DAFT: Differential Feature Extraction Network Based on Adaptive Frequency Transformer for Remote Sensing Change Detection

Zhaojin Fu, Jinjiang Li, Zheng Chen, Lu Ren, and Zhen Hua

***Abstract*—Remote sensing change detection is an important research direction in the field of remote sensing. It is mainly used to focus on the changing information on the ground over a period of time, and to identify the interested change targets from it. The rapid changes in ground information due to social development undoubtedly increase the importance of change detection. Currently, change detection methods still have some shortcomings in dealing with complex targets, environmental noise, and other aspects. Therefore, we propose a differential feature extraction network based on adaptive frequency transformer for remote sensing change detection (DAFT). Adaptive frequency transformer (AFFormer) is capable of separating change targets and environments from a frequency perspective and capturing long-range dependencies between feature information through self-attention. Therefore, in DAFT, we use AFFormer as the backbone network to extract feature information from bitemporal images, enhancing our focus on change targets while obtaining richer and more detailed information. To our knowledge, this is the first time that AFFormer has been applied in the field of CD. To address the issues of missing location information of change targets and insufficient local feature correlation, DAFT proposes a differential features enhancement module in the feature reconstruction stage of change targets. In addition, DAFT uses DO-Conv to enhance pixel correlation calculation in convolutional operations, allowing the network to focus on richer information. By outputting results at different scales during the feature reconstruction stage, DAFT computes multiple losses that are summed up to guide the training process for better performance. The experimental results prove that DAFT achieves high versus mainstream networks. On LEVIR-CD the F1 is 91.814 and the IoU is 84.866; on WHU-CD the F1 is 92.085 and the IoU is 85.330; on GZ-CD the F1 is 86.065 and the IoU is 74.512.**

***Index Terms*—Attention mechanism, change detection (CD), remote sensing, transformer.**

I. Introduction

Change detection (CD) tasks aim to identify change information in bitemporal images and annotate the change information in the form of dichotomous classification for subsequent analysis by researchers. CD tasks have become an important topic in remote sensing and have received a lot of attention from researchers in recent years. With the use of specialized remote sensing satellites [1], [2], [3], the captured image information in the remote sensing field has become rich and diverse. Depending on the application scenario, the CD task can be roughly divided into important subbranches such as farmland change detection [4], [5], forest change detection [6], postdisaster building damage assessment [7], and urban change detection [8], [9].

Currently, CD methods can be divided into two categories: 1) deep learning methods [10], [11], [12]; and 2) traditional methods [13], [14], [15]. Traditional methods are mainly developed from image processing algorithms [16], [17], [18] The processing process of traditional methods is often affected by noise factors such as lighting, climate, and clouds in the images. With the continuous pursuit of image processing accuracy in the field of remote sensing research, CD methods dominated by traditional methods have gradually lost their leading position, but still serve as important processing ideas and preprocessing processes to help other methods.

Over the past forty years, traditional algorithms have made important contributions to the data processing in the field of remote sensing CD. Since bitemporal images may come from different sensors and geographic locations, image registration has become a difficult problem in CD data processing. Feature detection algorithms such as SIFT [19] and Harris-Laplace [20] and feature matching algorithms provide a theoretical basis for solving the problem of image registration.

In recent years, deep learning methods receive a lot of attention from researchers and achieve excellent results [21], [22], [23], [24], [25], [26] in the field of remote sensing. Convolutional neural networks (CNN) possess strong recognition ability for feature information and can effectively extract features within their receptive field. However, due to the limitation of the receptive field of the convolutional kernel, CNN can only establish associations between local feature information and cannot establish associations between long-distance feature information. In the task of remote sensing CD, it not only involves small

Manuscript received 30 April 2023; revised 23 May 2023; accepted 24 May 2023. Date of publication 29 May 2023; date of current version 12 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62002200, Grant 61972235, Grant 62272281, and Grant 62202268, in part by the Shandong Natural Science Foundation of China under Grant ZR2021MF107 and Grant ZR2022QF151, and in part by Yantai Science and Technology Innovation Development Plan under Grant 2022JCYJ031. *(Corresponding author: Zhen Hua.)*

Zhaojin Fu and Zhen Hua are with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai 264005, China (e-mail: 1226187850@qq.com; huazhen@sdtbu.edu.cn).

Jinjiang Li, Zheng Chen, and Lu Ren are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: lijinjiang@gmail.com; chenzheng@sdtbu.edu.cn; renlu@sdtbu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3280589

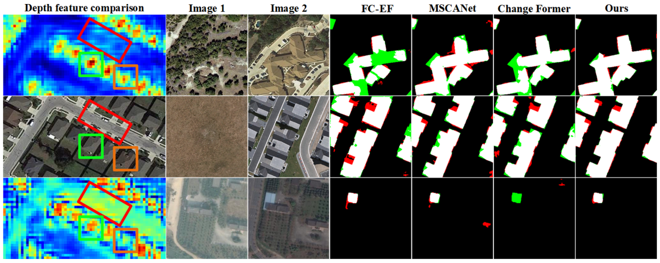


Fig. 1. Change map display. Top left side shows the high-level semantic information obtained by adaptive frequency transformer, and the bottom side shows the high-level semantic information obtained by ResNet_{34}. Right side from top to bottom shows the results obtained on LEVIR-CD, WHU-CD, and GZ-CD, respectively.

targets but also includes targets with large scales. Moreover, some targets in the samples often have features such as high density and large scope. In such sample environments, it is particularly important to establish the dependency relationships between long-distance features.

