{D'}$ and $\mathbf{b}_2 \in \mathbb{R}^{D}$ are the learnable biases, and σ denotes the GELU layer. The first layer maps each token into a higher dimension D' to obtain a high-dimensional representation, and the second layer maps the tokens into the original dimension D as the input for the next encoder block.

3) Prediction Head: Although the tokens are initially generated with a local manner, after the global context is modeled with the spatial and temporal MSA, the tokens will capture global-level information and be representational enough for the change detection task according to our visualization analysis on section IV-F. Therefore, we directly use several convolutional layers to output the final change map from these tokens. Different from other methods, there are no transformer decoders or skip connections between features and tokens in our proposed CDViT, which also makes the model more efficient for change detection. Specifically, the feature maps will be first reconstructed from the modeled tokens. Then, they will be upsampled to the original image size. Finally, a two-channel probability map $\mathbf{P} \in \mathbb{R}^{2\times H\times W}$ will be output by two convolutional layers. In the inference phase, the binary change detection results are generated with a pixelwise *argmax* operation over the channel.

IV. Experiments

We conduct four experiments on two change detection datasets.

1) To validate the effectiveness of CDViT, we carry out some comparison experiments with several SOTA methods, including purely convolutional-based, attention-based, and transformer-based methods.
2) To validate the efficiency of CDViT, we conduct some analytical experiments on parameters for the naive CDViT and a simplified variant model, named as CDViT_S.
3) To investigate the effect of some model hyperparameters, we conduct some ablation studies on CDViT, including the depth of transformer encoders, size of tokens, and size of patches.
4) To explore how the spatial and temporal MSA work in CDViT, we perform visualization analyses on the features and tokens.

A. Datasets Description

We conduct our experiments on two different change detection datasets, WHU-CD and LEVIR-CD.

WHU is a remote sensing dataset including multisource imagery for multiple tasks, of which *WHU-CD* is dedicated to change detection. WHU-CD consists of a pair of very high resolution (0.2 m) aerial images. Two images are registered and the image size is 32507×15354 pixels. To fully utilize the GPU memory for training, we cut the whole image pair into smaller patches with a size of 256×256 pixels. Then, we can get a total of 7620 patches. Following the random dataset split (training/validation/test) in [13], we can get 6096/762/762 pairs of patches, respectively. Although this dataset split cannot strictly make sure the areas of training/validation/test are spatially separated, it has been retained for consistency with the results of our compared methods.

LEVIR-CD is a remote sensing dataset for building change detection with 637 pairs of very high resolution (0.5) optical images. The size of each registered image pair is 1024×1024 pixels. Similarly, to make a fair comparison with existing methods, we adopt the dataset split and patch processing in [13]. After that, we can get 7120 patches for training and 1024/2048 patches for validation/test.

Notably, since no general dataset split scheme is available for WHU-CD, we randomly repeat the split operation four times and take the mean values of different metrics as results. To ensure the reliability of the results, we repeat each experiment with four

different random seeds. We will take the mean values of four experiments for evaluation. In addition, it should be pointed that as there is no public available multitemporal dataset, we have only conducted experiments on the bi-temporal dataset.

B. Experimental Setup

1) Loss Function: We take the most commonly used binary cross entropy loss (BCELoss) as the optimization goal during the training process. The BECLoss between target $\mathbf{T}$ and $\mathbf{P}$ is calculated as follows:

$$\mathcal{L}_{\text{bce}}(\mathbf{T}, \mathbf{P}) = -\left[\mathbf{T} \cdot \log \mathbf{P} + (1 - \mathbf{T}) \cdot \log(1 - \mathbf{P})\right] \quad (15)$$

where $\mathbf{T} \in \mathbb{R}^{2\times H\times W}$ is the one-hot encoded ground-truth.

2) Implemental Details: The proposed method and aforementioned experiments are implemented with PyTorch and trained using 8 NVIDIA GTX 1080Ti. Some regular data augmentation is applied to the input temporal images, including flipping, rotating, rescaling, and bright contrast adjustment. The stochastic gradient descent with momentum is used to optimize the model during training. The values for weight decay and momentum are set to 0.0005 and 0.9. The poly scheduler is used to adjust the learning rate with a power of 0.9. We set the initial learning rate to 0.05, and it will reduce to 0.000001 after 80 000 steps of the training process. To reduce the training time, we initialize the first three layers of the feature extractor with the weights of pretrained ResNet18 on ImageNet.

