Context and Difference Enhancement Network for Change Detection

Dawei Song, Yongsheng Dong, Senior Member, IEEE, and Xuelong Li, Fellow, IEEE

Abstract—At present, convolution neural networks have achieved good performance in remote sensing image change detection. However, due to the locality of convolution, these methods are difficult to capture the global context relationships among different-level features. To alleviate this issue, we propose a context and difference enhancement network (CDENet) for change detection, which can strongly model global context relationships and enhance the change difference. Specifically, our backbone is the dual TransUNet, which is based on U-Net and equipped with transformer block in the encoder. The dual TransUNet is used to extract bitemporal features. Then, the features are encoded as the input sequence, which is conducive to modeling the global context. Moreover, we design the content difference enhancement module to process the dual features of each layer in the encoder. The designed module can increase the spatial attention of difference regions to enhance the change difference features. In the decoder, we adopt a simple cross-layer feature fusion to combine the up-sampled features with the high-resolution features, which is used to generate more accurate results. Finally, we adopt a novel loss to supervise the accuracy of results in regions and pixels. The experiments on two public change detection datasets demonstrate that our CDENet has strong competitiveness and performs better than the state-of-the-art methods.

Index Terms—Change detection (CD), content difference enhancement, global context, remote sensing.

I. INTRODUCTION

Change detection (CD) is a technique for identifying land surface changes in remote sensing images. This technique is mainly to observe two images in different periods of the same region and extract the differences between them. Essentially, the CD is a binary classification task, which emphasizes the region whether changed. The value of the pixels in the CD result is 1 or 0, indicating whether the corresponding position has changed or not. It is very important for CD to quickly and automatically extract accurate change information from massive remote sensing data. Due to the importance and effectiveness of CD tasks, the CD has been widely used in land use [1], [2], environmental monitoring [3], [4], damage assessment [5], [6], and many other tasks.

Traditional CD methods can be divided into the image algebra method [7], [8] and the transformation-based methods [9], [10]. Image algebra methods include image difference, image quantification, and so on. They usually compare the pixel values between the bitemporal images to generate difference maps. The transformation-based methods include principal component analysis (PCA) and histogram trend similarity. From the available transformation change information, they can distinguish the changed region. In recent years, deep learning methods, especially convolutional neural networks (CNNs), have been widely used in various computer vision (CV) tasks, such as image segmentation [11], [12], [13], scene classification [14], [15], [16], object detection [17], [18], saliency detection [19], [20], image enhancement [21], [22], group detection [23], and so on. Because the CNNs can learn the multi-level features and semantic features of the bitemporal images, they are also introduced into CD to effectively describe the change information [24], [25], [26]. However, due to the locality of traditional convolution, these methods are difficult to capture the global semantic interaction and the context relationships among different-level features. Then, the new CNN-based methods [27], [28], [29] are equipped with dilated convolutions [30], [31] or attention mechanisms [32], [33] to alleviate this problem. They still have bottlenecks in extracting global information. Recently, the transformer [34] has been introduced into the CV fields. Due to the fantastic global context modeling ability, the transformer-based methods constantly refresh the best performance in various CV tasks [35], [36]. The transformer-based methods are also gradually applied to remote sensing image processing.

The existing methods have achieved great results in CD tasks, but these methods still have limitations on extracting global information. And the change features extracted by the existing methods contain much disturbing noise. Our motivation is to sufficiently capture the context relationships from different level of bitemporal features and to enhance the change difference features by increasing the spatial attention to the difference regions. We propose a context and difference enhancement network (CDENet) for CD, which is based on U-Net and equipped with transformer block in the encoder. The proposed CDENet...
can model the global context and strongly enhance the change difference. The flow chart of our CDENet is shown in Fig. 1. Specifically, the dual-stream structure can comprehensively extract the features of bitemporal images. The transformer block is equipped in the encoder, which is used to model the context relationships and enhance the ability of structure representation. Then, we design the content difference enhancement module (CDEM), which is used to extract the change difference features between bitemporal images and can enhance the content information in difference features. Besides, the simple cross-layer feature fusion can transfer the details of the difference features from the encoder layer to the corresponding decoder layer. Finally, we adopt a novel loss to supervise the results in regions and pixels. The proposed loss can improve the detection performance.

Our contributions are summarized as follows.

1) We propose CDENet for CD, which can strongly model global context relationships and enhance the change difference. The CDENet combines the ability of U-Net to extract spatial details and the ability of the transformer to model the context relationships. Besides, to fit the CD task, we specifically integrate the transformer block in the dual-stream network structure, which can capture the context relationships from the bitemporal features.

2) We design CDEM to process the dual features of each layer. The CDEM enhances the change difference features by increasing the spatial attention to the difference regions. It can preliminarily filter the noise and accurately locate the change regions.

3) We adopt a novel loss, which combines pixel- and region-level supervision. Therefore, it can ensure the accuracy of change results in regions and pixels.

4) We compare our CDENet with nine state-of-the-art (SOTA) methods on two public CD datasets. Visual and quantitative results demonstrate that our CDENet performs better. Besides, ablation experiments present the effectiveness of each module.

The rest of this article is organized as follows. Section II describes the related CD works. Section III introduces our CDENet. Section IV presents and analyzes the experimental results. Finally, Section V concludes this article.