As Transformer [27] achieves excellent results in the segmentation field, it is also being applied to the remote sensing CD field. With self-attention focusing on global feature information, the Transformer architecture can better weigh the information of a single pixel in the global context to the pixel itself. However, since self-attention involves matrix multiplication of pixels, it has a high computational complexity. Therefore, reducing the computational complexity becomes the goal of efforts to apply Transformer architecture to the remote sensing CD field. In addition, some methods aim to improve the description of edge features and feature continuity by enhancing global feature attention [27], [28], [29] and expanding network views [30], [31].

However, through extensive experiments, we find that current methods still have shortcomings in edge features and feature continuity. In addition, the feature extraction of backbone networks for bitemporal images is particularly important as it helps obtain feature maps with richer detail information and more complete representation of the change targets in the prediction map.

The top left position of Fig. 1 shows the feature map of high-level semantic information obtained by DAFT, while the bottom left position shows the feature map of high-level semantic information obtained by $ResNet_{34}$. Compared to $ResNet_{34}$, the feature information captured by DAFT is more concentrated, with less noise in the environment and the strength of nonfeature information is significantly lower than that of feature information. Additionally, FC-EF [32] adopts the U-Net architecture, in which the bitemporal images inputted are concatenated and learned by the encoder to identify the changed targets. The decoder then reconstructs the changed targets. However, as FC-EF adopts a pure convolutional structure, it can only correlate local features and cannot capture long-range dependencies. MSCANet [33] adopts the CNN-Transformer architecture and uses ResNet as the backbone network to extract the features of bitemporal images. Although the method performs well, ResNet [34] has limited feature-capturing ability and thus, fails to accurately describe the edge information in some samples. To obtain richer semantic information, Change Former [35] adopts Transformer as the backbone network. The Difference Module proposed by Change Former extracts the difference features from bitemporal images obtained by each layer of the backbone network and continuously integrates them. However, we found that Change Former has a large number of parameters and computational cost, and there are also shortcomings in target continuity, among other aspects, in experiments.

In terms of data, the use of platforms such as Global Human Settlement Layer, World Settlement Footprint evolution and the improvement of image acquisition techniques have resulted in CD data with higher accuracy, and at a larger scale. As a result, objects with complex colours and shapes are more clearly described, which in part makes the CD task more difficult.

Considering the current challenges in data processing in the field of CD, and the shortcomings of current CD methods in target edge feature processing and feature continuity, we propose DAFT. Fig. 2 shows the overall architecture of DAFT. DAFT uses AFFormer [36] as the backbone network to extract richer feature information from bitemporal images. The use of AFFormer solves the shortcomings of ResNet as the backbone network in feature extraction. To solve the problem of insufficient attention to edge feature information in the process of change target reconstruction, DAFT proposes DFEM, which first enhances the attention to local feature information by expanding the local field of view, thereby constructing richer local semantic information, and then uses Coordinate Attention (CoordAttention) to filter features and enhance attention to position information. In order to further strengthen the calculation of pixel correlations, DAFT replaces all traditional convolutions with DO-Conv [37] convolutions. To better supervise the network's reconstruction of change targets, DAFT calculates the loss for each layer of the accompanying output during the feature reconstruction phase.

In this article, our contributions are as follows.

1) The DFEM is proposed to effectively achieve the enhancement of bitemporal images feature information and extraction of difference features. DAFT uses AFFormer as a backbone network in the field of remote sensing CD. By learning the frequency information of different feature categories, AFFormer strengthens the differential features in bitemporal images.
2) The computation of pixel correlation by DAFT is enhanced using DO-Conv. A deep supervision mechanism is used to supervise the change target reconstruction process and enhance the network training process.
3) The results of the experiments on the LEVIR-CD, WHU-CD, and GZ-CD datasets show that the results achieved by DAFT in terms of edge feature processing and feature continuity are better than the current state-of-the-art methods in the field of CD.

II. Related Work

This section provides an overview of historical work on deep learning in the direction of remote sensing change detection and concludes with an introduction to the adaptive frequency transformer.