3) Evaluation Metric: To quantitatively and comprehensively assess the change detection results of different methods, some metrics are adopted for evaluation in our experiments. Concerning the changed pixels, the true positive (TP) represents the number of pixels that are both included in change detection results and the reference map. The true negative (TN) denotes the number of pixels that are neither included in change detection results nor the reference map. Similarly, the false positive (FP) means the number of pixels that are only included in change detection results, and the false negative (FN) indicates the number of pixels that are only included in the reference map [38]. Then, we mainly use overall accuracy (OA), precision (Pr), recall (Re), F1, intersection over union (IoU), and kappa coefficient (KC) for evaluation. They can be calculated as follows:

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (16)$$

$$\text{Pr} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (17)$$

$$\text{Re} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (18)$$

$$\text{F1} = \frac{2\text{Pr} \cdot \text{Re}}{\text{Pr} + \text{Re}} \quad (19)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (20)$$

$$\text{KC} = \frac{\text{OA} - \text{PRE}}{1 - \text{PRE}} \quad (21)$$

where PRE denotes the proportion of expected agreement between the ground-truth and predictions with the given class distributions and it is calculated by $\text{PRE} = ((\text{FP} + \text{FP})(\text{FP} + \text{FN}) + (\text{FN} + \text{TN})(\text{FP} + \text{TN}))/(\text{FP} + \text{TN} + \text{FP} + \text{FN})^2$ [39].

C. Compared With SOTA Methods

We compare the proposed CDViT with eight SOTA methods, and these methods are summarized as follows.

1) The fully convolutional early fusion (*FC-EF*) network [10] is a purely convolutional-based method. It first concatenates input images and then feeds them to the fully convolutional network.
2) The fully convolutional Siamese difference (*FC-Siam-Di*) network [10] is a purely convolutional-based method. It first extracts features from input images with a Siamese structure and then uses the feature difference as the input of a decoder to get the change detection results.
3) The fully convolutional Siamese concatenation (*FC-Siam-Conc*) network [10] is a purely convolutional-based method. Different from FC-Siam-Di, FC-Siam-Conc passes concatenated, rather than subtracted, features to the decoder.
4) The dual-task constrained deep Siamese convolutional network (*DTCDSCN*) [26] is an attention-based method. To obtain more representative features, it performs spatial and channel attention over the extracted features.
5) The spatial-temporal attention-based neural network (*STANet*) [14] is an attention-based method. It adopts the self-attention module to model the spatial-temporal relationships among input images to obtain more representative features.
6) The deeply supervised image fusion network (IFNet) [9] is an attention-based method. It introduces an additional difference discrimination network with a multilevel structure. At each level, it first makes a difference between paired features, and then performs spatial and channel attention to refine the features according to the labels.
7) The combination of Siamese network and Nested U-Net (*SNUNet*) [7] is an attention-based method. For the features at different semantic levels, it first uses the ensemble channel attention to refine them and then takes the concatenated features for change detection.
8) The BiT [13] is a transformer-based method. First, it projects the bitemporal images into some semantic tokens and uses a transformer encoder to model the context within them. Then, a transformer decoder is used to refine original features, followed by a prediction head to output final change detection results. In addition, we give the parameters and scores of its simplified variant BiT_S3 for a comprehensive analysis and comparison.
9) *ChangeFormer* [13] is a transformer-based Siamese network. It unifies hierarchically structured transformer encoder with multilayer perception decoder in a Siamese network architecture to efficiently render multiscale long-range details required for accurate CD [33].