II. RELATED WORK

In this section, we introduce the related work that mainly includes the traditional CD methods, CNN-based CD methods, and transformer-based methods in CV.

A. Traditional CD Methods

In the traditional CD methods, some simple algebraic methods, such as image differencing [37] and image regression [38], were applied to the bitemporal images. They detected changes by obtaining differences between the images as feature differences through threshold selection. Bruzzone and Prieto [39] proposed automatic techniques to obtain the statistical distributions of the pixel changes. Deng et al. [9] adopted the PCA, which used the transformation method to distinguish the effective difference information. Krinidis and Chatzis [40] modified the fuzzy C-means algorithm for image clustering. Zhang et al. [41] used spectral clustering and difference images, which clustered the bitemporal images into two clusters to get the change maps. Li et al. [42] proposed the locality adaptive discriminant analysis to learn a representative subspace of the data. Although these methods performed well, there were still many shortages.

B. CNN-Based CD Methods

Benefiting from the comprehensive feature extraction ability and strong expression of CNN, many CNN-based CD methods have achieved encouraging results. Daudt et al. [25] proposed end-to-end training CD methods, including FC-EF, FC-Siam-conc, and FC-Siam-diff, which were based on the fully convolutional network (FCN) with shared weights. To model the spatial–temporal relationships, Chen and Shi [43] proposed a novel Siamese-based spatial–temporal attention network. To improve the change information extraction of bitemporal image pairs, Zhang et al. [44] presented a deeply supervised image fusion network (DSIFN). Chen et al. [45] presented the dual attentive Siamese networks, which can extract long-distance dependence to represent more discriminative features. Li and Huo [46] improved the FCN and presented FC-Siam-diff-pyramid attention (PA), which added the PA layer in the FC-Siam-diff. Based on PCANet and saliency detection, Li et al. [47] proposed the SDPCANet. The SDPCANet improved the antinoise performance and detection precision. To distinguish object-level features and obtain complete CD maps, Liu et al. [48] used dual-attention to capture the relationships between channels and spatial positions. To sufficiently utilize the difference information, Lei et al. [49] proposed a difference enhancement and spatial–spectral nonlocal network in which their difference enhancement module was mainly based on channel attention. To maintain high-resolution and delicate representation, Fang et al. [50] aggregated and refined multi-level features. To alleviate pseudo-changes and noise in CD, Shi et al. [51] presented a deeply supervised attention metric-based network (DSAMNet). Besides, they created a new dataset (i.e., SYSU-CD) to overcome limitations in CD datasets.
To better utilize the spatial and temporal semantic correlations, Ding et al. [52] proposed the bitemporal semantic reasoning network.

C. Transformer-Based Methods in CV

The transformer was originally proposed for machine translation [34] and was widely applied in natural language processing. It can establish global dependencies in the sequence-to-sequence tasks. There were many attempts to use transformers in CV tasks, which was also a trend now. Dosovitskiy et al. [35] first applied the pure transformer for CV and proposed the ViT to image classification, which took the 2-D image patches with position embedding as the input. With the efforts of many researchers, the transformer-based methods performed well as the CNN-based methods. Yuan et al. [36] presented the tokens-to-token vision transformer to progressively structurize the image. Thus, surrounding tokens can be modeled and token length can be reduced. Chen et al. [53] proposed the TransUNet for medical image segmentation, which merits both transformers and U-Net. Fang et al. [54] presented the external attention-based TransUNet for crack detection. In addition, transformer-based methods were also used in remote sensing images. He et al. [55] embedded the Swin transformer into the U-Net for remote sensing semantic segmentation. Therefore, it can obtain global context relationships and improve feature discrimination. Ding et al. [56] proposed the wide-context network for remote sensing semantic segmentation, which used the CNNs to preserve the spatial information and used the transformer to model the semantic dependencies. Chen et al. [57] presented a bitemporal image transformer (BIT) for CD, which can efficiently model contexts and obtain semantic relationships. Bandara and Patel [58] utilized the hierarchical transformer encoder in a Siamese architecture for CD, which can capture multiscale details. Zhang et al. [59] designed a Siamese U-shaped pure transformer network (SwinSUNet) to get long-term global information in spacetime.

III. PROPOSED METHOD

First, we introduce our CDENet in detail. Then, we describe the transformer block that models the global context. Next, we introduce the CDEM that enhances the change difference. Finally, we present the loss function that ensures the accuracy of change results in regions and pixels.

A. Overview

We propose a novel CDENet for CD, which can strongly model global context relationships and enhance the change difference. The overall architecture of our CDENet is shown in Fig. 2, which can better process the bitemporal images $T_1$ and $T_2$. The green and blue boxes represent the features from the dual encoders of U-Net. The dual encoders have the same structures, but they do not share parameters (params). The final red boldface arrow represents the output of the changed pixel. The CDENet is based on the dual U-Net structure [60] and the transformer is integrated into the dual-branch encoder. Therefore, our CDENet...
can combine the ability of U-Net to extract spatial details and the ability of the transformer to model global context relationships. More importantly, we specially designed the CDEM to enhance the difference for bitemporal features. Besides, the cross-layer feature fusion can transfer the change information between the encoder and decoder. In general, the main parts of our network include the transformer block and the CDEM. Among them, the transformer block can better extract the global information from the bitemporal features. And the CDEM can enhance the features of the bitemporal images. The size of our input image is $256 \times 256$, and the sizes of the features in each layer extracted from U-Net are $256 \times 256 \times 128$, $128 \times 128 \times 256$, $64 \times 64 \times 512$, and $32 \times 32 \times 1024$, respectively. The size of the feature processed after the transformer block is $32 \times 32 \times 512$. After the dual-branch encoder and the transformer block, the CDEM is used to enhance and obtain the difference. Therefore, the obtained change features are more accurate and complete. Through the cross-layer feature fusion between the encoder and decoder, the network finally generates the change results. The results are the probability values of the predicted change pixels. Thus, we combine the binary cross-entropy (BCE) loss and the dice loss to supervise the CD results in pixels and regions.