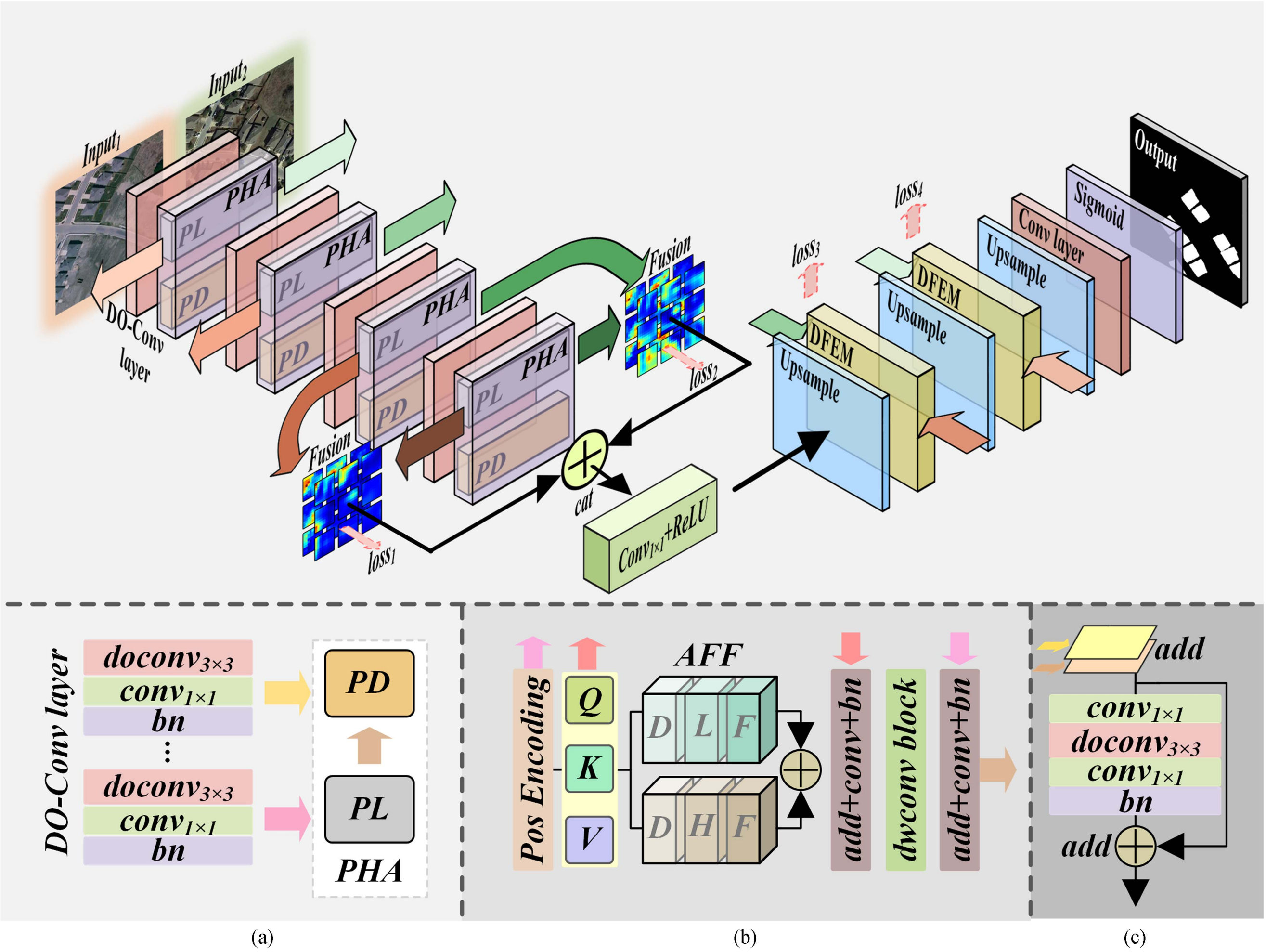


Fig. 2. Architecture diagram of DAFT: (a) Shows the structure in each layer of the backbone network, (b) shows the structure diagram for PL and (c) shows the structure diagram for pixel descriptor (PD).

A. Historical Work of Deep Learning in Remote Sensing Change Detection

In recent years, numerous excellent deep learning methods [38], [39], [40], [41], [42] have emerged in the field of CD. The “encoder–decoder” structure, named after the successful U-Net [43] model in the segmentation field, has become an important architecture in the CD field as well.

1) CNN-Based Method: FC-Siam-Conc [32] uses two identical U-Net encoders to perform feature extraction on bitemporal images. The skip connection in FC-Siam-Conc transfers the feature information produced by each layer of the two encoders to the decoder to supplement the lost features during the feature extraction process.

In contrast to FC-Siam-Conc, FC-Siam-Di [32] calculates the absolute difference value of the feature information connected by the skip connection in the same layer of the two encoders in order to enhance the attention to the difference features. This approach has played an important role in the development of CD tasks.

Since traditional convolutions are limited by the receptive field, they can only compute the relationship between pixels in a certain area, which often leads to discontinuous feature information for larger objects. Zhang et al. [39] used atrous spatial pyramid pooling (ASPP) to expand the network’s receptive field and enhance the attention to distant feature information.