Among these methods, the FC series, including FC-EF, FC-Siam-Di, and FC-Siam-Conc, are purely convolution-based

TABLE I
CHANGE DETECTION RESULTS OF DIFFERENT METHODS ON TWO DATASETS WITH ALL SCORES EXPRESSED IN PERCENTAGE

Methods	WHU-CD						LEVIR-CD					
	Pr	Re	F1	IoU	OA	KC	Pr	Re	F1	IoU	OA	KC
FC-EF [10]	71.63	67.25	69.37	53.11	97.61	68.12	86.91	80.17	83.40	71.53	98.39	82.56
FC-Siam-Di [10]	47.33	77.66	58.81	41.66	95.63	56.65	89.53	83.31	86.31	75.92	98.67	85.61
FC-Siam-Conc [10]	60.88	73.58	66.63	49.95	97.04	65.09	91.99	76.77	83.69	71.96	98.49	82.91
DTCDSCN [26]	63.92	82.30	71.95	56.19	97.42	70.78	88.53	86.83	87.67	78.05	98.77	87.02
STANet [14]	79.37	85.50	82.32	69.95	98.52	81.55	83.81	**91.00**	87.26	77.40	98.66	86.55
IFNet [9]	**96.91**	73.19	83.40	71.52	98.83	82.80	**94.02**	82.93	88.13	78.77	98.87	87.53
SNUNet [7]	85.60	81.49	83.50	71.67	98.71	82.82	89.18	87.17	88.16	78.83	98.82	87.54
BiT [13]	86.64	81.48	83.98	72.39	98.75	83.33	89.24	89.37	89.31	80.86	98.92	88.81
ChangeFormer [33]	-	-	-	-	-	-	92.05	88.80	90.40	82.48	99.04	89.89
CDViT	94.97	**89.87**	**92.35**	**85.80**	**99.37**	**92.02**	92.43	89.75	**91.07**	**83.61**	**99.10**	**90.60**

Note: Highest scores are marked in bold.

methods. DTCDSCN, STANet, and IFNet are attention-based methods. BiT is a recently proposed transformer-based method. Note that we directly use the experimental results in [13] to make a comparison.

1) Results on WHU-CD: Table I presents comparison results of these mentioned methods on WHU-CD. In terms of Pr scores, the IFNet gets the highest score of 96.91%, followed by the CDViT with 94.97%. However, the Re score of IFNet is much lower than other methods, which, thus, decreases the F1 score. For F1 scores, the CDViT achieves the highest score of 92.35%, that is, 8.37% higher than the followed BiT. Besides, the CDViT also obtains the highest IoU score of 85.80% and OA score of 99.37%. Moreover, for IoU scores, CDViT can get 13.41% higher than other methods. In addition, CDViT achieved a KC score of 92.02% on WHU-CD, which is a fairly high score compared to other methods.

2) Results on LEVIR-CD: The change detection results of different methods on LEVIR-CD are shown in the right side of Table I. According to the results in Table I, the CDViT achieves 83.61% and 99.10% in terms of IoU and OA scores, respectively, which is 2.75% and 0.18% higher than compared methods. For Pr scores, IFNet also achieves the highest score of 94.02%, which is 1.59% higher than the followed CDViT. The highest Re score of 91.00% is obtained by the STANet, followed by the CDViT with 89.75%. However, in terms of F1 scores, CDViT can obtain the highest F1 value, which is 1.76% higher than compared methods. In addition, CDViT gets 90.6% KC scores on LEVIR-CD.

3) Experimental Analysis: Similar to the transformer-based BiT [13], there are no complicated structures in our proposed method, and only several convolutional blocks are used to extract features from multitemporal images. In addition, there are no skip connections or transformer decoders added between features and tokens in CDViT. The compared results in Table I demonstrate the effectiveness of the proposed CDViT, and we attribute it to the better spatial and temporal modeling capabilities of global context.

We also display part of the visualization change detection results for the proposed method on two datasets in Fig. 4, including small [Fig. 4(a) and (d)], medium [Fig. 4(b) and (e)], and large [Fig. 4(c) and (f)] objects. For a better view, We use red, blue, white, and black to denote FP, FN, TP, and TN, respectively. As shown in Fig. 4(a) and (d), although there are two pooling operations used in CDViT, the changes of small objects can still be detected. We infer that this is because the feature extractor has sufficiently learned the features of the possible changed objects so that the corresponding changes can be detected even though the objects are small. For large objects, due to the global context being better modeled by the spatial and temporal MSA, the integrity can thus be ensured in detection results, as shown in Fig. 4(c) and (f). However, there are still some limitations that need to be improved, such as blurred boundaries, as shown in Fig. 5. This problem also exists in some other transformer-based methods [13], and we infer that it may be caused by the loss of spatial details during tokens embedding and features reconstructing, which will be included in our future research. Besides, for our proposed CDViT, this may also be due to the reduction of spatial size when extracting features from images.