B. Transformer Block

We consider that it is difficult for convolution to capture global context, and the transformer-based methods can effectively model context. To fully capture the global context relationship, we integrate the transformer block in the last layer of the encoder. The transformer was originally integrated [53] into a single-stream network for segmentation. To fit the CD task, we need to capture the context relationships from the bitemporal features. Therefore, we specifically integrate the transformer block in the dual-stream network structure, which can model the context relationships from the bitemporal features. The sub-branches of the encoder in CDENet are identical to the encoder in [53]. The transformer block can better build the dependency of context features and improves the feature representation, which is conducive to the generation of the change difference features.

First, we divide the input features $x$ into a series of patches, and then vectorize the patches as $x_p$. Then, we map $x_p$ to D-dimensional feature space via the trainable linear projection. Besides, we obtain the learnable position features from the added patch embeddings. The added patch embeddings are defined as

$$z_0 = [x_1; x_2; \ldots; x_p; \omega^2; \omega^3; \ldots; \omega^N] + \omega_{pos}$$

where $[\ldots]$ represents the concatenation, $\omega^1_p, \omega^2_p, \ldots, \omega^N_p$ present the vectorized patches, $N$ is the number, $\omega \in \mathbb{R}^{(p^2 \times C) \times D}$ presents the embedding projection, $P$ presents the size, $C$ presents the channel number, and $\omega_{pos} \in \mathbb{R}^{N \times D}$ presents the position features.

There are $n = 12$ transformer layers in the transformer block. Among them, multihead self-attention and multilayer perceptron modules constitute the transformer layer. The output $z_n$ of the $n$th transformer layer is defined as

$$z'_n = \text{MSA} (\text{LN}(z_{n-1})) + z_{n-1}$$
$$z_n = \text{MLP} (\text{LN}(z'_n)) + z'_n$$

where LN() represents the layer normalization operator, MSA represents multilhead self-attention, and MLP represents multilayer perceptron. The MSA can calculate the global self-attention from the feature of the last layer in the encoder, which calculates each head attention and concatenates the results together as the output. The MSA is defined as

$$\text{MSA}(Q, K, V) = \text{Cat}(\text{head}_1, \ldots, \text{head}_h) \cdot W^O$$
$$\text{s.t. head}_i = \text{Attention}(QW^Q_i, KW^K_i, VW^V_i)$$

where $Q, K, V$ represent the query, key, and value, respectively. $\text{Attention()}$ is defined as

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V$$

where $Q, K,$ and $V$ represent the query, key, and value, respectively, and $d_k$ denotes the channel dimension.

Through the abovementioned operation, the representation feature $z_n$ is reshaped to the same size as original input feature $x$.

C. Content Difference Enhancement Module

Conventional methods usually directly subtract the bitemporal features to obtain the difference features. However, this direct acquisition of difference features will produce serious noise. To alleviate this problem, we introduce the CDEM to enhance difference for bitemporal features.

We input bitemporal images into encoder to extract features and generate four levels of features. The CDEM is designed to extract the difference features in each layer and emphasize the difference. The CDEM first obtains the initial difference feature by subtraction, but this feature contains a lot of noise and is not suitable for the difference result. Therefore, the CDEM generates the difference attention region from the initial difference feature, which can locate the change regions and preliminarily filter the noise. Then, the generated attention region acts on the original bitemporal features and performs subtraction. Through these operations, accurate difference results can be obtained.

First, we subtract the bitemporal features to obtain the initial difference $D_i$, which is defined as

$$D_i = |F_{t1} - F_{t2}|$$

where $F_{t1}$ denotes the features of the $t1$th layer at $T_1$ time, $F_{t2}$ denotes the features of the $t2$th layer at $T_2$ time, and $| \cdot |$ represents the absolute value operation. It ensures the non-negativity of the change difference features.

Then, to fully utilize the spatial relationship of the obtained difference features, we first operate spatial attention on the differences. Specifically, we obtain the location of possible spatial differences in features through the max pooling and average pooling. Then, we use convolution and sigmoid operations to generate attention about the location. The spatial difference attention $A_{D_i}$ is defined as

$$A_{D_i} = \sigma (f^{7 \times 7} ([P_{\max}(D_i); P_{\text{ave}}(D_i)]))$$

where $P_{\max}$ and $P_{\text{ave}}$ represent the max pooling and average pooling operations, respectively.
where $P_{\text{max}}()$ means the global max pooling operation, $P_{\text{ave}}()$ means the global average pooling operation, $f_{7 \times 7}$ means the convolutional layers with $7 \times 7$ kernel size, which is used to generate the possible change attention, and $\sigma$ means the sigmoid operation.

