Dual-Attention Cross Fusion Context Network for Remote Sensing Change Detection

Yu Shangguan, Jinjiang Li, and Liang Chang

***Abstract*—Detecting changes in two remote sensing images of the same region but at different times is of great significance in applications, such as land management and urban planning, which also prompts the continuous development and progress of change detection (CD) technology. The current deep learning-based methods make full use of the excellent feature learning ability of deep convolution to show excellent detection performance. However, advances in remote sensing technology also mean that detected objects have higher resolution and more complex content, which is more challenging for CD techniques. Strengthening the model's ability to learn the context of the detected remote sensing image can effectively improve the model's ability to distinguish between changing features and nonchanging features, thereby achieving higher-precision detection results. In order to explore and utilize the contextual information of different levels of features as much as possible, we design a dual-attention cross fusion module in our method to realize the cross-learning of contextual information of different scales during the decoding process. It will be able to complementarily fuse feature content of different granularities. We also propose an Atrous Pyramid Difference Module (APDM) to efficiently capture the difference information of two refined features by exploiting receptive fields of different sizes. In addition, in order to further improve the context modeling ability of the model, we introduce a context transformer block (Cot). Different from other transformer-based self-attention methods, Cot dynamically guides the learning of the attention matrix by the contextual information of the input keys. Our method achieves F1-scores of 91.08%/91.93%/79.80% on the LEVIR-CD/WHU-CD/DSIFN-CD datasets, respectively. Extensive qualitative and quantitative experiments on these datasets validate the effectiveness of our method.**

***Index Terms*—Atrous pyramid difference module (APDM), change detection (CD), context transformer block, dual-attention cross fusion module (DCFM), high-resolution remote sensing image, siamese network.**

I. Introduction

CHANGE detection (CD) technology in the field of remote sensing compares and analyzes a pair of high-resolution remote sensing satellite images reflecting the same area at different times to obtain the changes of the environment and ground objects in this area during this period of time [1]. In practical applications, accurate prediction of these changes can effectively guide some land planning and management work, including urban planning, environmental monitoring, and disaster assessment [2], [3], [4]. As shown in Fig. 1, the detection results for these changes are usually described by a binary mask. It uses positive labels to denote pixels that have changed, appearing as white, and negative labels to denote pixels that have not changed, appearing black.

Table I summarizes the characteristics of CD in the field of current remote sensing images, including data types, detection methods, feature extraction methods, application scenarios, and existing challenges. In the early years of CD technology, manual visual analysis relying on professional knowledge and experience was the main way to distinguish the differences in image changes, which was time consuming and labor-intensive and difficult to handle large-scale geographical detection [5]. In addition, some commonly used algebra-based processing methods such as image regression, image difference, and change vector analysis [6] also have problems of low efficiency and strong scene limitations. In order to reduce labor and time costs and effectively deal with higher resolution dual-temporal remote sensing satellite images, artificial intelligence methods have been widely tried in CD tasks and initially demonstrated their advantages. The method based on machine learning has been actively applied to the CD of remote sensing images and has effectively improved the accuracy of CD due to its powerful adaptive and automatic capabilities [7], [8], [9], [10], [11]. The representative methods mainly include: naive Bayesian [12], [13], support vector machine [14], [15], random forest [16], [17], and decision tree [18], but they still have the characteristics of low computational efficiency. Currently, methods based on deep convolutional neural networks (CNNs) combine multilayer operations to explore feature representations at the abstract level, which have achieved satisfactory results in many computer vision (CV) tasks including remote sensing image CDs [2], [19], [20], [21], [22], [23], [24], [25], [26], [27]. It thanks to CNNs' superior feature extraction ability, and the proposal of ResNet [28] and UNet [29] structures has made CNNs more widely and more effectively applied. However, general CNNs

Manuscript received 22 May 2023; revised 8 July 2023, 6 August 2023, 3 September 2023, and 14 September 2023; accepted 14 September 2023. Date of publication 20 September 2023; date of current version 3 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62272281, Grant 62002200, Grant 62202268, and Grant 61972235, in part by the Shandong Natural Science Foundation of China under Grant ZR2023MF026, and Grant ZR2022MA076, and in part by the Shandong technology and business university wealth management interdisciplinary research project (2023TSXM003), Yantai science and technology innovation development plan under Grant 2022JCYJ031. *(Corresponding author: Liang Chang.)*

Yu Shangguan and Jinjiang Li are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: 1359304997@qq.com; lijinjiang@gmail.com).

Liang Chang is with the School of Public Administration, Shandong Technology and Business University, Yantai 264005, China (e-mail: 201913692@sdtbu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3317488

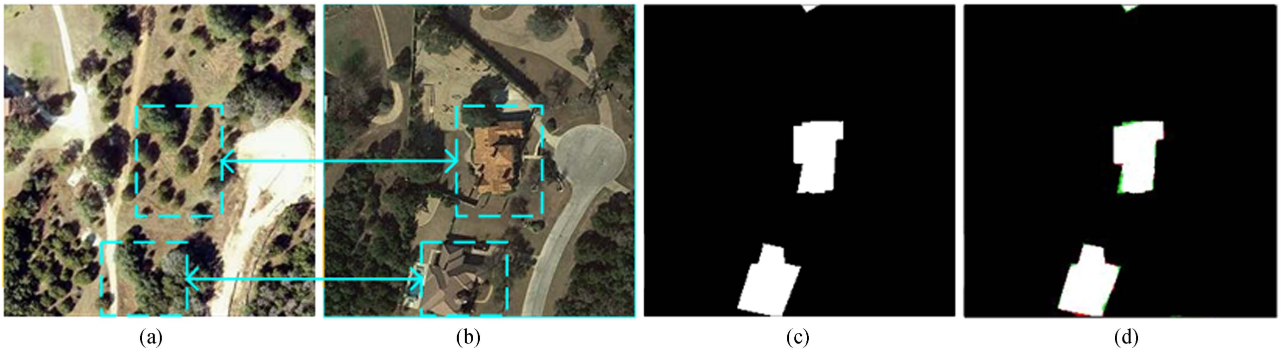

Fig. 1. (a) Image before change. (b) Image after change. (c) Ground truth. (d) Results of our method.

TABLE I
CHARACTERISTICS OF THE FIELD OF CD IN REMOTE SENSING IMAGES

characteristics	Description
Date type	Synthetic aperture radar (SAR), multispectral, hyperspectral, very high spatial resolution (VHR) imagery and heterogeneous imagery.
Methods for change detection	Pixel-based change detection and object-based object change detection
Methods of feature extraction	Traditional hand-designed features, convolutional neural networks in deep learning, transform-based feature extraction
Application scenarios	Urban sprawl, building change detection, land use change detection, disaster damage assessment, agricultural surveys, and map revisions
Existential challenges	The complexity of objects and spectra in the image scene, the differences caused by imaging under different periods and conditions, and the existence of various noise sources and artifacts in remote sensing images

still have some insurmountable problems: shallow CNNs can effectively extract low-level features, such as edges and shapes, but cannot capture broader contextual information, which affects the model's ability to perceive the real world. The deep network implemented by using skip connections proposed by ResNet can obtain more effective features and enhance the global exploration performance of the model, but it usually contains a large number of parameters and requires more computing resources. It is also difficult for CNNs to balance the combination of low-level features and high-level features. In the CD task of remote sensing images, in order to identify and segment changing targets more accurately, it is necessary to distinguish the objects and backgrounds in the images. Effective exploration of contextual information can help the model understand the contextual relationship of the target in the surrounding environment, so as to achieve a reasonable distinction between the environment and the target. Therefore, the context modeling capability of the model becomes particularly critical. The use of various attention mechanisms, such as channel attention [22], [23], [30], [31], spatial attention [22], [23], [30], cross-attention [26], [32], [33], and self-attention [2], [19] can directionally strengthen the context understanding ability of the model at different levels. They also play an important role in the work of efficient CDs. However, simply redistributing the weights of a single-level feature map in the spatial or channel dimension is not enough to fully capture rich contextual information, because it is limited to a single granularity of feature information.

In order to fully exploit the spatial context information of remote sensing images and the long-range connections between changing objects and backgrounds to achieve high-precision detection, we propose a dual-attention cross fusion context network (DCFCNet). A dual-attention cross fusion module (DCFM) is designed to explore the complementary relationship between adjacent-level features based on multiscale work. In this process, spatial attention is used to preserve the spatial information of low-granularity features, whereas channel attention is used to preserve the channel information of high-granularity features. These complementary relations can effectively guide learning an efficient fusion of these features to reasonably and comprehensively explore the contextual information of the input image. We also consider the effectiveness of the fusion method of bitemporal features and propose an atrous pyramid difference module (APDM) to enhance the model's difference computation performance. After obtaining the change feature containing difference information, we additionally use a context transformer block (Cot) modified based on the transformer's self-attention module to optimize the change feature according to its spatial context information. Cot uses the connection of adjacent keys to match the query and aggregates with the value to explore the dynamic and static context of the obtained change feature.

The rest of this article is organized as follows. We next introduce the deep learning-based CD technique together with the attention mechanism in the relevant part of Section II. In Section III, methods we detail the implementation principles and reasons for using each module of our network. In the experimental part of Section IV, we introduce the used datasets, evaluation metrics, experimental implementation details and show our experimental results including comparison experiments with advanced methods and ablation experiments.Finally, Section V concludes this article.

II. Related Work

A. Deep Learning-Based CD Techniques

In the past period of time, deep learning-based CD technology has recently become a mainstream approach to solve CD in high-resolution remote sensing images due to its outstanding performance and convenient automation capabilities. The method based on deep learning can automatically extract multilayer rich feature information from the input target image and use them to efficiently model the relationship between feature objects and the real world, which can help detect change information more effectively. Deep learning-based methods can be divided into the following three categories according to the availability of labeled data during the training phase.

1) Unsupervised methods.
2) Semisupervised methods.
3) Supervised methods [34].