D. Analysis on Parameters

To verify the efficiency of CDViT, with the same or even fewer parameters, we propose a variant of CDViT, that is, CDViT_S. For CDViT_S, the number of encoding blocks and the size of token are reduced to 1 and 128. Similarly, the sizes of two convolutional layers in the prediction head are reduced from $64 \times 32 \times 3 \times 3$ and $32 \times 2 \times 3 \times 3$ to $16 \times 16 \times 3 \times 3$ and $16 \times 2 \times 3 \times 3$. Table II gives the change detection results of CDViT_S and CDViT on WHU-CD and LEVIR-CD datasets. Due to the simplification of the model, we can find that the number of parameters of CDViT_S gets greatly reduced from 21.98 M to 1.08 M compared to CDViT. Only a slight drop in F1 scores (i.e., 0.94% and 1.02%) and IoU scores (i.e., 1.65% and 1.18%) are caused in change detection results on two datasets. Only with one encoder and a few convolutional layers, 91.41%/90.05% F1 scores and 84.15%/82.43% IoU scores can be obtained, indicating that the proposed global context modeling method is highly competent and well suited for the change detection task.

Besides, we compare the proposed CDViT with some attention-based methods and a SOTA transformer-based method

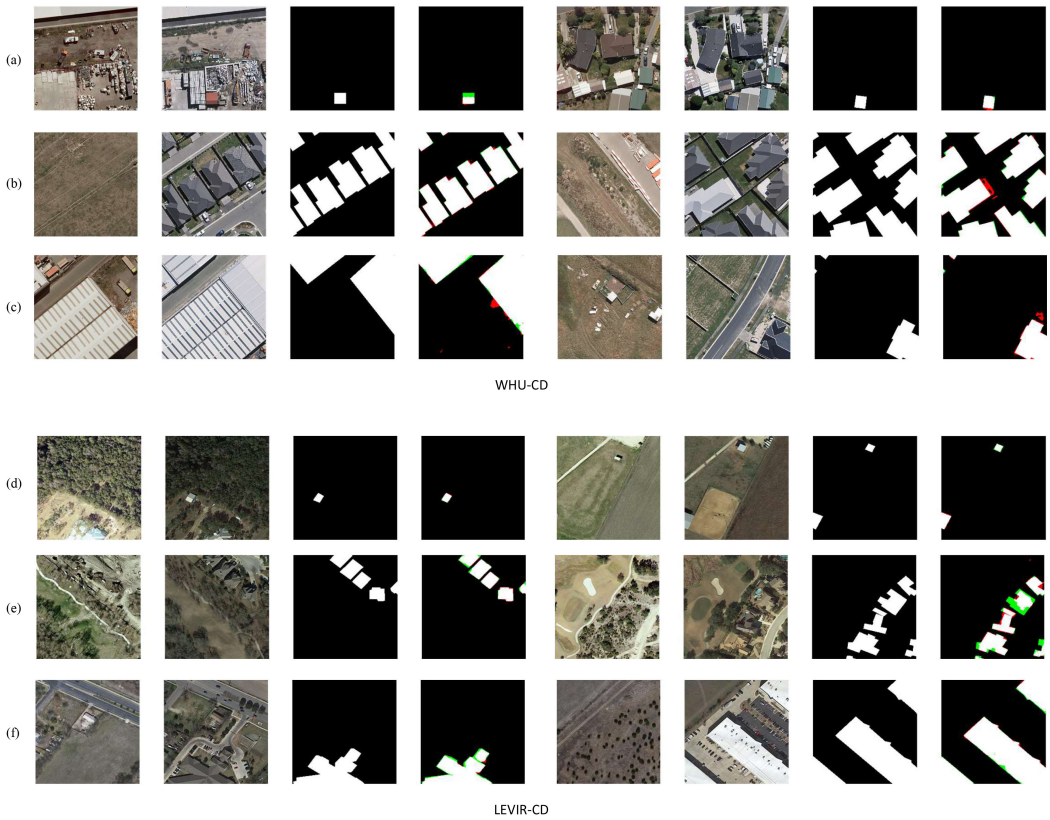

Fig. 4. Results on WHU-CD and LEVIR-CD datasets of the proposed method. Each sample consists of four images, image 1, image 2, ground-truth, and detection results. (a) and (d) are the results for small objects. (b) and (e) are for medium objects. (c) and (f) are for large objects. We use red to represent false-positive pixels and green for false-negative pixels.