Finally, the spatial difference attention $A_{D_i}$ acts on the bitemporal features and generates enhanced difference features $O_{D_i}$ through subtraction. The $O_{D_i}$ is defined as

$$O_{D_i} = |(A_{D_i} + 1) \cdot F_{i1} - (A_{D_i} + 1) \cdot F_{i2}|$$

where $\cdot$ represents the element-wise multiplication.

The CDEM is suitable for enhancing the difference features. It captures the detailed location information of the difference features through spatial attention and applies the spatial attention weight to the bitemporal features. The CDEM can obtain enhanced difference features from these operations.

Besides, some details information will be lost during the transmission from the encoder to the decoder. To alleviate the problems, we introduce the simple cross-layer feature fusion, which fuses the features by concatenation operation. The simple cross-layer feature fusion can transfer the details of the difference features between the encoder layer and the corresponding decoder layer. It mainly fuses the downsampling features in the encoder with the upsampling features in the decoder to obtain more accurate context relationships.

**D. Loss Function**

Essentially, CD is a binary classification task, that is, the pixels in remote sensing images are divided into changed with the value of 1 and unchanged with the value of 0. Therefore, we can use classification loss to supervise our network training. Specifically, to better supervise the training of CD, we adopt a novel loss, which is composed of the BCE loss and the dice loss.

The BCE loss is the pixel-level supervision between the prediction results and the ground truth (GT). It is the point-to-point change supervision and is defined as

$$L_{\text{bce}}(X, Y) = - \sum_i (Y_i \log(X_i) + (1 - Y_i) \log(1 - X_i))$$

where $X$ denotes the prediction results and $Y$ denotes the GT.

The dice loss is the region-level supervision between the prediction results and the GT. It is the regional change supervision and is defined as

$$L_{\text{dice}}(X, Y) = 1 - 2 \cdot \frac{|X \cap Y|}{|X \cup Y|}$$

where $X$ denotes the prediction results and $Y$ denotes the GT.

The total loss function $L_{\text{total}}$ can be denoted as

$$L_{\text{total}}(X, Y) = \lambda_1 \cdot L_{\text{bce}}(X, Y) + \lambda_2 \cdot L_{\text{dice}}(X, Y)$$

where $\lambda_1$ and $\lambda_2$ are the corresponding hyperparameters of the $L_{\text{bce}}$ loss and $L_{\text{dice}}$ loss, respectively.

The BCE loss can supervise the prediction results at the pixel level and the dice loss can supervise the prediction results at the regional level [61]. Because the total loss combines the BCE loss and the dice loss, it can conduct hybrid supervision on the predicted change results at the pixel level and the regional level.

E. Prediction of the Proposed CDENet

We display the representative CD results, which are shown in Fig. 3. Note that, different from the difference enhancement module in [49], our content enhancement module is based on spatial attention rather than channel attention. That is, our module enhances spatial content differences rather than channel differences. Our content enhancement module mainly focuses on the change region, and spatial attention can enhance content information. We also replace our CDEM with the channel attention enhancement module (CAEM) and display their difference.

We can see that the change region in Fig. 3(e) is very complete, while the change region in Fig. 3(g) is missing. In addition, we also show the effect of superimposing heat maps on the T2 image. We can see more clearly that the superimposed result in Fig. 3(f) can cover more change regions and maintain better details than Fig. 3(h). Therefore, the result of our CDENet is closer to the GT than the CAEM, which means our CDENet can better enhance the content difference.

IV. EXPERIMENTS

First, we present the datasets and evaluation metrics. Then, we clearly introduce the experimental settings. Next, we show and analyze the comparison result between the proposed CDENet and other methods. The visual and quantitative results show the great performance of our CDENet. Finally, the ablation experiments show the effectiveness of each module in the CDENet.

A. Datasets

Our CDENet is evaluated on two popular CD datasets: 1) LEVIR-CD dataset [43]; and 2) SYSU-CD dataset [51].

LEVIR-CD is a large-scale building CD dataset. It consists of 637 VHR Google Earth image pairs, with the image resolution of 0.5 m and the image size of $1024 \times 1024$. These image pairs are bitemporal images, and their time span is 5–14 years. It shows significant building-related land changes in 20 different regions.
of Texas, USA. The samples of the dataset are binary labels annotated with high-quality annotations, where 1 represents change and 0 represents unchanged. In our experiments, we first crop original images of size $1024 \times 1024$ into nonoverlapping images of size $256 \times 256$. According to the official division, we divide the processed image pairs into three parts: 1) 7120 samples are used as the training set; 2) 2048 samples are used as the testing set; and 3) 1024 samples are used as the validation set.

SYSU-CD is a CD dataset reflecting the cover changes in Hong Kong from 2007 to 2014. The types of the dataset include new urban buildings, vegetation changes, road expansion, sea construction, and so on. SYSU-CD contains 20,000 aerial image pairs, with the image resolution of 0.5 m and the image size of $256 \times 256$. According to the official division, 12,000 samples are used as the training set, 4000 samples are used as the testing set, and 4000 samples are used as the validation set. For fair comparisons, we use the same dataset to train all methods and evaluate their performance.