It is expensive to produce ground truth that reflects real change information for paired high-resolution remote sensing images, which makes unsupervised methods practically applied in CD tasks in the field of remote sensing [35]. Unsupervised deep learning methods can autonomously mine and explore the inherent structure and regularity of datasets, which can be more flexibly applied to unlabeled data samples. Zhange et al.[36] designed an unsupervised network for the CD task on multispectral images. It first converts the spectral channel of the image into an abstract feature space through the deep belief network to distinguish the changed area from the nonchanged area to obtain effective features, and then uses the feature change analysis to analyze and obtain different types of changes. Correa et al.[37] used the linear change features based on tasseled caps and orthogonal equations to uniformly represent the image information acquired by multiple sensors when processing the CD task of multitemporal very high spatial resolution (VHR) images obtained by different spectral sensors. This process makes these input images comparable in time for efficient CD in urban areas. Li et al. [38] proposed an end-to-end unsupervised CD method applied to hyperspectral images (HSI). This method uses the existing FCN framework to learn image features and uses noise reduction modeling to enhance the robustness of the model during training.

The semisupervised deep learning method has the characteristics of both supervised learning and unsupervised learning, which is reflected in the fact that the dataset objects it processes contain a small amount of labeled data and a large amount of unlabeled data. While semisupervised methods make full use of unlabeled data to improve the generalization ability of the model, they can also be extended with labeled datasets to improve the diversity and anti-interference performance of the dataset. Gao et al. [39] proposed a semisupervised network based on convolutional-wavelet neural networks (CWNNs), which detects sea ice changes in synthetic aperture radar (SAR) images based on the change type of pixels. In this method, they applied the dual-tree complex wavelet transform to the CNN to eliminate the influence of speckle noise in SAR images. On the other hand, the strategy of CWNNs virtual sample generation can solve the problem of limited samples to strengthen the training process. Saha et al. [40] proposed a semisupervised network based on graph convolution, which encodes the input multitemporal image into a graph structure through multiscale parcel segmentation and uses graph convolutional neural network [41] to further model the relationship between them. This method propagates information from labeled nodes to unlabeled nodes through an iterative training process to improve CD performance.

Compared with unsupervised and semisupervised deep learning methods, supervised methods rely on more labeled data to correlate the matching degree of accuracy between the output obtained by the input and the expectation, thereby helping the model to learn useful information in a directional manner. This has more general applicability in the background of increasingly abundant remote sensing image technology and data sets. Since the idea of residual in ResNet [28] was proposed, the CNN-based network can achieve a deeper structure and obtain multilevel feature information and thus achieve a significant effect improvement. Early methods based on supervised deep learning also achieved superior performance by frequently relying on pure CNN architectures [42], [43]. The UNet structure [29] realizes the discovery of global and local information of input image features through the structure of encoding and decoding, and it has also been proved that it can effectively handle other CV tasks including CD to achieve scalable applications. Peng et al. [44] proposed an improved UNet++ network based on UNet to achieve effective detection of changes in remote sensing images. It uses dense skip connections to combine multiple scale feature information to comprehensively obtain change information. Moustafa et al. [45] designed a CD workflow architecture for CD in hyperspectral data. This workflow evaluates four variants based on the UNet structure, including residual UNet, residual recurrent UNet, attentional residual UNet, and attentional residual recurrent UNet. It proves that deep neural networks can combine complex features to enhance CD performance on HSI data.

In this article, we hope to effectively combine different levels of useful features to improve the model's perception of complex changes before computing the difference information of bitemporal features. However, it is not enough to simply use concatenation or addition operations to aggregate features as in previous CD work based on the UNet structure. There is an inevitable internal connection between these mutually derived features, so we use the idea of complementary learning to design a dual-attention intersection module to strengthen feature learning. On the other hand, in order to highlight the feature representation that reflects the changing content of two input images at different times, we utilize multiscale receptive fields for difference calculation and a contextual self-attention module to strengthen the self-learning of single changing features.

B. Attention Mechanism

The attention model first proposed by Bahdanau et al. [46] in machine translation tasks led to the concept of attention learning and is currently a research hotspot within the scope of deep neural networks [47]. From the perspective of human biological systems, attention is a complex cognitive ability that can quickly and effectively help us capture interesting and valuable information and filter out unimportant content accordingly. Similarly, attention-based methods can dynamically guide the model to filter high-value information and weaken the influence of unimportant information. It is widely used and improved because of its superior performance in natural language processing (NLP) and CV tasks. Jaderberg et al. [48] pioneered the design of a spatial transformer network (STN) considering the computational inefficiency of ordinary convolutional networks. This network can spatially transform deep convolutional features to focus on the most relevant spatial regions of input objects and task goals. STN can be flexibly inserted into convolutional models and has shown impressive performance in multiple tasks. In the CD task of remote sensing images, attention-based methods also show excellent performance. For example, Liu et al. [22] proposed a Siamese dual-task-constrained convolutional CD network, which additionally introduces a dual-attention module (DAM). DAM combines spatial attention and channel attention to explore the interdependence of features in both channel and spatial dimensions. Zheng et al. [6] proposed a high-frequency attention siamese network to achieve CD for architectural objects. In this method, a high-frequency attention block is applied to amplify the high-frequency information of buildings to optimize the edge detection of objects. In addition, when ordinary attention methods deal with large high-resolution images, the correlation between pixels far away in the image tends to weaken and is only limited to a single image information association, which will also lead to the loss of important contextual information. Huang et al. [49] improved the general global attention and proposed criss-cross network for image semantic segmentation tasks. The network enhances the ability of global context aggregation for images and has more efficient computational performance by learning the context information on each pixel cross path. Song et al. [33] proposed an axial cross-attention to model the global representation in different position dimensions by axial attention and cross-attention. It is designed for semantic segmentation and remote sensing image CD tasks and achieves decent performance. The self-attention-based transformer [50], which was originally applied to NLP, has good context learning ability. It has also been proven to be capable of many CV tasks [51], [52], [53], [54], [55], [56], [57] and has become one of the hottest artificial intelligence technologies at the moment. The vision transformer proposed by Dosovitskiy et al. [58] applied the transformer to the image classification task for the first time. He divides the input image into nonoverlapping image patches and encodes them positionally, and then captures and models the global information among these image patches through multihead self-attention. Chen et al. [59] utilize a transformer-based approach to acquire semantic tokens in input images and use them to model contextual information. These tokens are finally fed back into the pixel space to refine the feature representation. Liu et al. [60] combined CNN and transformer to propose a network applied to farmland CD. After using CNN to extract features, it uses the transformer module to perform context aggregation on multiscale features to obtain effective CD effects. Li et al. [61] proposed a Cot module based on transformer self-attention. Cot contains two learning branches including static context and dynamic context feature learning, which no longer only rely on isolated queries and the relationship between key pairs, but further explore the context information of adjacent keys. Cot is flexible and can replace ordinary convolutional layers, so we deploy it in the second half of the network to strengthen the context learning of single changing features to highlight the changing parts.

III. Methodology

In this section, we elaborate our network framework and introduce the main modules contained in our method one by one: DCFM, dilated pyramid difference module and context transformer block. We illustrate how these modules are designed and what they do. In addition, we introduce the loss function used by the model and summarize our method in algorithmic form.

A. Overall Framework of the Network

The UNet structure exhibits its excellent performance in CV tasks, which can effectively combine multiscale information of features. Inspired by it and considering the potential of the connection relationship between adjacent scales, the backbone of the network we designed adopts a Siamese structure similar to UNet to further explore multiscale information. It shares weights for bitemporal inputs. We use a simple residual convolution block for different levels of feature extraction without using the backbone of the currently popular ResNet series, which allows our network to restart training and effectively reduce training parameters. For the obtained features at different levels, in order to effectively retain the context information of different scales contained in them and explore the relationship between different granularity information, we give up directly using splicing or element-level addition operations to fuse features. Instead, we utilize a DCFM during decoding to strengthen the connection of multiscale information. For the difference calculation of bitemporal features, we also consider using a multilevel strategy

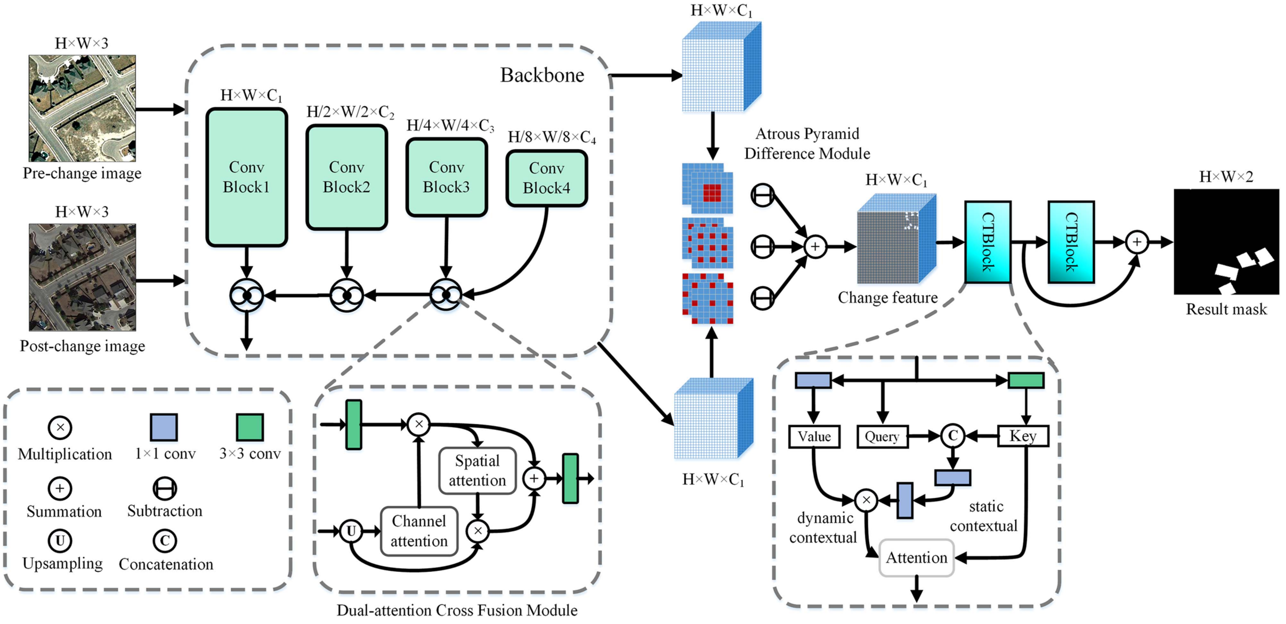

Fig. 2. Description of the overall architecture of DCFCNet. Two input images at different times are fed into a Siamese backbone and use different residual blocks to obtain multi-scale features. DCFM is then used to pairwise and reciprocally learn adjacent size feature information and fuse them. We leverage an atrous pyramid difference block to enhance the difference computation and a Cot to enrich the contextual relevance of changing features.