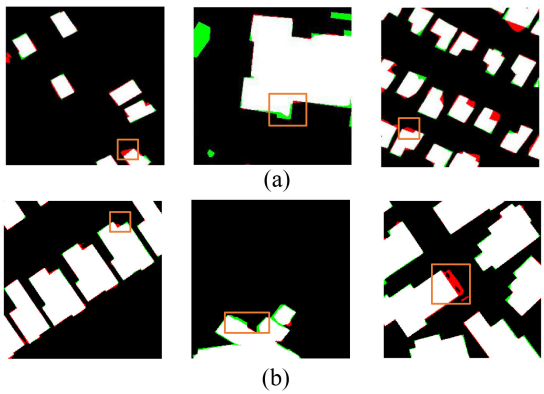

Fig. 5. Illustration of the blurred boundaries. We use yellow boxes to mark part of the blurred boundaries. (a) Some samples of BiT. (b) Some samples of CDViT.

TABLE II
PARAMETRIC ANALYSIS OF DIFFERENT MODELS

Methods	Params (M)	WHU-CD		LEVIR-CD	
		F1	IoU	F1	IoU
DTCDSCN	41.07	71.95	56.19	87.67	78.05
STANet	16.93	82.32	69.95	87.26	77.40
IFNet	50.71	83.40	71.52	88.13	78.77
SNUNet	12.03	83.50	71.67	88.16	78.83
BIT_S3	1.45	81.38	68.60	88.51	79.39
BIT	3.55	83.98	72.39	89.31	80.68
CDViT_S	**1.08**	91.41	84.15	90.05	82.43
CDViT	21.98	**92.35**	**85.80**	**91.07**	**83.61**

Note: The number of parameters (Params), F1, and IoU scores of different models on two change detection dataset are reported. The best performance are marked in bold.

(BIT_S3 and BIT). Attention-based methods mainly pay attention to features and require a complex design on convolutional layers to extract more refined features. Although this allows attention-based models to obtain better features and thus improve change detection results, the number of parameters gets heavily increased with many features possibly being redundant. For transformer-based methods, it is mainly the self-attention among patches and the modeled contextual information of temporal images that greatly contribute to detecting changes. Once the global context are well modeled, the transformer-based methods can detect changes with much smaller parameters. As shown in Table II, our proposed CDViT_S outperforms the compared methods both in F1/IoU scores and parameters on two change detection datasets. Compared with attention-based methods, CDViT_S can get higher F1/IoU scores with much smaller (i.e., 10–50 times) parameters. For example, CDViT_S achieves 7.91%/12.48% higher F1/IoU scores with ten times

TABLE III
ABLATION STUDY ON DIFFERENT DEPTH OF TRANSFORMER ENCODER WITH CDVIT ON WHU-CD

encoder depth	F1	IoU	OA	KC
1	91.56	84.42	99.30	91.20
2	91.81	84.86	99.32	91.16
4	**92.35**	**85.80**	**99.37**	**92.02**
8	91.12	83.65	99.26	90.73

Note: All scores are represented in percentage.
The best performance are marked in bold.

TABLE IV
ABLATION STUDY ON DIFFERENT TOKEN SIZES OF CDVIT ON LEVIR-CD

token size	F1	IoU	OA	KC
128	90.83	83.22	99.09	90.35
256	90.98	83.45	99.09	90.50
512	91.07	83.61	99.10	90.60
768	**91.13**	**83.71**	**99.11**	**90.66**

Note: All scores are represented in percentage.
The best performance are marked in bold.