2) Attention-Based Method: Although expanding convolution can improve CNN’s attention to feature information, it still cannot consider the correlation of feature information from a global perspective. Therefore, attention mechanisms [28], [29], [44], [45], [46] have been gradually applied to the field of CD and have given rise to numerous excellent algorithms [47], [48], [49], [50], [51].

IFNet [47] proposes a difference discrimination network (DDN) for extracting difference features from bitemporal images. In DDN, bitemporal features are first fused, and then channel attention is used to obtain the information weight of each channel and weight it to each pixel of the fused feature. After being processed by convolution layers, the spatial attention labels the positions where the weight of the fused feature is concentrated through pooling, thus further enhancing the strength of the difference information in the fused feature. SNUNet [49] uses a deep supervision mechanism to supervise the feature extraction process of the network’s hidden layer. It adopts the same processing idea as U-Net++ [52], but outputs multiple feature information of different scales and proposes an Ensemble Channel Attention Module to fuse multiple scale feature information, and finally obtains the prediction image.

MTCNet [53] also applies CBAM [29] and decomposes it into a spatial attention module (SAM) and a channel attention module (CAM). SAM is applied to weak features with low weights before multiscale Transformer to enhance their attention, while CAM is applied to highlight change targets in the feature reconstruction stage. However, because MTCNet uses only one set of feature maps output by ResNet, and the multiscale features obtained by MTCNet all come from this set of feature maps, the attention to feature information is not ideal.

3) Transformer-Based Method: The self-attention mechanism proposed by Transformer calculates the weights of each pixel in the feature map through matrix multiplication and applies these weights to the pixel itself, thereby achieving the attention of global semantic information. As a result, Transformer effectively improves the issue of insufficient attention to long-range dependencies in CD tasks. In recent years, Transformer has gained widespread application in CD tasks.

The self-attention proposed by Transformer can obtain richer semantic information by calculating the global correlation of pixels. BIT [54] combines CNN with Transformer, first using ResNet as the backbone network to complete feature extraction, then using the Transformer encoder to globally relate the deep semantic information, and finally obtaining the predicted image after processing through convolution layers. Although BIT achieves good results, it still has instability in handling small targets and details due to only semantic correlation of deep feature information. ACABFNet [55] compensates for the local semantic information obtained by CNN using the global semantic information obtained by Transformer and proposes an axial cross-attention to focus on the changing targets from both horizontal and vertical directions.

Inspired by Transformer's ability to capture long-range dependencies, RCDT [56] proposes the Relational Cross Attention Module (RCAM) to obtain change information in bitemporal features. RCAM abandons self-attention in favor of cross-attention, using earlier image features as "Query" and later image features as "Key" and "Value" to complete cross-attention. To address the limitations of using a standalone CNN or Transformer architecture for feature extraction, ICIF-Net [57] adopts a parallel architecture of CNN and Transformer to complete bitemporal image feature extraction and proposes the intrascale cross-interaction module to complement different feature information. However, ICIF-Net still lacks information interaction in the feature extraction stage.

Although attention mechanisms significantly improve the network's ability to focus on feature information, most attention mechanisms focus on features only from a channel perspective, neglecting to focus on pixel location information. In contrast, in the CD domain, obtaining the location information of features helps to locate discrepant features in bitemporal images.

Therefore, in DFEM, we use CoordAttention [58] to obtain position information and weight it as a weight to bitemporal images. In addition, most of the current mainstream CD methods are based on pixels and channels to enhance the focus on feature information, with little focus on frequency information. In contrast, frequency information contains information that is not focused on by vision, and the use of models to obtain different frequency information can better enhance the focus on feature information in bitemporal images. Therefore, in DAFT, AFFormer is used as the backbone network to complete the feature extraction.

B. Adaptive Frequency Transformer

AFFormer is initially proposed for semantic segmentation, which abandons the traditional "encoder–decoder" structure and removes the decoder, using a parallel architecture to accomplish semantic information extraction and feature reconstruction. AFFormer proposes Frequency Similarity Kernel (FSK), a Transformer variant with linear complexity O(n). First, the feature G is encoded with relative positional encoding through a convolutional layer to obtain feature $X \in \mathbb{R}^{(h\times w)\times C}$. FSK uses a fixed-size similarity kernel to obtain correlations between different frequency components and enhance important frequency components. *X* is then transformed into *keys K*, *values V*, and *query Q* through linear layers in FSK. As in (1), the results of the linear computation of *keys K* and *values V* are normalized by Softmax to obtain the similarity kernel $A_{i,j}$ for the frequency components, where k_i represents the frequency components of *keys K* and v_i represents the frequency components of *values V*. Finally, the similarity kernel is computed linearly with *query Q*, which in turn enhances the frequency information of different categories.

$$A_{i.j} = e^{k_i v_i^T} \Bigg/ \sum_{j=1}^{n} e^{k_i}. \tag{1}$$

Different object categories have their own unique frequency information, which is not perceivable by human vision. Obtaining frequency information can better distinguish between categories [59], [60]. In order to enhance the network's ability to distinguish category boundaries in images, AFFormer uses FSK to create adaptive frequency filter (AFF) in prototype learning (PL), as in Fig. 2. In AFF, dynamic low-pass filters (DLF) and dynamic high-pass filters (DHF) are proposed to extract frequency information from different bands, thereby obtaining low-frequency information and high-frequency information from the spatial domain.