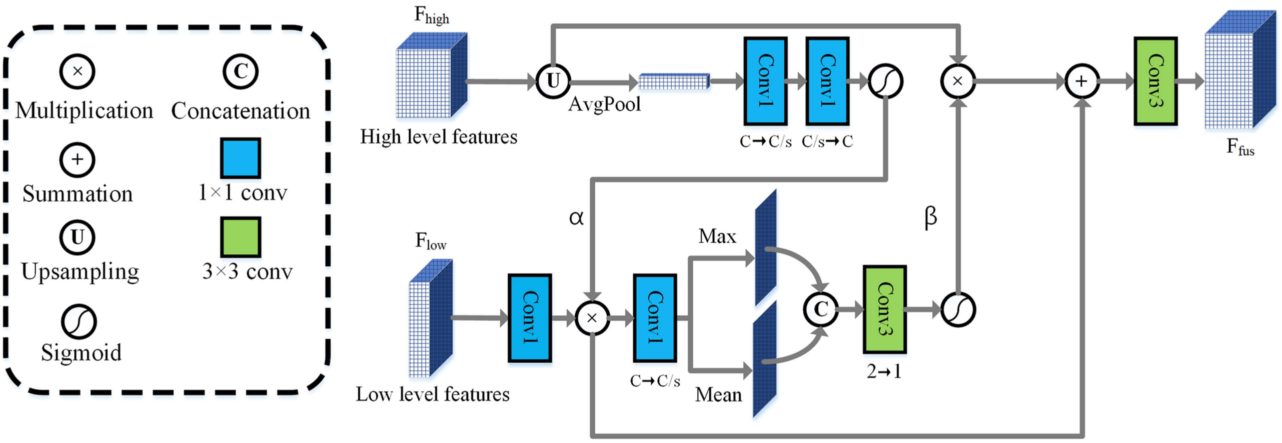

Fig. 3. Description of the DCFM.

and propose a APDM to strengthen and obtain more effective change information. After that, we introduce the context transformer block to further explore the context information of the changing features before obtaining the prediction results. A skip connection between two context transformer blocks is used to implement residual learning. The overall architecture of our network is shown in Fig. 2.

The procedure of our method can be illustrated by Algorithm 1.

B. Dual-Attention Cross Fusion Module

Low-level features contain more detailed content but lack rich global semantic information. As the features are further down-sampled and new features are learned through convolution operations, higher-level features also pay more attention to the high-level semantics of the image and lose low-grained features. However, for CD tasks, low-level feature information, such as edges and shapes, are also important factors to achieve high-precision segmentation. In order to better balance the information of different resolutions of adjacent-level features and retain meaningful information more effectively, we design a DCFM to replace ordinary concatenation and addition operations in the backbone to achieve more efficient integration. This module is shown in Fig. 3. For two input feature maps of different levels, DCFM first implements unified preprocessing to ensure that they have the same size and number of channels. Then, to learn rich information in the channel dimension of higher level features, we

Algorithm 1: Implementation Steps of DCFCNet.

```
Input: $\mathbf{I_1}, \mathbf{I_2}$ $(A\ pair\ of\ satellite\ remote\ sensing$
  $images\ at\ different\ times)$
Output: Out
// step1 : $Obtaining\ features\ through\ multi-$
  $level\ residual\ blocks$
for $i\ in\ \{1, 2\}$ do
    for $j\ in\ \{1, 2, 3, 4\}$ do
        $\mathbf{f_{i,j}} = ConvBlock(\mathbf{I_i})$
    end
// step2 : $Use\ DCFM\ to\ achieve\ efficient\ fusion$
  $of\ adjacent\ level\ features$
    $\mathbf{F'_i} = \mathbf{F_{i,4}}$
    for $k\ in\ \{3, 2, 1\}$ do
        $\mathbf{F'_i} = DCFM(\mathbf{F'_i}, \mathbf{F_{i,k}})$
    end
end
// step3 : $Use\ APDM\ to\ get\ the\ change\ feature$
$\mathbf{F_d} = APDM(\mathbf{F'_1}, \mathbf{F'_2})$
// step4 : $Use\ Cot\ to\ enhance\ contextual\ learning$
$\mathbf{F_c} = Cot(\mathbf{F_d})$
$\mathbf{F'_c} = \mathbf{F_c} + Cot(\mathbf{F_c})$
// step5 : $Use\ convolution\ to\ get\ the\ final\ result$
Out $= Conv(\mathbf{F'_c})$
```

employ a channel attention module to generate a channel weight that reweights the dimensionally expanded low-level features. A spatial attention module is used to directionally discover useful spatial information of low-level features, which also generates a spatial weight and reweights the upsampled high-level features. After cross-exchanging feature information, finally we fuse the two reattended features and restore the channel through convolution operation.

Specifically, given two inputs $F_{\text{low}} \in \mathbb{R}^{C\times H\times W}$ and $F_{\text{high}} \in \mathbb{R}^{C'\times H/r\times W/r}$, where H and W represent height and width, respectively, C and C' represent the number of channels of the two input features, and r represents the downsampling multiple. We first use a bilinear operation to upsample F_{high} to the same size as F_{low} and use a 1 × 1 convolutional layer to change the number of channels of F_{low} from C to C'. This process can be described by the following formula:

$$\begin{aligned} F'_{\text{high}} &= \text{Upsample}\left(F_{\text{high}}\right) \\ F'_{\text{low}} &= \text{Conv}_{1\times 1}\left(F_{\text{low}}\right). \end{aligned} \tag{1}$$

An average pooling operation is used to perform a Squeeze operation on F'_{high} with a dimension of $\mathbb{R}^{C'\times H\times W}$ and obtain a feature with a dimension of $\mathbb{R}^{C'\times 1\times 1}$, followed by two 1 × 1 convolution operations to realize the excitation operation of the compressed feature. In this process, the number of channels of the feature is changed from C' to C'/s, and then changed from C'/s to C' to reduce the amount of calculation, where the value of s is set to four. After this, a sigmoid function is used to generate weights α that model the correlation between feature channels. We formulate it as

$$\begin{aligned} F_{\text{se}} &= \text{AvgPool2d}(F'_{\text{high}}) \\ F_{\text{ex}} &= B((R(B(F_{\text{high}}W_1)))W_2) \\ \alpha &= \text{F}_{\text{ex}} \end{aligned} \tag{2}$$

where W_1 and W_2, respectively, represent the parameters learned by two 1 × 1 convolution layers, B(.) represents BatchNormal, and R represents the activation function ReLu. We multiply the obtained weight α and F'_{low} element-wise to let the low-level features learn the semantic responses contained in different channels and use a convolutional layer to compress the channel dimension. Next, we use mean and max operations along the channel dimension to obtain two features with dimension $\mathbb{R}^{1\times H\times W}$ and concatenate them. A 3 × 3 convolutional layer is used to compress the channel dimension from 2 to 1 and to model local spatial relationships. Likewise, we use a sigmoid function to generate a spatial weight that learns channel correlations. This process can be described as the expression

$$\begin{aligned} F_{l2} &= \text{Conv}_{1\times 1}(F'_{\text{low}} * \alpha) \\ F_{l3} &= \text{Concat}(\text{Mean}_{\text{dim}=1}(\text{F}_{\text{l}2}), \text{Max}_{\text{dim}=1}(\text{F}_{\text{l}2})) \\ \beta &= \text{Sigmoid}(\text{Conv}_{3\times 3}(\text{F}_{\text{l}3}))\,. \end{aligned} \tag{3}$$

We alternately use weights α and β to reweight F'_{low} and F'_{high} and perform element-wise summation. Finally, we obtain the fusion result after a layer of 3 × 3 convolution operation. We use the following expressions to describe:

$$\begin{aligned} F''_{\text{low}} &= F'_{\text{low}} * \alpha \\ F''_{\text{high}} &= F'_{\text{high}} * \beta \\ F_{\text{fus}} &= R(B(\text{Conv}_{3\times 3}(F''_{\text{low}} + F''_{\text{high}}))). \end{aligned} \tag{4}$$

C. Atrous Pyramid Difference Module

After obtaining the two paired bitemporal features, it is necessary to further calculate the difference information between them to obtain the changing state of the two remote sensing images at different times. Previous work on CD simply uses pixel-level subtraction or summation operations to obtain change features, which cannot fully combine the global spatial relationship to calculate more effective change information, which affects the final detection accuracy. In order to effectively extract the contextual difference content between bitemporal images, we consider using multiple convolutional layers with different dilation rates to enhance the model's perception of changing objects to improve the performance of difference detection. We thus propose an APDM to extract the variation features of the refined bitemporal features. This module obtains difference information under different receptive fields from multiple levels and finally fuses these information to obtain rich context and effective change features. As shown in Fig. 4, APDM contains four parallel branches. Specifically, we perform the same convolution operation on the two features F_{pre} and F_{post} obtained in the backbone network to obtain their local spatial connections statically. We set the kernel size of the convolutional layers of the first three parallel branches to three, and set their dilation rates to [1, 6, 12]. These