TABLE V
ABLATION STUDY ON DIFFERENT PATCH SIZES OF CDVIT ON LEVIR-CD

patch size	F1	IoU	OA	KC
2	91.01	83.50	99.09	90.53
4	**91.07**	**83.61**	**99.10**	**90.60**
8	90.95	83.40	99.08	90.47
16	90.57	82.76	99.05	90.07

Note: All scores are represented in percentage.
The best performance are marked in bold.

smaller parameters than SNUNet. Compared with the efficient transformer-based method BIT_S3, CDViT_S can obtain better change detection results in terms of F1/IoU scores with smaller parameters (i.e., 1.08 M), which further demonstrates the effectiveness of the proposed context modeling approach. Moreover, in practice, we can easily make a tradeoff between accuracy and efficiency by adjusting the token size and encoder depth, which is critically important for practical applications.

E. Ablation Study

1) Depth of Transformer Encoder: Fixing the dimension of the encoder to 512, we perform an ablation over the depth of the encoder by 1/2/4/8. Results of CDViT on WHU-CD are shown in Table III. As the depth of the transformer encoder increases from 1 to 4, F1/IoU/KC scores get 0.79%/1.38%/0.82% improvement. However, with this comes a noticeable increase in the number of parameters of the model, from 6.23 to 42.99 M. As shown in Table III, we can see that despite the encoder depth of 1, the model can learn the relations between temporal and spatial patches, and model the global context for change detection. It is also consistent with the results observed in BiT [13]. In addition, we also see that the scores decrease when the depth of the encoder increases from 4 to 8. As proved in [34] and [40], with the increase of depth and parameters, it will become harder to train the transformer-based model from scratch. Therefore, we infer that the decrease in scores is caused by the transformer encoder not being pretrained on large-scale remote sensing data.

2) Token Size: When the height and width of the patch are determined, as the size of input tokens increases, the tokens will become more representational [28], [34]. For this reason, we conduct an ablation experiment on the size of tokens. We test different token sizes of 128, 256, 512, and 768 on LEVIR-CD dataset for the transformer encoder. The comparison results are presented in Table IV. When the token size is set to 128, we can still get a better change detection result of 90.93% F1 scores and 82.22% IoU scores. Notably, the proposed model does not have a skip connection operation between the feature maps and any outputs of transformer encoders. Therefore, this indicates that the tokens can characterize the corresponding patch regions at a specific space and time only with a token size of 128. When the token size increases from 128 to 768, the change detection results of the model are also gradually improved, and the F1 score and IoU score can achieve 91.13% and 83.71%, respectively. It also illustrates that longer tokens can bring up stronger representational abilities and better change detection results. However, considering the efficiency of the model, the token size is usually set to 512, which is consistent with the traditional transformer [28].

3) Patch Size: For patch-based change detection methods [15], [41], patch size generally determines the computational time and performance of models. Fixing the depth of the transformer and the token size to 4 and 512, we perform an ablation study over the patch size of input images on LEVIR-CD dataset. Notably, since the size of input images is determined, the number of patches will decrease with the patch size increasing. For example, when the patch size is reduced by two times, the number of patches will increase four times. We set the patch size, width, and height of the patch to 2, 4, 8, and 16, respectively, and the corresponding results are illustrated in Table V. We can observe that as the patch size decreases, all the metrics get slightly increased. For example, when the patch size decreases from 16 to 4, the IoU scores increases from 82.76% to 83.61%. As the patch size decreases, the same token will represent a smaller region, and a more fine-grained representation will be fed to the encoder to get better change detection results. However, when the patch size continually decreases from 4 to 2, the performance of the model gets slightly decreased, e.g., the IoU scores decrease from 83.61% to 83.50%. It may be because that using a token with size 512 to represent a patch with size 2×2 will contain too much redundant information and thus make it hard to train a transformer-based model. Finally, we set the patch size to 4 for our proposed CDViT.