DLF is mainly used to extract low-frequency information from semantic information. Specifically, DLF uses average pooling to process the spatial domain. After grouping the channels, different average pooling kernels are used to simulate frequency thresholds, and thus obtain low-frequency information in different frequency ranges. Finally, the frequency information is restored to the same size as X through bilinear interpolation. The low-frequency information of the mth group can be represented as in (2), where $\mathit{\Gamma}(\cdot)$ stands for adaptive average pooling, $s \times s$ stands for different pooling kernels, and v_m stands for the mth group of *values V* in FSK.

$$D_m^{lf}\left(v^m\right) = \mathit{\Gamma}_{s\times s}\left(v^m\right). \tag{2}$$

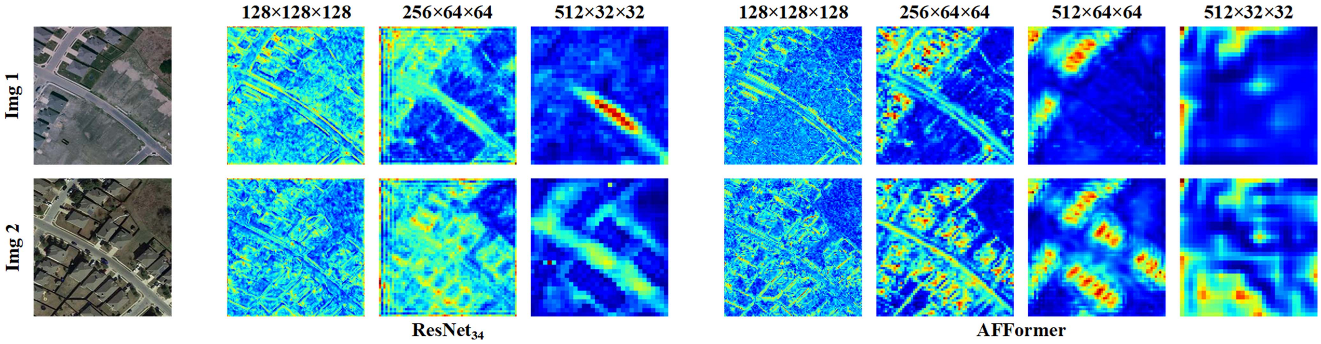


Fig. 3. Comparison of the acquisition characteristics of different backbone networks.

DHF is mainly used to extract high-frequency information, and convolution can retain information with higher intensity in pixel semantic information. Therefore, DHF uses convolutional layers to extract high-frequency semantic information. Like DLF, *values V* is also grouped, and different convolutional kernels are used to extract high-frequency semantic information from different groups, as in (3), where $\Phi(\cdot)$ represents the DO-Conv layer and $k \times k$ represents the size of the convolutional kernel.

$$D_n^{hf}\left(v_n\right) = \Phi_{k\times k}\left(v_n\right). \tag{3}$$

Finally, the frequency information extracted by the DLF as well as the DHF is superimposed on the globally enhanced frequency information obtained by FSK, as in (4).

$$AFF(X) = \sum_{cat\ H}^{i=1} D_h^{fsk}(X) + \sum_{cat\ M}^{j=1} D_m^{lf}(X) + \sum_{cat\ N}^{k=1} D_n^{hf}(X). \tag{4}$$

As in Fig. 3, we compare the feature maps obtained by $ResNet_{34}$ and AFFormer as the backbone network under the same number of training epochs. It is observed that AFFormer has a higher ability to capture feature information, and the feature information is relatively more concentrated. Therefore, in DAFT we use AFFormer as the backbone network for feature extraction of bitemporal images.

III. Method

This section focuses on the overall architecture of the network, the Differential Features Enhancement Module.

A. Network Structure

As in Fig. 2, the DAFT adopts a CNN-Transformer architecture, which can be divided into three stages. In the first stage, the feature extraction of the bitemporal images is completed, and AFFormer is used as the backbone network in this stage. The processing method of AFFormer adopts parallel heterogeneous architecture (PHA) and describes pixel semantic information through prototype learning. In each layer of AFFormer, the network first aggregates feature information $F \in \mathbb{R}^{H\times W\times C}$ using a 3 × 3 convolutional layer to generate a new pixel matrix $G \in \mathbb{R}^{h\times w\times C}$, and then feeds the feature information into the PHA. The PHA uses a parallel architecture in which the PL uses the AFF for prototype learning and constantly updates each aggregation centre to obtain $G' \in \mathbb{R}^{h\times w\times C}$. PD recovers the abstract semantic information from the PL and fuses the abstract semantic information (G') with the pixel semantic information (F).