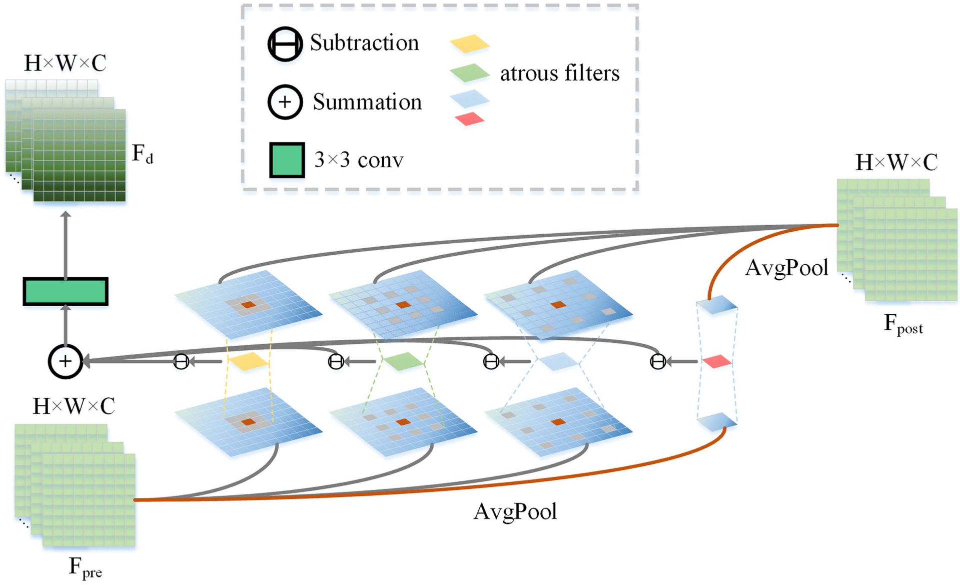

Fig. 4. Illustration for the APDM.

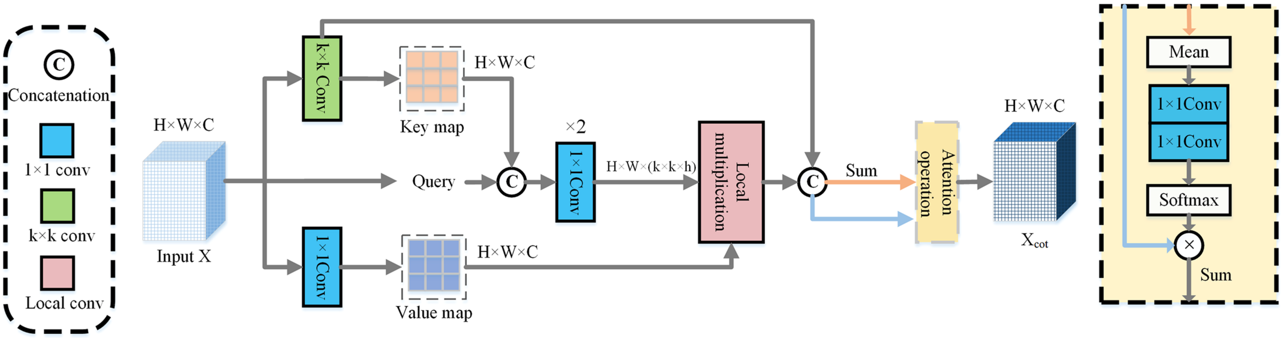

Fig. 5. Description of the Cot, which combines dynamic context and static context to enhance self-attention learning.

atrous convolutions are used to ensure as much as possible the integrity of the captured changing objects. In addition, in order to combine the global spatial information, we additionally use an average pooling branch to compress the spatial content of the bitemporal features. After that, we perform element-level subtraction and absolute value operations on the multibranch dual features to obtain multilevel difference information. Finally, an element-wise addition operation and a 3 × 3 convolutional layer are used to obtain the final changed features. This process can be expressed by the following formula:

$$\begin{aligned}F_{d1} &= \mathrm{abs}(\mathrm{Conv}_{3\times3}(F_{\mathrm{post}}) - \mathrm{Conv}_{3\times3}(F_{\mathrm{pre}}))\\ F_{d2} &= \mathrm{abs}(\mathrm{Conv}_{3\times3}^{a=6}(F_{\mathrm{post}}) - \mathrm{Conv}_{3\times3}^{a=6}(F_{\mathrm{pre}}))\\ F_{d3} &= \mathrm{abs}(\mathrm{Conv}_{3\times3}^{a=12}(F_{\mathrm{post}}) - \mathrm{Conv}_{3\times3}^{a=12}(F_{\mathrm{pre}}))\\ F_{d4} &= \mathrm{abs}(\mu(\rho(\mathrm{F}_{\mathrm{post}})) - \mu(\rho(\mathrm{F}_{\mathrm{pre}})))\\ F_d &= \mathrm{Conv}_{3\times3}(F_{d1} + F_{d2} + F_{d3} + F_{d4})\end{aligned} \quad (5)$$

where abs$(.)$ represents the absolute value operation, $\rho(.)$ represents average pooling operation, $\mu(.)$ represents upsampling operation, $\mathrm{Conv}_{3\times3}$represents the 3 × 3 convolution with BatchNormal and ReLu, and a is the atrous rate.

D. Context Transformer Block

In order to further contextually explore feature representations of changing information, we introduce a Cot before obtaining prediction results. It is adapted from the self-attention in transformer, as shown in Fig. 5. Unlike the traditional transformer's self-attention, which relies on independent query-key pair relationships, Cot utilizes the relationship between adjacent keys of the feature map to explore the contextual information of the feature for more effective dynamic self-attention learning. Specifically, for the input $X \in \mathbb{R}^{C\times H\times W}$, the acquisition method of key, query, and value is described by the following expressions:

$$K = XW_k$$

$$Q = X$$
$$V = XW_v. \tag{6}$$

Among them, $K \in \mathbb{R}^{C\times H\times W}$ represents key map, $Q \in \mathbb{R}^{C\times H\times W}$ represents queries and $V \in \mathbb{R}^{C\times H\times W}$ represents value map. Embedding matrix W_k and W_v are learned through a k × k group convolutional layer (k is set to 3) and a 1 × 1 convolutional layer, respectively. Cot uses k × k convolution to obtain a context key map, It reflects the local contextual relations between adjacent keys within a k × k spatial grid and is represented as a static context. Key map and queries are then concatenated and passed through two 1 × 1 convolution operations to obtain a dynamic self-attention matrix. This process is expressed as follows:

$$A_{kq} = (\mathrm{Concat}(\mathrm{K}, \mathrm{Q})\, W_\psi) W_\tau \tag{7}$$

where W_ψ and W_τ represent two consecutive 1 × 1 convolutional parameter matrices with ReLU activation functions. $A_{kq} \in \mathbb{R}^{H\times W\times (k\times k\times h)}$ (h indicates the number of heads, we set it to 8) represents the multihead self-attention matrix. The local attention matrix (size: k × k) of each head for each spatial position of A_{kq} is obtained through query feature and key map learning including feature context. Compared with the traditional self-attention method based on isolated query-key pairs, this additionally strengthens the context mining ability of features. After that, Cot uses local matrix multiplication just like typical self-attention [61], [62] to aggregate A_{kq} and the value map, which matrix-multiplies h local attention matrices of shape k × k at each spatial location of A_{kq} and all values in the value map with a k × k grid, respectively

$$X_{qkv} = \delta(A_{kq}, V) \tag{8}$$

where $X_{qkv} \in \mathbb{R}^{C\times H\times W}$ indicates the attended feature map and $\delta(.)$ indicates local matrix multiplication. Note that the local attention matrix corresponding to each head is only used to aggregate the value maps that are evenly partitioned along the channel dimension, and X_{qkv} is the concatenation of the aggregated feature maps of all heads. X_{qkv} is represented as a dynamic context because it dynamically learns self-relationships of features based on key and query. Cot concatenates the dynamic context representation X_{qkv} with the static context representation K and utilizes a channel attention to efficiently obtain the final result. This process is described as follows:

$$X_{\mathrm{cat}} = \mathrm{Sum}_{\mathrm{dim}=2}(\mathrm{Concat}(K, X_{qkv}))$$
$$X'_{\mathrm{cat}} = \eta(\theta(\mathrm{Mean}_{\mathrm{dim}=2,3}(X_{\mathrm{cat}})))$$
$$X_{\mathrm{cot}} = \sigma(X'_{\mathrm{cat}}) * X_{\mathrm{cat}} \tag{9}$$

where θ and η represent 1 × 1 convolutional layers with ReLU activation function, σ represents a Softmax operation and $X_{\mathrm{cot}} \in \mathbb{R}^{C\times H\times W}$ represents the final output of the context transformer block.

E. Loss Function

We use the minimized cross-entropy loss to optimize the model during the training phase, which is formally defined as follows:

$$L = \frac{1}{H_0 \times W_0} \sum_{h=1,w=1}^{H,W} l\left(P_{hw}, Y_{hw}\right) \tag{10}$$

where H_0 and W_0 represent the height and width of the original image, respectively, $l(P_{hw}, y) = -\log\left(P_{hwy}\right)$ is the cross-entropy loss, and Y_{hw} is the label of the pixel at (*h*,*w*) position.

IV. Experiments

The content of this part will introduce the remote sensing image datasets used, details of experimental implementation, evaluation indicators, comparative experiments with advanced methods, and ablation experiments of different modules.

A. Datasets

In this article, we use three public high-resolution datasets for comparison experiments and ablation experiments, including LEVR-CD [2], WHU-CD [63], and DSIFN [23]. With these datasets, we will present qualitative and quantitative experimental results to illustrate the effectiveness of our method.

LEVIR-CD [2] is a large-scale data set obtained from Google Earth and applied to remote sensing CD, which contains 637 pairs of remote sensing satellite images with high spatial resolution (0.5 m/pixel) at different times, and the size of these images is 1024× 1024. The dataset records changes in various regions and scene types between 2002 and 2018 and mainly reflects the growth and disappearance of urban buildings. We cropped the original dataset into 7120 images of 256 × 256 without overlap and divided them according to the ratio of 7:2:1 for model training/testing/validation.