F. Visual Features in CDViT

To better explain how features are extracted and how the attention is concentrated on the changed objects with the proposed CDViT, we give a sample visualization of features maps and

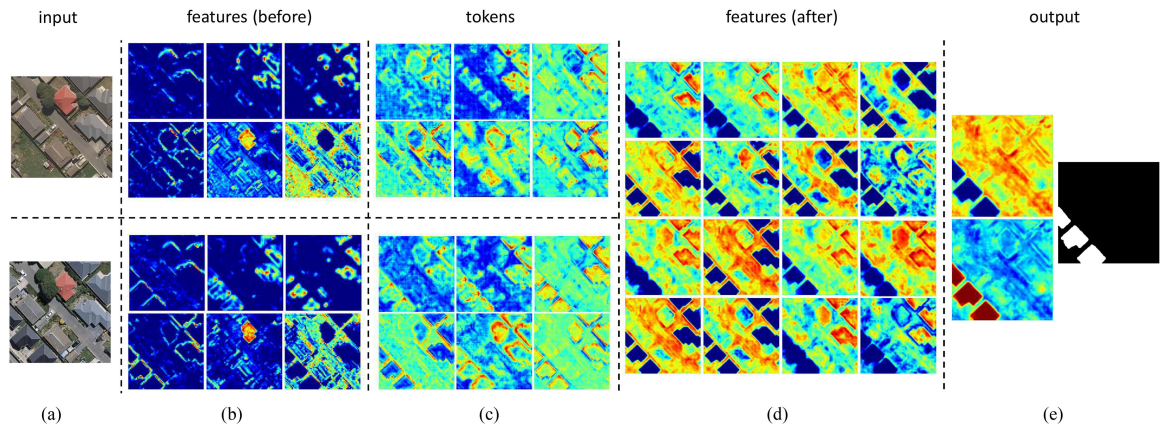

Fig. 6. Visualization analysis of the feature maps and tokens. Pixels with red represent higher values while blue for lower values. (a) Pair of images in WHU-CD. (b) Feature maps extracted by the first feature extractor (before global context modeling). (c) Visualization of reconstructed tokens. (d) Feature maps decoded by the prediction head (after global context modeling). (e) Final two-channel probability map and the binary results which are only generated during the inference.

reconstructed tokens throughout the model in Fig. 6. It should be pointed that we just sample those features and tokens with more significant responses to kernels, i.e., larger summation values for presentation. As shown in Fig. 6(b), the extracted feature maps correspond to some basic features, such as edges and colors. After tokens embedding and global context modeling, the tokens in Fig. 6(c) become more representational and semantically informative. They become concentrated in the specific objects, such as roads, buildings, and lands. Finally, after feature reconstruction and prediction head, the features in Fig. 6(d) are mainly concentrated in the changed buildings. Moreover, the inference phase can produce binary change detection results with the argmax operation over the two-channel probability maps. The visualization results show that the proposed method can effectively make the model focus on specific objects with spatial and temporal context rather than heavy convolutional layers to extract refined features. It is also the reason why the model can achieve superior performance but with fewer parameters.

V. Conclusion

In this article, we introduce the transformer to the change detection task and propose an effective and efficient transformer-based method. CDViT does not require highly refined features but only uses several convolutional layers as a feature extractor. The change detection results are directly obtained over the token space with a prediction head, and there are no additional complex decoders or skip connections needed between features and tokens, which can reduce the parameters and make the model more efficient. We adopt a divided spatial and temporal transformer encoder to model the global context. We first use the spatial MSA to make the tokens implicitly contain the corresponding image information, then incorporate temporal information into tokens via the temporal MSA to make tokens more representational and suitable for change detection. Considering the limitations for practical applications, we also introduce a simplified variant CDViT_S only with 1.08 M parameters. Due to the superior global context modeling capabilities, there is only a slight drop in F1 scores (i.e., 0.94% and 1.02%) and IoU scores (i.e., 1.65% and 1.18%) compared with CDViT, despite the significant decrease (i.e., 20.9 M) in the number of parameters. According to our experiments on two datasets, the CDViT and CDViT_S both outperform the recently proposed attention- and transformer-based methods in terms of efficiency and effectiveness. As the excellent performance of the transformer in change detection, in future work, we will along CDViT to explore a pure transformer approach for change detection, i.e., removing all the convolutional layers.

References

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

[2] R. Manonmani and G. Suganya, "Remote sensing and GIS application in change detection study in urban zone using multi temporal satellite," *Int. J. Geomatics Geosciences*, vol. 1, no. 1, pp. 60–65, 2010.