### B. Evaluation Metrics

To evaluate the experimental results, we adopt five quantitative indicators: 1) precision; 2) recall; 3) F1-score (F1); 4) overall accuracy (OA); and 5) intersection over union score (IoU).

TP represents the number of changed pixels with correct classification. FP represents the number of changed pixels with incorrect classification. FN represents the number of unchanged pixels with incorrect classification. TN represents the number of unchanged pixels with correct classification.

Precision indicates the proportion of the correct change pixels in prediction result and the total change pixels in prediction result. The precision is defined as

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (12)$$

Recall indicates the proportion of the correct change pixels in prediction result and the total change pixels in GT. The recall is defined as

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (13)$$

F1 combines accuracy and recall. Its formula is

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (14)$$

OA is the proportion of pixels correctly predicted, and its formula is

$$\text{OA} = \frac{TP + TN}{TP + TN + TP + FN}. \quad (15)$$

IoU indicates the region overlay between the prediction result and the GT. Its formula is

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \quad (16)$$

Because the evaluation of CD needs to comprehensively consider the results of accuracy and recall, the F1, OA, and IoU can more comprehensively reflect the performance of the methods.

### C. Implementation Details

Our CDENet is implemented with Pytorch, and all experiments are conducted on the NVIDIA GeForce RTX 3090Ti GPU (24 GB memory). There are 19,120 samples in the training dataset, including 7120 samples from LEVIR-CD and 12,000 samples from SYSU-CD. The size of our input and output images is $256 \times 256$. The training details are presented as follows. We use the Adam gradient descent [62] to optimize our CDENet. We train our CDENet for 200 epochs and retain the optimal model. The batch size is 12 and the initial learning rate is $5 \times 10^{-3}$. The hyperparameters in our loss are set as follows. $\lambda_1$ is 0.8 and $\lambda_2$ is 0.2. Because the prediction results are the probability values of change pixels, in the test phase, the values of each pixel are binarized to 0 or 1 according to the threshold value of 0.5.

### D. Comparative Experiments

1) Comparison Methods: We compare the proposed CDENet with some existing methods. The comparison methods include FC-EF [25], FC-Siam-conc [25], FC-Siam-diff [25], DSIFN [44], DTCDCSN [48], SNUNet [50], DSAMNet [51], BIT [57], and ChangeFormer [58].

FC-EF [25] concatenated the bitemporal images and fed them into a U-shaped structure. FC-Siam-conc [25] was based on a Siamese U-shaped structure, which extracted multi-level features and concatenated the features of each layer. FC-Siam-diff [25] was also based on a Siamese U-shaped structure, which extracted multi-level features and utilized the difference of features for skip connection. The DSIFN [44] adopted the two-stream FCN to extract representative multi-level features from bitemporal images. Then, DSIFN performed difference discrimination on representative features to obtain the results. The DTCDCSN [48] used dual-attention to capture the relationships between channels and spatial positions. The SNUNet [50] combined the NestedUNet and the Siamese network, which can aggregate and refine multi-level features. The DSAMNet [51] integrated the CBAM module [32] to learn a change map. Besides, an auxiliary supervision module was used to enhance features, which made the features contain more spatial information. BIT [57] modeled the context relationships with the transformers, which can better identify the change regions. ChangeFormer [58] utilized the hierarchical transformer encoder in a Siamese architecture, which can capture multiscale details.

2) Experiments on the LEVIR-CD Dataset: The visual experiment results are shown in Fig. 4. For small change object (see first row), our method can completely detect the region and structure, but the comparison methods detect incomplete regions and even some methods miss the small object. For the large building (see second row), our method can detect the correct region and keep the boundary smooth, but the comparison methods detect the wrong regions, the correct regions detected are discontinuous, and some methods even cannot detect the object. For sparse multiple buildings (see third row), our method can maintain the structure and details of the sparse multiple buildings, but the comparison methods often lose the structure of
the object and blur the details. For dense multiple buildings (see fourth row), our method can accurately detect the dense multiple buildings and keep their boundaries clear, but the comparison methods often mistakenly detect the buildings, and there will be adhesion between the multiple buildings. We can see that our method has achieved excellent performance in maintaining the complete structure and accurate details.

To directly show the performance of our CDENet, we conduct quantitative experiments. The quantitative results are given in Table I, where boldface numbers present the best value. From Table I, we can see that our method performs best in F1, OA, and IoU indexes. Specifically, in terms of the F1 index, our method is 0.8844 and is 0.98% higher than the second-best method (BIT). In terms of the OA index, our method is 0.9864 and is 0.07% higher than the second-best method (BIT). In terms of the IoU index, our method is 0.7927 and is 1.55% higher than the second-best method (BIT). Qualitative and quantitative results on the LEVIR-CD dataset show that our CDENet has achieved satisfactory results and better performances.

3) Experiments on the SYSU-CD Dataset: The visual experiment results are shown in Fig. 5. For the small building (see first row), our method can completely detect the region and structure, but the comparison methods often detect the wrong regions and even some methods miss the small object. For multiple objects (see second row), our method can accurately detect the multiple objects and keep their profile, but the comparison methods often mistakenly detect the objects and blur their profile. For the large continuous vegetation (see third row), our method can correctly detect the large continuous vegetation and maintain great details and contours. But the comparison methods often detect the discontinuous regions and inaccurate contours. For the vegetation in large discontinuous regions (see fourth row), our method can correctly detect the vegetation in different regions and maintain great consistency and details. But the comparison methods cannot correctly detect the vegetation in different regions, and the continuity and details of the detected vegetation are poor. We can see that our method has achieved excellent performance in maintaining accurate regions and details.