WHU-CD [63] contains two aerial images with a size of 32 507 × 15 354 and a spatial resolution of 0.075 m, which were taken in 2012 and 2016, respectively. This dataset records changes in 16 077 buildings covering 20.5 km^2. We also obtained 6096/762/762 pairs of subimages with a resolution of 256 × 256 by random cropping and image enhancement for training/validation/testing.

DSIFN-CD [23] is a public data set (2 m/pixel) obtained from Google Maps. It covers six different cities in China including Beijing, Chengdu, Chongqing, Shenzhen, Xi'an, and Wuhan. The paired bitemporal images in this dataset mainly describe the changing conditions of land cover, such as roads, farmland, and buildings. We first cropped and enhanced images of cities other than Xi'an to obtain 3940 pairs of 512 × 512 resolution subimages for training and validation sets. We crop the Xi'an image into 48 subimage pairs for the test set.

B. Implementation Details and Evaluation Metrics

Implementation details: We build our models on PyTorch version 1.7 and train our models on NVIDIA RTX TITAN GPUs with 24 GB memory. To improve the generalization ability of the model to data, we perform data augmentation on the input image including random flipping, Gaussian blurring, random color dithering, random rescaling, and random cropping. We use AdamW as the optimizer and set its weight decay to 0.01 with a beta of (0.9, 0.999). The training epochs is 200, the batchsize is set to 12, and the learning rate is initially set to 0.0001, which decays to 0 through linear decay as the training progresses.

Evaluation metrics: In comparison with other advanced methods, we chose F1-score and Intersection over Union (IoU) of the change category as the primary evaluation metrics. In addition, we also used precision, recall, and overall accuracy (OA) as additional evaluation metrics. The definitions of these metrics are as follows:

$$
\begin{aligned}
F1 &= 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{IoU} &= \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \\
\text{OA} &= \frac{\text{TP} + \text{PN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}
\end{aligned}
\tag{11}
$$

where TP indicates the number of true positives, TN indicates the number of true negatives, FN indicates the number of false negatives, and FP indicates the number of false positives.

C. Comparison With State-of-the-Art Methods

State-of-the-art methods: We conduct qualitative and quantitative comparisons with ten state-of-the-art methods to verify the effectiveness of our method, including four methods based on CNN and attention: FC-EF [64], DTCDSCN [22], IFNet [23] and SNUNet [65]. Six transformer-based methods: Bit [59], Changeformer [66], CropLand [60], PaFormer [67], ICIF [68] and ACABF [33].

1) *FC-EF [64]:* A network based on FCN and UNet structures. It first concatenates two input images and then uses a single branch of convolution processing to get the final result.
2) *DTCDSCN [22]:* A method based on dual attention and FCN, which adds spatial attention and channel attention to Siamese FCN structure to obtain more salient discriminative features.
3) *IFNet [23]:* A multiscale feature fusion method based on attention and FCN, which utilizes the attention module to effectively fuse the differential features of bitemporal images and achieves deep supervision by introducing a CD loss.
4) *SNUNet [65]:* It is a network based on the NestedUNet [69] structure, which uses dense connections to fuse multilevel features and uses channel attention for deep supervision.
5) *Bit [59]:* A transformer-based approach that abstracts the input image into semantic tags containing high-level concepts of regions of interest and leverages transformers to model context.

TABLE II
QUANTITATIVE COMPARISON OF OUR PROPOSED METHOD WITH OTHER STATE-OF-THE-ART METHODS ON THE LEVIR-CD DATASET(%)

Method	LEVIR-CD				
	Precision	Recall	F1	IoU	OA
FC-EF	86.91	80.17	83.40	71.53	98.39
DTCDSCN	88.53	86.83	87.67	78.05	98.77
IFNet	**94.02**	82.93	88.13	78.77	98.87
SNUNet	89.18	87.17	88.16	78.83	98.82
BiT	89.24	89.37	89.31	80.68	98.92
ChangeFormer	92.05	88.80	90.40	82.48	99.04
CropLand	89.79	87.57	88.67	79.64	98.86
PaFormer	90.75	87.23	88.95	80.10	98.87
ICIF	92.07	89.30	90.81	83.31	99.08
ACABF	91.30	88.71	89.99	81.79	98.99
DCFC(Ours)	92.71	**89.51**	**91.08**	**83.62**	**99.11**

The best ones are marked in bold.

6) *Changeformer [66]:* Transformer-based multiscale feature connection method, which also combines a multilayer perception decoder.
7) *CropLand [60]:* A CNN- and transformer-based approach that aggregates context information of multiscale features through three token encoders and token decoders built on the transformer structure.
8) *PaFormer [67]:* A transformer-based end-to-end method for building CD. It combines prior extraction and context fusion by learning prior-aware transformers
9) *ICIF [68]:* An intrascale cross-interaction and interscale feature fusion network based on transformer and CNN, which can fully exploit the potential of CNN and transformer integration.
10) *ACABF [33]:* A method combining CNN and transformer, which utilizes them to effectively combine global and local information. An axial cross-attention module is used to fuse global feature information along the height and width dimensions of the image.

Results of comparative experiments: We compared LEVIR-CD, WHU-CD, and DSIFN-CD with advanced methods, including quantitative evaluation index comparisons as given in Tables I– III and qualitative visual comparisons as shown in Figs. 6–8.

According to the results in Table II, in the LEVIR-CD dataset, DCFCNet achieves the best performance compared with other methods except precision. Among these indicators, specifically, the F1/IoU/OA of our method obtained 91.08%/83.62%/99.11% respectively, compared with the second-ranked ICIF, which increased by 0.27%/0.31%/0.03%. In comparison with bit in Recall, our method achieves 89.51%, which is 0.14% higher than BIT. In addition, in the visualization results shown in Fig. 6, DCFCNet achieves a higher accuracy, which is reflected in having fewer false positive parts (red) and false negative parts (green). Take (f) of Fig. 6 as an example. Affected by lighting conditions and environmental factors, it is difficult to distinguish the building targets above and below that will be demolished in image A, which leads to the loss of the building

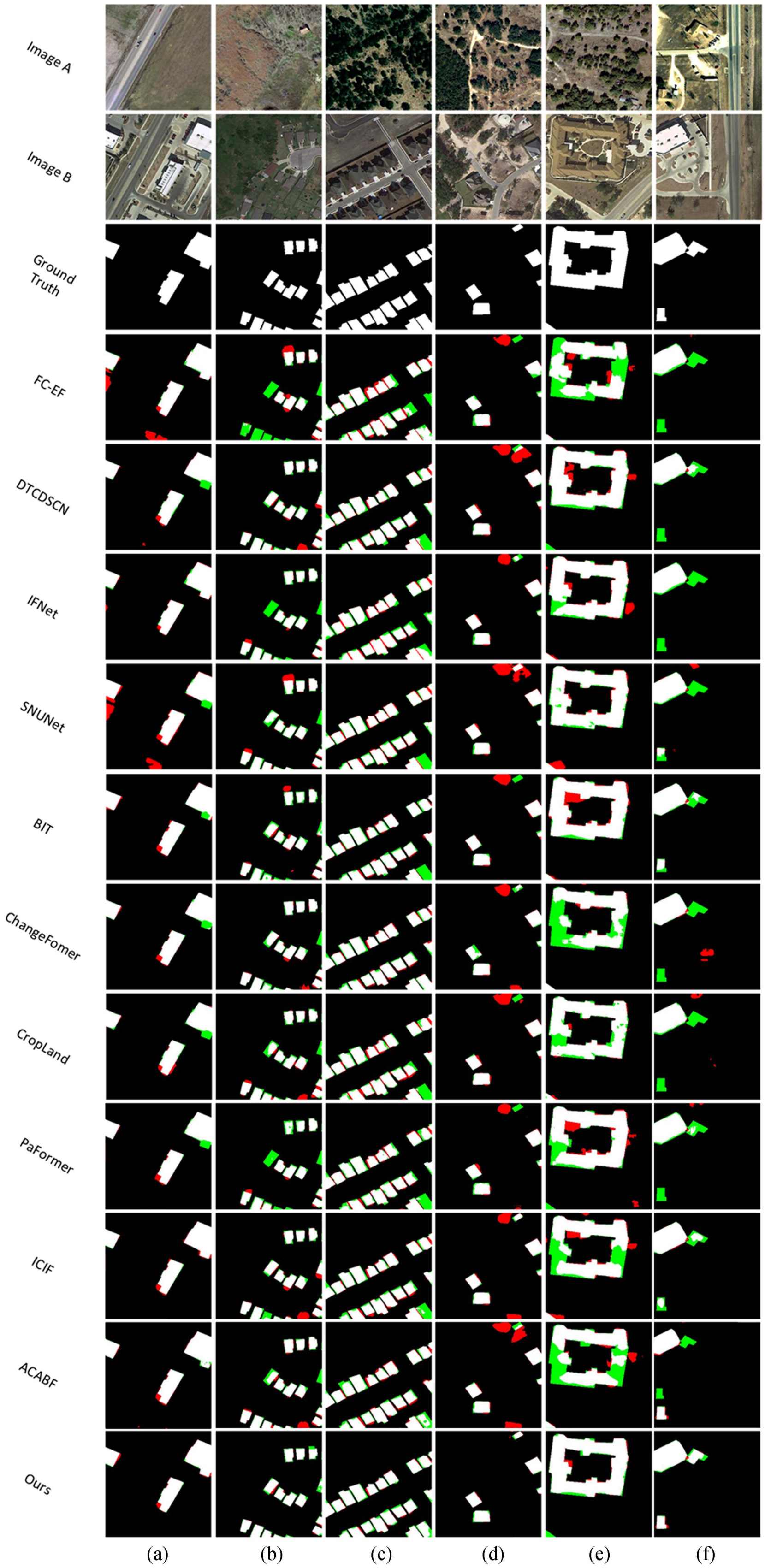

Fig. 6. Visualization of the results of our method compared with other methods on LEVIR-CD. For easier visualization, different colors are used to explain the variation, where white indicates true positives, black indicates true negatives, red indicates false positives, and green indicates false negatives.