[3] R. S. Lunetta, J. F. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy, "Land-cover change detection using multi-temporal MODIS NDVI data," *Remote Sens. Environ.*, vol. 105, no. 2, pp. 142–154, Nov. 2006.

[4] Z. Zhu and C. E. Woodcock, "Continuous change detection and classification of land cover using all available landsat data," *Remote Sens. Environ.*, vol. 144, pp. 152–171, 2014.

[5] L. Gueguen and R. Hamid, "Toward a generalizable image representation for large-scale change detection: Application to generic damage analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3378–3387, Jun. 2016.

[6] 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, 2021.

[7] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8007805.

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

[9] C. Zhang *et al.*, "A deeply supervised image fusion network for change detection in high resolution Bi-temporal remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 183–200, 2020.

[10] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.
[11] N. Shi, K. Chen, G. Zhou, and X. Sun, "A feature space constraint-based method for change detection in heterogeneous images," *Remote Sens.*, vol. 12, no. 18, 2020, Art. no. 3057.
[12] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.
[13] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.
[14] 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, 2020, Art. no. 1662.
[15] Y. Sun, L. Lei, X. Li, X. Tan, and G. Kuang, "Patch similarity graph matrix-based unsupervised remote sensing change detection with homogeneous and heterogeneous sensors," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4841–4861, Jun. 2021.
[16] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.
[17] 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, 2020, Art. no. 484.
[18] 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.
[19] F. I. Diakogiannis, F. Waldner, and P. Caccetta, "Looking for change? roll the dice and demand attention," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3707, doi: 10.3390/rs13183707.
[20] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A conditional adversarial network for change detection in heterogeneous images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 45–49, Jan. 2019.
[21] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.
[22] K. Sakurada and T. Okatani, "Change detection from a street image pair using CNN features and superpixel segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2015, vol. 61, pp. 1–12.
[23] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Auton. Robots*, vol. 42, no. 7, pp. 1301–1322, 2018.
[24] E. Guo *et al.*, "Learning to measure change: Fully convolutional siamese metric networks for scene change detection," 2018, *arXiv:1810.09111*.
[25] K. Li, Z. Li, and S. Fang, "Siamese nestedunet networks for change detection of high resolution satellite image," in *Proc. Int. Conf. Control, Robot. Intell. Syst.*, 2020, pp. 42–48.
[26] 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 2020.
[27] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.
[28] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[29] K. Han *et al.*, "A survey on visual transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2022.3152247.
[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
[31] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.
[32] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[33] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," 2022, *arXiv:2201.01293*.
[34] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.
[35] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[37] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," 2016.
[38] M. Gong, Y. Yang, T. Zhan, X. Niu, and S. Li, "A generative discriminatory classified network for change detection in multispectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 321–333, Jan. 2019.
[39] A. M. El Amin, Q. Liu, and Y. Wang, "Zoom out CNNs features for optical remote sensing change detection," in *Proc. 2nd Int. Conf. Image, Vis. Comput.*, 2017, pp. 812–817.
[40] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 813–824.
[41] T. Bao, C. Fu, T. Fang, and H. Huo, "PPCNet: A combined patch-level and pixel-level end-to-end deep network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1797–1801, Oct. 2020.
[42] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 228–239, 2022.
[43] 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.

Nian Shi received the B.S. degree in electronic engineering from the Beijing Institute of Technology, Beijing, China, in 2020. He is currently working toward the M.S. degree in pattern recognition and intelligent system with the Institute of Electronics, Chinese Academy of Sciences, Beijing.

His research interests include remote sensing image processing and pattern recognition.

Keming Chen received the M.S. degree in automatic control from the Wuhan University of Technology, Wuhan, China, in 2006 and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently an Associate Professor with the Institute of Electronics, Chinese Academy of Sciences. His current research interests include remote sensing image processing and pattern recognition.

Guangyao Zhou received the M.S. degree in pattern recognition and intelligent system from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2013.

He is currently a Senior Engineer with the Institute of Electronics, Chinese Academy of Sciences. His current research interests include remote sensing image processing and pattern recognition.