In the second stage, we fuse the high-level semantic information from the third and fourth layers of the AFFormer. Specifically, $I_{i3} \in \mathbb{R}^{H/4\times W/4\times 512}$ is the feature information output from the third layer and $I_{i4} \in \mathbb{R}^{H/8\times W/8\times 512}$ is the feature information output from the fourth layer. I_{i3} is downsampled and then summed with I_{i4} after channel attention processing respectively, as in (5) and (6). Where σ is Sigmoid, α is ReLU, $h = H/4, w = W/4$, $i \in [1, 2]$, and $f_{1\times 1}(\cdot)$ is 1 × 1 convolution. The feature information of the bitemporal images is then fused using dimensional stitching and dimensionality reduction using 1 × 1 convolution to obtain deep semantic information D, as in (7).

$$Z_{i3} = \sigma f_{1\times 1}\left(\alpha f_{1\times 1}\left(\frac{1}{h\times w}\sum_{k=1}^{h}\sum_{j=1}^{w} I_{i3}(k,j)\right)\right) \tag{5}$$

$$X_i = Z_{i3} \otimes I_{i3} + Z_{i4} \otimes I_{i4} \tag{6}$$

$$D = \alpha f_{1\times 1}\left(cat\left(X_1, X_2\right)\right). \tag{7}$$

The third stage completes the detection of the difference features between the bitemporal images. We propose DFEM to enhance the difference features. In DFEM, CoordAttention is used to obtain the position information of the features and embed it into the pixel semantics, thereby enhancing the prediction and reconstruction of the difference feature information.

We combine the idea of deep supervision mechanism and set up accompanying outputs for the feature fusion process in the second stage and the DFEM processing in the third stage. After calculating the loss between the accompanying outputs and corresponding Ground Truth (GT) with different weights, we add them to the total loss function to better supervise the hidden layers of the network and optimize the overall training effect of the network. DO-Conv combines the calculation ideas of traditional convolution and depth convolution, focusing on more pixel correlations when calculating feature correlations.

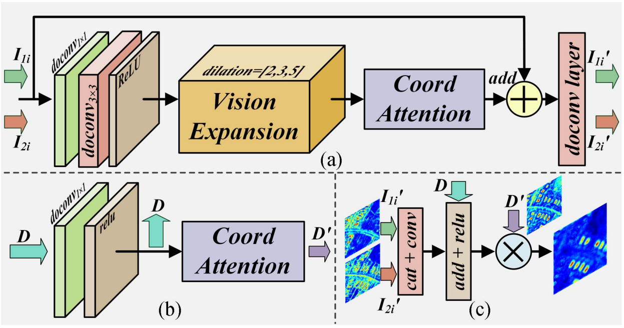


Fig. 4. Differential Features Enhancement Module.

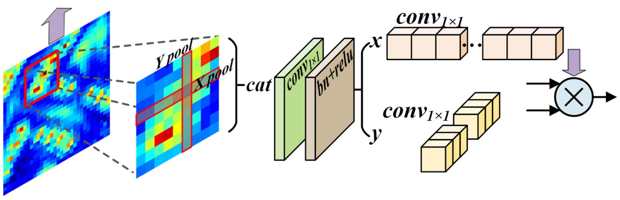


Fig. 5. Coordinate Attention.

Therefore, to achieve better training results, traditional convolution and depth convolution involved in the network are both replaced by DO-Conv.

B. Differential Features Enhancement Module

In the first and second phases of the network, the backbone network acquires rich and varying frequency of feature information from bitemporal images, which contain both change and nonchange features. The main objective of the CD task is to complete the reconstruction of the change target. Therefore, in the third stage of the network, we propose DFEM mainly for the extraction of variation features from bitemporal images, as in Fig. 4.

The semantic information captured by the shallow backbone network contains a large amount of detailed information, but some of these details have weak intensities. Therefore, the primary task of DFEM is to enhance the feature information. Specifically, in the upper-level network of DFEM [Fig. 4(a)], influenced by the idea of field expansion, we propose Vision Expansion. This module expands the receptive field of the convolutional layer by changing the distance between 3 × 3 convolutional kernel operators, allowing the network to focus on long-range dependencies in local processing. As a result, the local features are enhanced, as in (8), where $\Phi_{3\times3}^{dilation}(\cdot)$ represents the DO-Convolution with an expanded field of view.

In the process of local feature enhancement, some of the nonchange features may be similarly enhanced, and these features will ultimately affect the accuracy of the change target reconstruction. In order to be able to enhance the focus on location information while suppressing the channels where nontarget features are located. We use Coordinate Attention (Fig. 5) to process the results obtained from Vision Expansion.