To directly show the performance of our CDENet, we conduct quantitative experiments. The quantitative results are given in Table II, where boldface numbers present the best value. From Table II, we can see that our method performs best in F1, OA, and IoU indexes. Specifically, in terms of the F1 index, our method is 0.7593 and is 0.21% higher than the second-best method (ChangeFormer). In terms of the OA index, our method is 0.8966 and is 0.47% higher than the second-best method (ChangeFormer). In terms of the IoU index, our method is 0.6119 and is 0.27% higher than the second-best method (ChangeFormer).

| Method          | Precision | Recall | F1 | OA   | IoU   |
|-----------------|-----------|--------|----|------|-------|
| FC-EF [25]      | 0.7358    | 0.5458 | 0.6267 | 0.9669 | 0.4564 |
| FC-Siam-conc [25] | 0.8497    | 0.4978 | 0.6278 | 0.9699 | 0.4576 |
| FC-Siam-diff [25] | 0.9160    | 0.5820 | 0.7118 | 0.9760 | 0.5525 |
| DSIFN [44]      | 0.8486    | 0.8945 | 0.8710 | 0.9865 | 0.7714 |
| DTCDSCN [48]    | 0.8749    | 0.8544 | 0.8645 | 0.9864 | 0.7614 |
| SNUNet [50]     | 0.8818    | 0.8552 | 0.8683 | 0.9868 | 0.7673 |
| DSAMNet [51]    | 0.8314    | 0.8462 | 0.8387 | 0.9834 | 0.7222 |
| BIT [57]        | 0.9057    | 0.8457 | 0.8746 | 0.9877 | 0.7772 |
| ChangeFormer [58] | 0.8724    | 0.8597 | 0.8660 | 0.9864 | 0.7637 |
| Ours            | 0.8975    | 0.8716 | **0.8844** | **0.9884** | **0.7927** |
Qualitative and quantitative results on the SYSU-CD dataset show that our CDENet has achieved satisfactory results and better performances.

**E. Efficiency Studies**

Table III presents the efficiency studies of the methods on the LEVIR-CD dataset. Specifically, we conduct efficiency studies on our CDENet and the SOTA methods in multiple quantitative indicators, including the number of params, floating point operations (FLOPs), and F1. Our CDENet is based on the dual-encoder of U-Net, and they do not share the params. The bitemporal features are processed by a dual-encoder of U-Net, respectively. Ours (Siamese network) is based on one encoder of U-Net, and the bitemporal features are only processed by the single encoder of U-Net. Compared with SOTA methods, the params of our CDENet are large, and the params of the Siamese network are normal. It can be seen that the dual-flow structure causes an increase in network params. In addition, compared with SOTA methods, the FLOPs of our CDENet are medium, which means the proposed network is not very complex. And the F1 of our CDENet is the highest, which indicates that the detection performance of our CDENet is the best. From the efficiency studies, we can see that the params of our CDENet are acceptable, the FLOPs of our CDENet are reasonable, and our performance is optimal.
TABLE IV
QUANTITATIVE ABLCATION RESULTS ON THE LEVIR-CD DATASET

| Method                  | Params | FLOPs | Precision | Recall | F1    | IoU  |
|------------------------|--------|-------|-----------|--------|-------|------|
| Baseline               | 42.12  | 29.52 | 0.7941    | 0.7948 | 0.7945 | 0.6590 |
| + CLFF                 | 43.67  | 33.15 | 0.9022    | 0.8471 | 0.8738 | 0.7759 |
| + Transformer Block    | 129.68 | 55.24 | 0.8994    | 0.8585 | 0.8785 | 0.7834 |
| + CDEM                 | 129.68 | 55.24 | 0.8909    | 0.8738 | 0.8822 | 0.7893 |
| Ours (Siamese network) | 66.78  | 55.24 | 0.9085    | 0.6027 | 0.7247 | 0.5682 |
| Ours                   | 129.68 | 55.24 | 0.8975    | 0.8716 | 0.8844 | 0.7927 |

F. Ablation Studies

To study the factors affecting the results of the CDENet and the effectiveness of different modules, we perform ablation experiments on the LEVIR-CD dataset. The quantitative ablation results are given in Table IV, where boldface numbers present the best value.