TABLE III
QUANTITATIVE COMPARATIVE EXPERIMENTS BETWEEN OUR PROPOSED METHOD AND OTHER STATE-OF-THE-ART METHODS ON THE WHU-CD DATASET(%)

Method	WHU-CD				
	Precision	Recall	F1	IoU	OA
FC-EF	71.63	67.25	69.37	53.11	97.61
DTCDSCN	63.92	82.30	71.95	56.19	97.42
IFNet	**96.91**	73.19	83.40	71.52	98.83
SNUNet	85.60	81.49	83.50	71.67	98.71
BiT	86.64	81.48	83.98	72.39	98.75
ChangeFormer	90.50	79.61	84.51	73.18	98.59
CropLand	83.87	75.81	79.64	66.17	94.11
PaFormer	85.99	**93.77**	89.71	81.34	98.96
ICIF	92.93	88.70	90.77	83.09	99.13
ACABF	92.91	88.61	90.71	83.00	99.12
DCFC(Ours)	93.72	90.20	**91.93**	**85.06**	**99.24**

The bold entities are the best indicator value.

TABLE IV
QUANTITATIVE COMPARATIVE EXPERIMENTS BETWEEN OUR PROPOSED METHOD AND OTHER STATE-OF-THE-ART METHODS ON THE DSIFN-CD DATASET(%)

Method	DSIFN-CD				
	Precision	Recall	F1	IoU	OA
FC-EF	72.61	52.73	61.09	43.98	88.59
DTCDSCN	53.87	77.99	63.72	46.76	84.91
IFNet	67.86	53.94	60.10	42.96	87.83
SNUNet	60.60	72.89	66.18	49.45	87.34
BiT	68.36	70.18	69.26	52.97	**89.41**
ChangeFormer	74.01	76.45	75.10	62.76	85.22
CropLand	71.19	79.12	73.58	60.45	82.07
PaFormer	70.53	78.41	74.19	62.38	86.14
ICIF	75.39	79.51	77.13	64.99	86.02
ACABF	**79.50**	76.84	78.06	66.50	88.17
DCFC(Ours)	79.03	**80.67**	**79.80**	**68.39**	88.28

The bold entities are the best indicator value.

targets in image A in the results of other methods. Our method can accurately preserve these targets. On the other hand, it can be observed from Fig. 6(d) that other methods mistakenly judge the nonbuilding above the image B as the detection target. Some methods also ignore nearby building objects resulting in too many false positive and false negative regions. Our method can significantly alleviate these situations and achieve more complete detection and segmentation. These results demonstrate that our method outperforms state-of-the-art methods on CD on the LEVIR-CD dataset, probably because DCFCNet can comprehensively and effectively capture context information of images. At the same time, the edge details of the resulting mask can be improved because we complementarily learn information of different granularities from features of different resolutions.

Table III gives the quantitative comparison between DCFCNet and other methods in the WHU-CD dataset. Compared with other methods, DCFCNet has achieved a significant improvement. In the comparison of F1/IoU indicators, our method obtains 91.93%/85.06%, which is 1.16%/1.97% higher than the second best ICIF, respectively. As for OA and Recall, DCFCNet got the first ranking of 99.11% and 90.20% respectively. Fig. 7 shows qualitative visualization results on WHU-CD. Since WHU-CD mainly includes the growth and disappearance of buildings, in order to more fully demonstrate the CD capability of our method, we select some images containing objects of different sizes and styles for comparison. Fig. 7(a)–(f) shows that DCFCNet has an advantage in yOA and has better details. Fig. 7(a) shows that our method has better localization ability and can accurately detect the positions of two white rectangular buildings. In Fig. 7(b)–(f), our method shows more complete and accurate results than other methods when facing irregular objects. These results show that DCFCNet can achieve better performance than other methods on the WHU-CD dataset.

The quantitative comparison results on DSIFN-CD are given in Table IV, and DCFCNet performs best except for Precision and OA. In the comparison with F1/IoU indicators, DCFCNet achieved 79.80%/68.39%, which is 1.74%/1.89% higher than the second-ranked one. As for Recall, DCFCNet obtained 80.67%, which is 1.16% higher than the second-ranked ICIF. Fig. 8 shows the visualization results of our method and other comparative methods on the DSIFN-CD dataset. We also selected some images with different structures in the test set of DSIFN-CD with lower spatial resolution for comparison. As can be seen from Fig. 8(a)–(f), DCFCNet can achieve effective detection for both simple and complex objects and can suppress the appearance of false positive regions.

Fig. 9 presents the details of the visualization results of the qualitative experiments on the three datasets. From these results, it can be seen that our method has better edge details and can minimize the misjudgment of objects.In addition, through the test quantitative results obtained in different data sets, it can be found that compared with other methods, our method has achieved significant improvements in IoU and F1-score. However, like other state-of-the-art methods, our method ignores some subtle object changes on the DISFN-CD dataset with low spatial resolution, as shown in Fig. 8(a), (b), and (d). This may be mainly because DSIFN-CD's division is based on the entire building area as target, which makes it difficult to detect small changing objects with indistinct characteristics in complex scenes. On the other hand, LEVIR-CD and WHU-CD are aimed at different architectural targets and have more significant image details. According to Figs. 6 and 7, we can find that our method shows more excellent results for higher resolution LEVIR-CD and WHU-CD, and our method will also be applicable to the high resolution dataset of advanced remote sensing technology in the future.

D. Ablation Experiments

Our proposed method aims to enhance the contextual exploration of bitemporal images and the cross-learning of features at adjacent scales. We conduct ablation experiments on the three modules included in DCFCNet to demonstrate their effectiveness. Table V gives the quantitative ablation experimental results for DCFM, APDM, and Cot, where the baseline (NO.1) uses an addition-based fusion module and a subtraction-based difference module to replace DCFM and APDM.

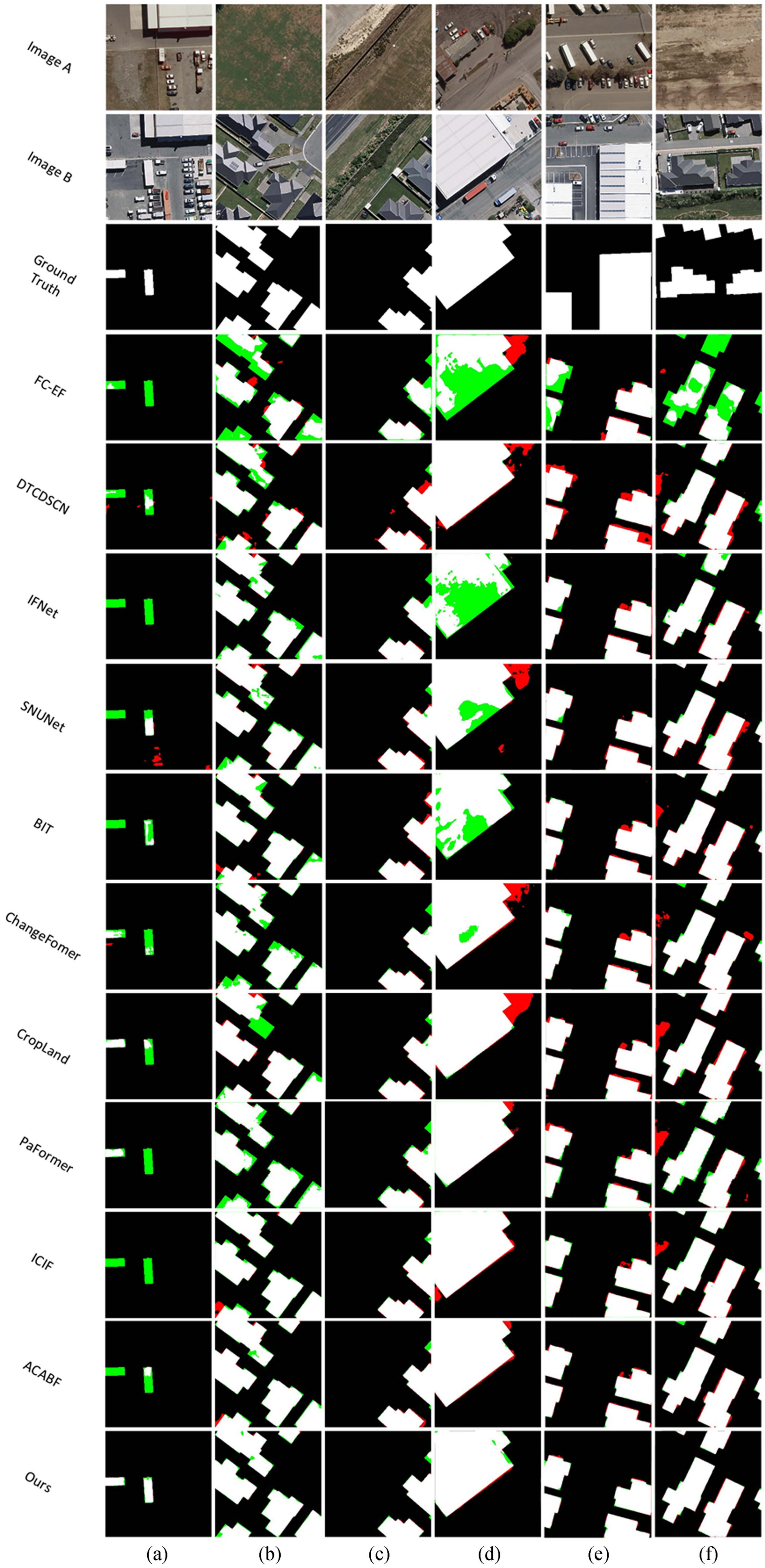

Fig. 7. Visualization of the results of our method compared with other methods on WHU-CD. For easier visualization, different colors are used to explain the variation, where white indicates true positives, black indicates true negatives, red indicates false positives, and green indicates false negatives.