Coordinate Attention uses horizontal pooling and vertical pooling to obtain horizontal features and vertical features, and encodes them separately as in (9) and (10). Then, the horizontally

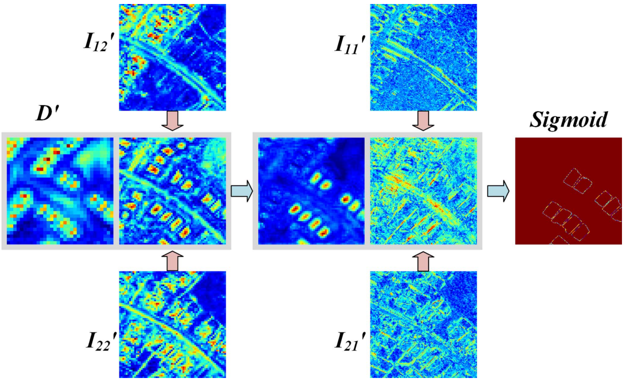


Fig. 6. Results of DFEM's focus on feature information.

and vertically pooled features containing global information are fused together through 1 × 1 convolution to obtain feature F, as in (11).Where δ is ReLU and $f_{1\times1}(\cdot)$ is 1 × 1 convolution. Feature F is decomposed into horizontal and vertical components in the spatial domain, which are processed by horizontal 1 × 1 convolution and vertical 1 × 1 convolution, respectively, to obtain horizontal and vertical weights. Finally, the weights are assigned to the original feature through multiplication, as in (12), where σ is the Sigmoid, $f_{1\times1}^{h}(\cdot)$ is the horizontal 1 × 1 convolution, $f_{1\times1}^{w}(\cdot)$ is the vertical 1 × 1 convolution, F^h is the vertical component, and F^w is the horizontal component.

$$I^{ve} = \sum_{cat\ [2,3,5]}^{i=2} \Phi_{3\times3}^{dilation=i}(I) \tag{8}$$

$$Z_I^h(h) = \frac{1}{W} \sum_{0\le i\le W} I(h,i) \tag{9}$$

$$Z_I^w(w) = \frac{1}{H} \sum_{0\le j\le H} I(j,w) \tag{10}$$

$$F = \delta f_{1\times1}\left(cat\left(Z_I^h, Z_I^w\right)\right) \tag{11}$$

$$I(i,j)' = I(i,j) \times \sigma\left(f_{1\times1}^h\left(F^h\right)\right) \times \sigma\left(f_{1\times1}^w\left(F^w\right)\right). \tag{12}$$

Since the deep semantic information D from the backbone network also contains nonchanging features, in the second stage of DFEM [Fig. 4(b)], the features of the changing target are also augmented using Coordinate Attention to obtain D'. The detailed information in the shallow features is important for the edge reconstruction of the change target, while the deep features are able to localize the change target from a macroscopic perspective. Therefore, in order to highlight the features of the change target, in the third stage [Fig. 7(c)], the features with rich detail information (I_{1i}', I_{2i}') obtained in the first stage and the features with rich semantic information (D') obtained in the second stage are fused, as in (13).

$$Out = \delta\left(cat\left(I_{1i}', I_{2i}'\right) + D\right) \times D'. \tag{13}$$

As in Fig. 6, we demonstrate the processing results of the DFEM stage. After the DFEM processing, the differential features in the spatiotemporal feature information are significantly enhanced. After being processed by the Sigmoid function, the differential features are significantly highlighted.

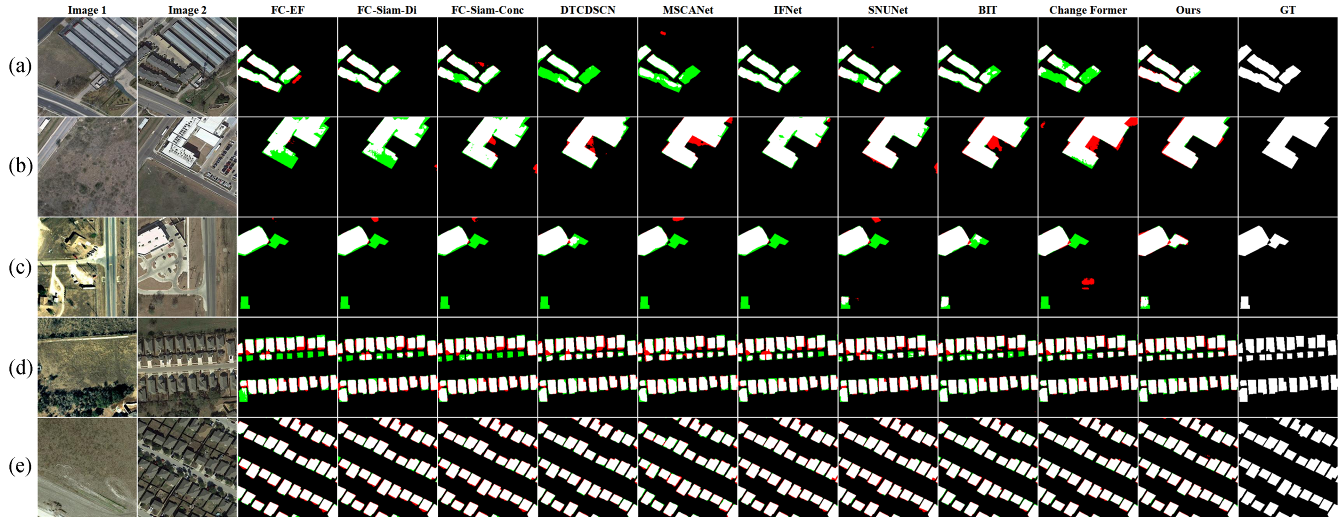


Fig. 7. Processing results of each method on LEVIR-CD.