Specifically, we mainly research the impact of the cross-layer feature fusion (+CLFF), the transformer block (+Transformer Block), the CDEM (+CDEM), the loss function (Ours), and the Siamese network [Ours (Siamese network)]. First, we adopt the dual-stream FCN and a simple feature subtraction method as the baseline. We use the simple cross-layer feature fusion to transfer the details of the changed regions from the encoder layer to the corresponding decoder layer. The +CLFF makes the F1 increase by 7.93% and makes the IoU increase by 11.69% more than the baseline. So, the +CLFF can better transfer the details from the encoder layer to the corresponding decoder layer. To capture the global context of the bitemporal images, the transformer block is integrated into the last convolution layer of the dual-stream. The +Transformer Block makes the F1 increase by 0.47% and makes the IoU increase by 0.75% more than the +CLFF. So, it can better capture the global context relationships. Then, the CDEM is proposed to enhance the difference for bitemporal features. The +CDEM makes the F1 increase by 0.37% and makes the IoU increase by 0.59% more than the +Transformer Block. So, it can better enhance the difference for bitemporal features. The proposed loss can better supervise the training of CD. The Ours makes the F1 increase by 0.22% and makes the IoU increase by 0.34% more than the +CDEM, which means the proposed loss can appropriately improve the detection performance. Besides, we also compare the dual-stream network (Ours) and Siamese network [Ours (Siamese network)] on our model. The Ours makes the F1 increase by 15.97% and makes the IoU increase by 22.45% more than the Ours (Siamese network). And the Ours makes the F1 increase by 8.99% and makes the IoU increase by 13.37% more than the baseline. We also can see that the params and FLOPs in +CDEM and Ours are the same as the +Transformer Block. The params and FLOPs in +CLFF are a little higher than baseline. The params and FLOPs in +Transformer Block are much higher than +CLFF. The quantitative ablation results reveal the effectiveness of the proposed network.

We provide the visual results to demonstrate the effect of the transformer block, which are shown in Fig. 6. Without the transformer block, our method cannot identify the part of the change regions. With the transformer block, our method can completely detect the change regions. Therefore, the transformer block can better capture the global context relationships for each bitemporal feature and help locate the change regions.

We also show the quantitative comparison between our CDEM and the CAEM [49]. The quantitative comparison results are given in Table V, where boldface numbers present the best value. We can see that Ours makes the F1 increase by 0.89% and makes the IoU increase by 1.41% more than the result of CAEM. The comparison results present that our CDEM performs better than the CAEM.

Besides, we provide the visual results to demonstrate the performance of the proposed loss. The visual results are shown in Fig. 7. We can see that the change region in Fig. 7(d) is inaccurate and chaotic, while the change region in Fig. 7(e) is complete and continuous. The proposed loss can detect more accurate regions than the BCE loss. Therefore, we can conclude that the proposed loss can better maintain the accuracy of regions and pixels.

We have designed experiments to determine the hyperparameters of the loss function. Experimental results of loss parameters on the LEVIR-CD dataset are given in Table VI, where boldface
numbers present the best value. \( \lambda_1 \) and \( \lambda_2 \) are the corresponding hyperparameters of the \( L_{bce} \) loss and \( L_{dice} \) loss, respectively. The experimental results show that our performance is optimal when \( \lambda_1 = 0.8 \) and \( \lambda_2 = 0.2 \).

V. CONCLUSION

In this article, we proposed a CDENet for remote sensing image CD. The proposed CDENet is conducive to modeling global context relationships and enhancing the change difference. Specifically, the CDENet is based on U-Net and equipped with transformer block in the encoder. The U-Net structure can well utilize the bitemporal features. And the transformer block can capture the global context relationships among different-level features. Moreover, we design the CDEM to process the dual features of each layer, which increases the attention of difference regions. Therefore, we can enhance and obtain the change difference features in the encoder. With the cross-layer feature fusion, we can transfer the details of the changed regions from the encoder layer to the corresponding decoder layer. It is beneficial to transmit more effective change difference information and get more accurate results. Finally, the proposed loss can better supervise the accuracy of change results in regions and pixels. The visual and quantitative experimental results on two popular CD datasets demonstrate that our CDENet is effective and superior to other SOTA methods. Our CDENet is mainly used to improve the detection performance, but the limitation is that the params of our CDENet are large. For future work, we will consider reducing the params of the model while maintaining the accuracy of CD.

REFERENCES

[1] X. Zheng et al., “Full parameter time complexity (FPTC): A method to evaluate the running time of machine learning classifiers for land use/land cover classification,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 2222–2235, Jan. 2021.

[2] S. Shi, Y. Zhong, J. Zhao, P. Lv, Y. Liu, and L. Zhang, “Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery,” IEEE Trans. Geosci. Remote Sens., vol. 60, Nov. 2022, Art. no. 5600116.

[3] F. Nirchio, “Speditive COSMO-SkyMed SAR quick-look analysis for change detection in support to disaster management and environmental monitoring,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2015, pp. 4809–4812.

[4] A. Reigber, M. Jäger, and E. Krogager, “Polarimetric SAR change detection in multiple frequency bands for environmental monitoring in arctic regions,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2016, pp. 5702–5705.

[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] W. Zhang, L. Shen, and W. Qiao, “Building damage detection in VHR satellite images via multi-scale scene change detection,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2021, pp. 8570–8573.

[7] W. A. Mallia, “Change vector analysis: An approach for detecting forest changes with landsat,” in Proc. Mach. Process. Remotely Sensed Data Symp., 1980, pp. 326–335.

[8] F. Bovolo and L. Bruzzone, “A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain,” IEEE Trans. Geosci. Remote Sens., vol. 45, no. 1, pp. 218–236, Jan. 2007.

[9] J. Deng, K. Wang, Y. 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.

[10] 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.

[11] S. Liu, J. Cheng, L. Liang, H. Bai, and W. Dang, “Light-weight semantic segmentation network for UAV remote sensing images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 8287–8296, Aug. 2021.