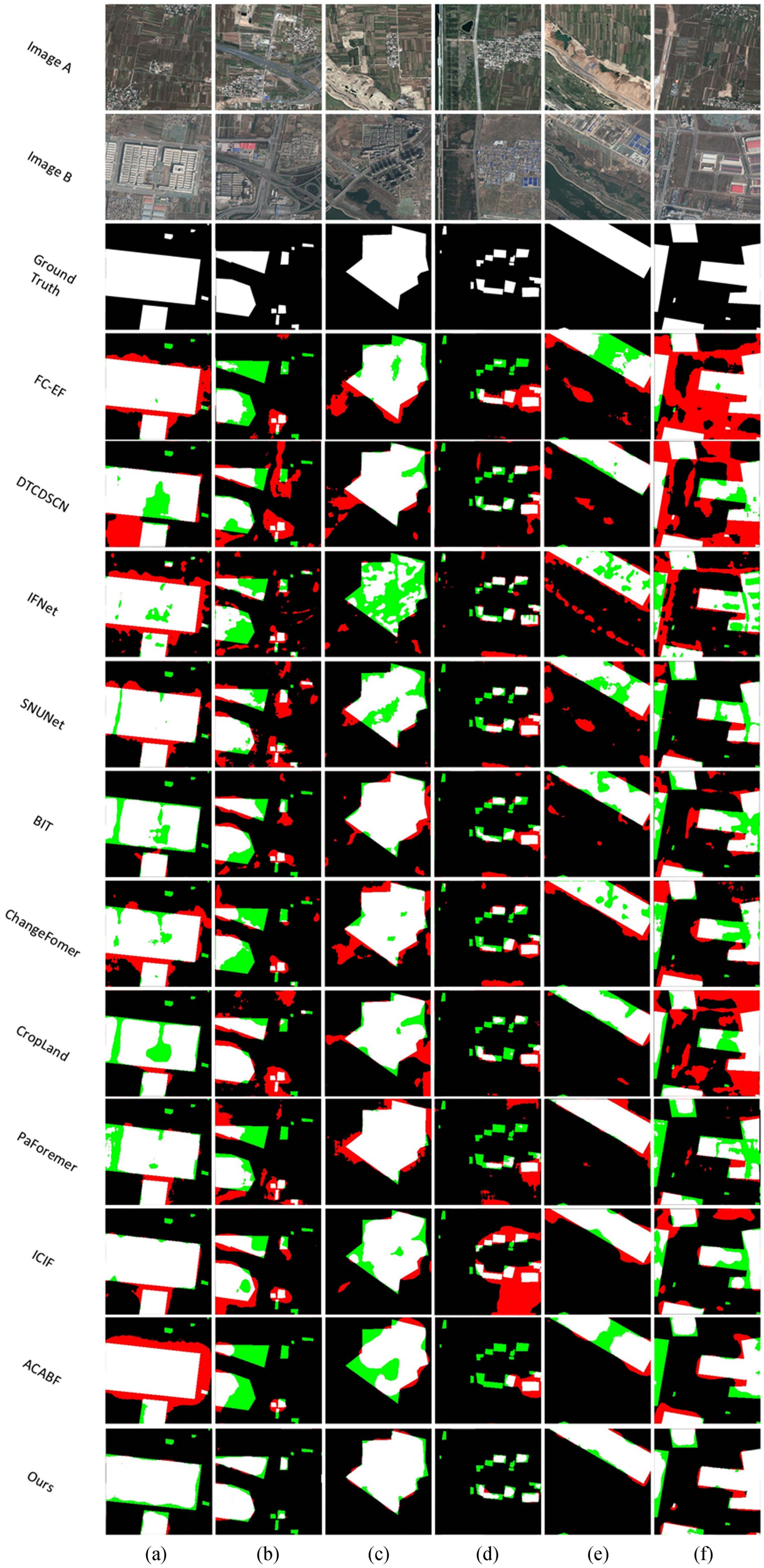

Fig. 8. Visualization of the comparison results of our method and other methods on DSIFN-CD. For easier visualization, different colors are used to explain the variation, where white indicates true positives, black indicates true negatives, red indicates false positives, and green indicates false negatives.

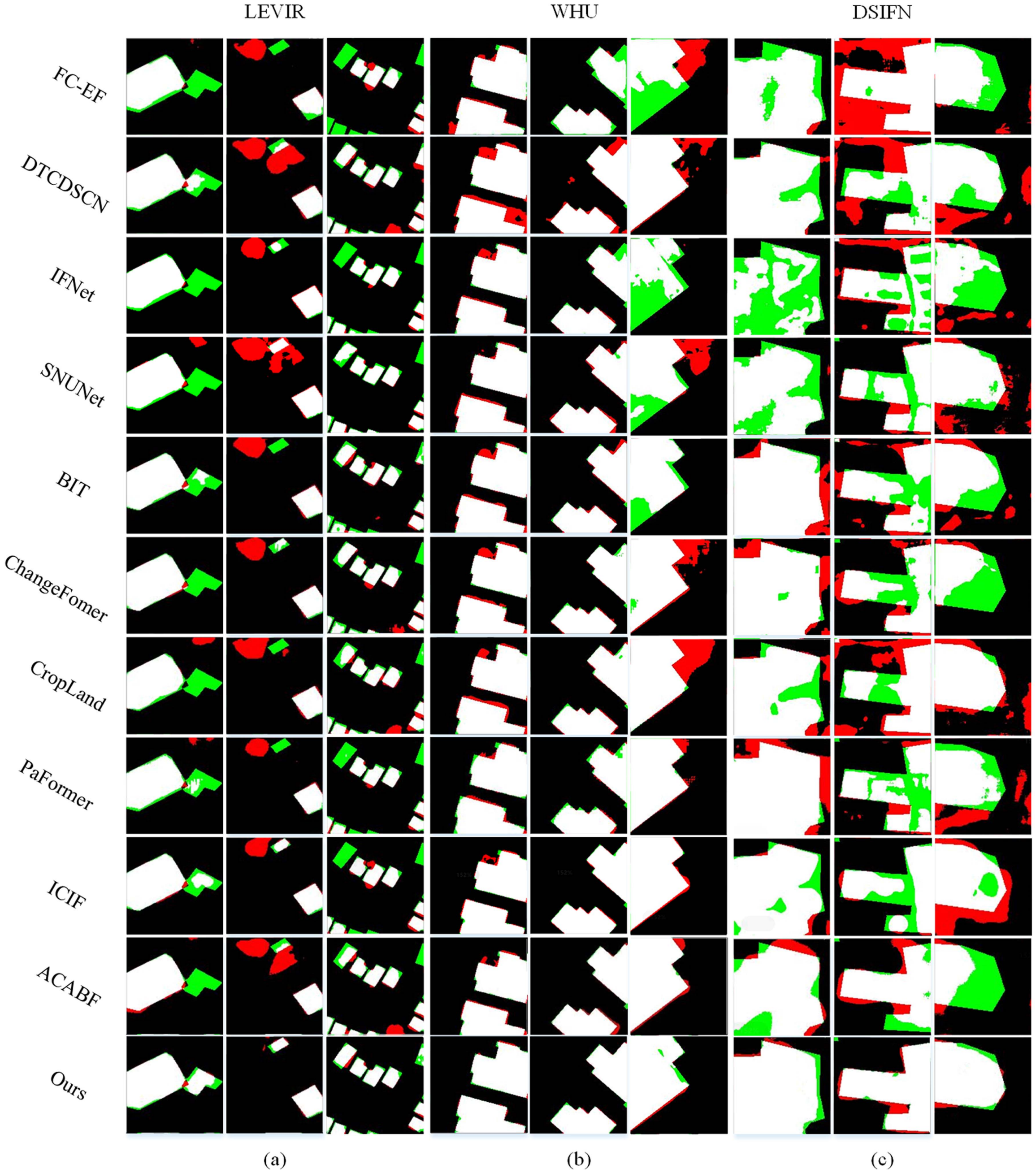

Fig. 9. Detail presentation of visualization images of different methods on three datasets. (a) Detail image on LEVIR-CD, which corresponds to (b),(d), and (f) in Fig. 6. (b) Detail image on WHU-CD, which corresponds to (c),(d), and (e) of Fig. 7. (c) Detail image on DSIFN-CD, which corresponds to (b),(c), and (f) of Fig. 8.

Ablation on DCFM: DCFM is used to fuse and decode features of two adjacent resolutions to complement each other in terms of channel and spatial dimensions. We replace DCFM with a simple fusion module based on addition operations to achieve its ablation. In this alternative fusion process, the input low-level features with larger size are expanded by a 1 × 1 convolution with ReLU activation function to expand the number of channels and the high-level features are upsampled to the same size as the low-level features by a bilinear interpolation operation. After that, an element-wise addition operation and a 3 × 3 convolutional layer are used to combine and adjust the two processed features. NO.5 in Table V gives the results of DCFCNet on three data sets after removing DCFM alone. Compared with the complete DCFCNet (NO.8), the F1-score of DCFCNet is reduced by 1.16%/3.84%/5.33% on LEVIR-CD/WHU-CD/DSIFN-CD when using common fusion modules, and the IoU is reduced by 3.27%/4.73%/6.12%, respectively. In addition, NO.2 in Table V gives the results of adding DCFM alone on the baseline

TABLE V
ABLATION EXPERIMENTS (%) OF DIFFERENT MODULES ON DIFFERENT DATA SETS IN OUR METHOD REPORT THE RESULTS OF F1-SCORE AND IOU

No.	DCFM	APDM	Cot	LEVIR-CD		WHU-CD		DSIFN-CD	
				F1	IoU	F1	IoU	F1	IoU
1	×	×	×	85.18	74.22	79.16	68.35	63.27	46.14
2	✓	×	×	87.24	76.08	84.66	73.04	67.19	50.43
3	×	✓	×	87.33	75.29	83.14	72.55	68.06	52.17
4	×	×	✓	88.17	77.16	85.91	76.38	72.12	58.62
5	×	✓	✓	89.92	80.35	88.09	81.42	74.47	62.27
6	✓	✓	×	88.82	79.57	88.67	81.93	73.18	61.59
7	✓	×	✓	90.15	81.99	90.38	83.14	76.22	64.46
8	✓	✓	✓	91.08	83.62	91.93	85.06	79.80	68.39

TABLE VI
ABLATION EXPERIMENT OF APDM

No.	Dilation interval	LEVIR-CD		WHU-CD		DSIFN-CD	
		F1	IoU	F1	IoU	F1	IoU
1	0	90.41	82.04	91.22	83.59	77.11	65.07
2	2	90.79	83.02	91.75	84.35	77.99	67.73
3	4	90.94	83.39	91.76	84.28	78.27	68.11
4	6	91.08	83.62	91.93	85.06	79.80	68.39
5	12	90.85	83.23	91.57	84.29	79.56	68.34
6	24	90.71	82.88	91.66	83.93	78.90	66.41

We tried different dilation intervals, including 0, 2, 4, 6, 12, and 24.

TABLE VII
ABLATION EXPERIMENTS (%) OF COT IN ENCODER (EC) AND DECODER (DC)

No.	Cot in EC	Cot in DC	LEVIR-CD		WHU-CD		DSIFN-CD	
			F1	IoU	F1	IoU	F1	IoU
1	×	×	88.82	79.57	88.67	81.93	73.18	61.59
2	✓	×	90.36	82.41	90.84	84.17	77.22	66.48
3	×	✓	91.08	83.62	91.93	85.06	79.80	68.39
4	✓	✓	90.88	83.01	91.52	84.25	79.66	68.40

(NO.1). F1-score and IoU increased by 2.06%/5.50%/3.92% and 1.86%/4.69%/4.29% respectively. These results show that DCFM can effectively improve the CD capability of the model.

Ablation on APDM: APDM uses multibranch atrous convolution and pooling operations to mine the difference features of dual-time features under different sizes of receptive fields, which is conducive to enhancing the difference calculation performance of the model. To verify the effectiveness of APDM, we replace it with a subtraction-based difference module to obtain variation features. This module uses element-wise subtraction and absolute value operations to compute differences and utilizes stacked convolution operations to enhance the representation of changing features. NO.7 in Table V gives the performance of our method on three datasets when ablating APDM alone. Compared with NO.8, when the model removes APDM, F1-score, and IoU are reduced by 0.93%/1.55%/3.58% and 1.63%/1.92%/3.93%, respectively, on the three data sets. On the other hand, when adding APDM alone on the baseline (NO.3), F1-score and IoU increased by 2.15%/3.98%/4.79% and 1.07%/4.20%/4.29%, respectively. In addition, we also conducted experiments on different configurations of the dilation rate of the parallel hollow convolution branch in APDM on different data sets, aiming to select the most suitable configuration parameters. As given in Table VI, we experimented with different dilation intervals, including 0, 2, 4, 6, 12, and 24. According to the results in Table VI, when the dilation rate is 6, the model achieves the best results on the three data sets, so we choose it as the final parameter configuration of APDM.

Ablation on Cot: Cot performs further contextual self-attention learning on the obtained change features to highlight the difference, which strengthens the dynamic context learning ability of the model. We verified its effectiveness by removing and adding Cot separately as given in NO.4 and NO.6 of Table V. When we remove Cot from DCFCNet, the F1-score and IoU are reduced by 2.26%/3.26%/6.62% and 4.05%/3.13%/6.8% on the three datasets, respectively. When Cot is added on the baseline alone (NO.4), F1-score and IoU are improved by 2.99%/6.75%/8.85% and 2.94%/8.03%/12.48% on the three datasets, respectively. In addition, we, respectively, use Cot in the encoder and decoder to perform ablation experiments on its location as shown in Table VII. This verifies that our strategy for using Cot is reasonable. These results indicate that Cot plays an important role for DCFCNet.

V. Conclusion

In this article, we propose a DCFCNet called DCFCNet applied to the task of CD in high-resolution remote sensing satellite images. In this network, we propose a novel DCFM to make up for the information loss brought by the extraction of multiscale features. DCFM complementarily helps adjacent-level features learn their respective advantages in spatial and channel dimensions, which helps to fully explore and preserve contextual information of multiscale images to enhance feature representation. To obtain the difference information of paired images to a greater extent, we also propose a novel APDM. This module realizes the feature difference calculation of different sizes of receptive fields through different dilation ratios to obtain a change feature containing multi-scale context information. A Cot is used to combine static and dynamic context information in the change feature to further enrich the discovery of contextual content of the change feature. We conduct extensive experiments on different datasets for qualitative and quantitative comparison with state-of-the-art methods. We also verify the effect of each module individually using ablation experiments. These experimental results demonstrate the effectiveness of our method.

References

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

[2] 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.

[3] J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, "Building damage detection in satellite imagery using convolutional neural networks," 2019, arXiv:1910.06444.

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

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

[6] H. Zheng et al., "HFA-Net: High frequency attention siamese network for building change detection in VHR remote sensing images," *Pattern Recognit.*, vol. 129, 2022, Art. no. 108717.

[7] A. Sharifi, "Development of a method for flood detection based on sentinel-1 images and classifier algorithms," *Water Environ. J.*, vol. 35, no. 3, pp. 924–929, 2021.

[8] A. Sharifi, J. Amini, and R. Tateishi, "Estimation of forest biomass using multivariate relevance vector regression," *Photogrammetric Eng. Remote Sens.*, vol. 82, no. 1, pp. 41–49, 2016.

[9] A. Sharifi, "Estimation of biophysical parameters in wheat crops in golestan province using ultra-high resolution images," *Remote Sens. Lett.*, vol. 9, no. 6, pp. 559–568, 2018.

[10] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[11] N. Zerrouki, F. Harrou, Y. Sun, and L. Hocini, "A machine learning-based approach for land cover change detection using remote sensing and radiometric measurements," *IEEE Sensors J.*, vol. 19, no. 14, pp. 5843–5850, Jul. 2019.

[12] I. Rish et al., "An empirical study of the naive bayes classifier," in *Proc. Workshop Empir. Methods Artif. Intell.*, 2001, pp. 41–46.

[13] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Proc. 10th Eur. Conf. Mach. Learn.*, Chemnitz, Germany, Springer, 1998, pp. 4–15.

[14] J. A. Suykens et al., "Least squares support vector machine classifiers: A large scale algorithm," in *Proc. Eur. Conf. Circuit Theory Des.*, Citeseer, 1999, pp. 839–842.

[15] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 388–394.

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[17] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, 2006.

[18] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May/Jun. 1991.

[19] J. Chen et al., "DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2020.

[20] 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.

[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] 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.

[23] 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.

[24] J. Li, X. Feng, and Z. Hua, "Low-light image enhancement via progressive-recursive network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4227–4240, Nov. 2021.

[25] X. Su, J. Li, and Z. Hua, "Transformer-based regression network for pan-sharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5407423.

[26] Z. Li, J. Li, F. Zhang, and L. Fan, "CADUI: Cross attention-based depth unfolding iteration network for pan-sharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5402420.

[27] X. Xu, Z. Yang, and J. Li, "AMCA: Attention-guided multiscale context aggregation network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5908619.

[28] 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.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Munich, Germany, Springer, 2015, pp. 234–241.

[30] X. Peng, R. 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.

[31] 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.

[32] Y. Zhou, C. Huo, J. Zhu, L. Huo, and C. Pan, "DCAT: Dual cross-attention-based transformer for change detection," *Remote Sens.*, vol. 15, no. 9, 2023, Art. no. 2395.

[33] L. Song, M. Xia, L. Weng, H. Lin, M. Qian, and B. Chen, "Axial cross attention meets CNN: Bibranch fusion network for change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 32–43, Nov. 2022.

[34] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, "Deep learning-based change detection in remote sensing images: A review," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 871.

[35] C. Cao, S. Dragićević, and S. Li, "Land-use change detection with convolutional neural network methods," *Environments*, vol. 6, no. 2, 2019, Art. no. 25.

[36] H. Zhang, M. Gong, P. Zhang, L. Su, and J. Shi, "Feature-level change detection using deep representation and feature change analysis for multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1666–1670, Nov. 2016.

[37] Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "An approach for unsupervised change detection in multitemporal VHR images acquired by different multispectral sensors," *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 533.

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

[39] F. Gao, X. Wang, Y. Gao, J. Dong, and S. Wang, "Sea ice change detection in SAR images based on convolutional-wavelet neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1240–1244, Aug. 2019.

[40] S. Saha, L. Mou, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Semisupervised change detection using graph convolutional network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 607–611, Apr. 2021.

[41] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, arXiv:1609.02907.

[42] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 3708–3712.

[43] O. A. Penatti, K. Nogueira, and J. A. Dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 44–51.

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

[45] M. S. Moustafa, S. A. Mohamed, S. Ahmed, and A. H. Nasr, "Hyperspectral change detection based on modification of UNet neural networks," *J. Appl. Remote Sens.*, vol. 15, no. 2, 2021, Art. no. 028505.

[46] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, arXiv:1409.0473.

[47] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, 2021, Art. no. 53.

[48] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[49] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[50] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[51] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 1691–1703.

[52] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12873–12883.

[53] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 10347–10357.

[54] B. Wu et al., "Visual transformers: Token-based image representation and processing for computer vision," 2020, arXiv:2006.03677.

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

[56] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021, Paper 1041, pp. 1–12.

[57] W. Liu, S. Chen, L. Guo, X. Zhu, and J. Liu, "CPTR: Full transformer network for image captioning," 2021, arXiv:2101.10804.

[58] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, arXiv:2010.11929.

[59] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2021, Art. no. 5920416.

[60] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, May 2022.

[61] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489–1500, Feb. 2023.

[62] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 68–80.

[63] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[64] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[65] 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, Feb. 2021, Art. no. 8007805.

[66] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[67] M. Liu, Q. Shi, Z. Chai, and J. Li, "PA-former: Learning prior-aware transformer for remote sensing building change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Aug. 2022, Art. no. 6515305.

[68] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 4410213.

[69] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2018, pp. 3–11.