[12] X. Li, Z. Zhao, and Q. Wang, “ABSSNet: Attention-based spatial segmentation network for traffic scene understanding,” IEEE Trans. Cybern., vol. 52, no. 9, pp. 9352–9362, Sep. 2022.

[13] S. Girisha, U. Verma, M. M. M. Pai, and R. M. Pai, “U-vid-Net: Enhanced semantic segmentation of UAV aerial videos by embedding temporal information,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 4115–4127, Mar. 2022.

[14] J. Fang, Y. Yuan, X. Lu, and Y. Feng, “Robust space–frequency joint representation for remote sensing image scene classification,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 10, pp. 7492–7502, Oct. 2019.

[15] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, “Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 3735–3756, Jun. 2020.

[16] J. Ma, M. Li, X. Tang, X. Zhang, F. Liu, and L. Jiao, “Homo–heterogenous transformer learning framework for RS scene classification,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 2223–2239, Mar. 2022.

[17] K. Liu, Z. Jiang, M. Xu, M. Perc, and X. Li, “Tilt correction toward building detection of remote sensing images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 5854–5866, May 2021.

[18] J. Huang, K. Liu, M. Xu, M. Perc, and X. Li, “Background purification framework with extended morphological attribute profile for hyperspectral anomaly detection,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 8113–8124, Aug. 2021.

[19] X. Li, D. Song, and Y. Dong, “Hierarchical feature fusion network for salient object detection,” IEEE Trans. Image Process., vol. 29, pp. 9165–9175, Sep. 2020.

[20] D. Song, Y. Dong, and X. Li, “Hierarchical edge refinement network for saliency detection,” IEEE Trans. Image Process., vol. 30, pp. 7567–7577, Aug. 2021.

[21] X. Li, G. Li, and B. Zhao, “Low-light hyperspectral image enhancement,” IEEE Trans. Geosci. Remote Sens., vol. 60, Aug. 2022, Art. no. 5537013.

[22] Y. Dong, W. Tan, D. Tao, L. Zheng, and X. Li, “CartoonLossGAN: Learning surface and coloring of images for cartoonization,” IEEE Trans. Image Process., vol. 31, pp. 4885–4908, Dec. 2022.

[23] X. Li, M. Chen, F. Nie, and Q. Wang, “A multiview-based parameter free framework for group detection,” in Proc. AAAI Conf. Artif. Intell., 2017, pp. 4147–4153.

[24] L. Zhang, L. Zhang, and B. Du, “Deep learning for remote sensing data: A technical tutorial on the state of the art,” IEEE Geosci. Remote Sens. Mag., vol. 4, no. 2, pp. 22–40, Jun. 2016.

[25] R. C. Daudt, B. Le Saux, and A. Boulch, “Fully convolutional Siamese networks for change detection,” in Proc. Int. Conf. Image Process., 2018, pp. 4063–4067.
G. Liu, Y. Yuan, Y. Zhang, Y. Dong, and X. Li, “Style transformation-based spatial–spectral feature learning for unsupervised change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5401515.

X. Peng, K. Zhong, Z. Li, and Q. Li, “Optical remote sensing image change detection based on attention mechanism and image difference,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.

M. Gong et al., “A spectral and spatial attention network for change detection in hyperspectral images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 5521614.

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, Nov. 2022, Art. no. 5406216.

F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” 2015, arXiv:1511.07122.

X. Wang, F. Ding, J.-W. Chen, L. Jiao, and L. Wang, “A lightweight convolutional neural network for bitemporal image change detection,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2551–2554.

S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

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

A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.

L. Yuan et al., “Tokens-to-token ViT: Training vision transformers from scratch on ImageNet,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 538–547.

Karthik and B. R. Shivakumar, “Change detection using image differencing: A study over area surrounding Kumta, India,” in *Proc. 2nd Int. Conf. Elect., Comput. Commun. Technol.*, 2017, pp. 1–5.

R. D. Molin, R. A. Rosa, F. Mayer, M. I. Pettersson, and R. Machado, “A change detection algorithm for SAR images based on logistic regression,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 1514–1517.

L. Bruzzone and D. F. Prieto, “Automatic analysis of the difference image for unsupervised change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

S. Krinidis and V. Chatzis, “A robust fuzzy local information C-means clustering algorithm,” *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1328–1337, May 2010.

X. Zhang, Z. Li, B. Hou, and L. Jiao, “Spectral clustering based unsupervised change detection in SAR images,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2011, pp. 712–715.

X. Li, M. Chen, F. Nie, and Q. Wang, “Locality adaptive discriminant analysis,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2201–2207.

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.

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.

J. Chen et al., “DASNNet: 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, Nov. 2021.

S. Li and L. Huo, “Remote sensing image change detection based on fully convolutional network with pyramid attention,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4352–4355.

M. Li, M. Li, P. Zhang, Y. Wu, W. Song, and L. An, “SAR image change detection using PCANet guided by saliency detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 16, no. 3, pp. 402–406, Mar. 2019.

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.

T. Lei et al., “Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2022, Art. no. 4507013.

S. Fang, K. Li, J. Shao, and Z. Li, “SNUNet-CD: A densely connected Siamese network for change detection of VHR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5604816.