IV. Experiment

In this section, the specific details of the experimental procedure are outlined, including: experimental parameters, comparison methods, experimental datasets, comparison experiments, and ablation experiments.

A. Experimental Parameters

All participating experimental methods are trained in a Linux environment, with computing power provided by NVIDIA TITAN RTX, Pytorch version 1.4.0, and Python version 3.8.

To ensure the comparability of experiments, the maximum number of training epochs for all comparison methods is set to 200, and the current best model is saved after each iteration. All comparison methods use 0.001 as the initial learning rate and utilize Adam as the optimizer.

We use binary cross-entropy loss as the loss function [as in (14)]. As in Fig. 2, we set accompanying outputs for the network feature fusion stage and DFEM module, and calculate the accompanying losses $Loss_1$, $Loss_2$, $Loss_3$, $Loss_4$, where $Loss_1$, $Loss_2$ are the losses of different spatiotemporal images. $Loss_3$, $Loss_4$ are the accompanying losses of the DFEM module. After experimenting with the two types of losses, we set different parameters to control their proportion in the total loss, as in (15), where θ is 0.2 and φ is 0.5.

$$L_{CE(p,\hat{p})} = -\frac{1}{WH}\sum_{x=0}^{W-1}\sum_{y=0}^{H-1} p(x,y)\log\hat{p}(x,y) + (1-p(x,y))\log(1-\hat{p}(x,y)) \tag{14}$$

$$L = \theta(Loss_1 + Loss_2) + \varphi(Loss_3 + Loss_4) + Loss. \tag{15}$$

For the objective evaluation of the performance of each network, we calculate five evaluation metrics using PyCharm: *OA* [as in (16)], *F1* [as in (17)], *IoU* [as in (18)], *Precision* [as in (19)], and *Recall* [as in (20)]. In the Equation, TP for true positive, TN for true negative, FP for false positive, and FN for false negative.

$$OA = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} \tag{16}$$

$$F1 = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{17}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{FP}+\mathrm{FN}+\mathrm{TP}} \tag{18}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{19}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{20}$$

B. Experimental Datasets

LEVIR-CD [61] is currently the most widely used and valuable dataset in the field of CD. The dataset is collected from the Google Earth platform, with a resolution of 0.5 meters/pixel and a size of 1024 × 1024. LEVIR-CD collects ground environment change images in Texas from 5 to 14 years. Due to the large time span, LEVIR-CD has a complex sample environment. Different functions of buildings lead to different sizes and shapes of the acquired building groups. This puts higher demands on the performance of the network. We divide the original images into 256 × 256 size images without overlap by default cropping method, and divide them into training set (7120), validation set (1024), and test set (2048).

WHU-CD [62] is a dataset consisting of aerial images with a size of 32507 × 15354, covering approximately 22,000 buildings in Christchurch, New Zealand, with a resolution of 0.075 meters/pixel. This dataset has a high resolution, providing clearer contour information for the buildings on the ground, but also contains more noise such as lighting and shadows. Therefore, WHU-CD presents a challenge for the network's ability to handle noise. To adapt to the GPU memory, we applied default cropping to the WHU-CD, obtaining a training set (6096), a validation set (762), and a test set (762), all with a size of 256 × 256.

GZ-CD [63] is collected using the Google Earth platform and contains VHR (Very High Resolution) images ranging in size from 1006 × 1168 to 4936 × 5224. This dataset records the environmental changes in Guangzhou, China over a period of ten years, covering buildings with different colors and having diverse samples, with a resolution of 0.55 meters/pixel. We have cropped the samples in GZ-CD into nonoverlapping image blocks of size 256 × 256 and divided them into training set (2834), validation set (400), and test set (325).

C. Comparison Method

To validate the performance of DAFT, we select nine mainstream methods that are representative of the CD. Among them, three CNN-based methods (FC-EF, FC-Siam-Conc, FC-Siam-Di), three attention-based methods (IFNet, SNUNet, DTCD-SCN), and three Transformer-based methods (MSCANet, BIT, Change Former).

FC-EF [32] uses a U-Net architecture, where feature information is fused through dimension concatenation before entering the network.

FC-Siam-Conc [32] employs a dual-encoder structure to process two spatiotemporal images separately, and fuses the difference feature with the bitemporal images in the decoder stage.

FC-Siam-Di [32] passes the bitemporal images features used for skip connections to the decoder stage by taking the absolute difference.

IFNet [47] uses a deep supervision mechanism to supervise the hidden layers of the network and provide feedback through the loss function. Channel attention and spatial attention are used to enhance the feature focus in the bitemporal images.

DTCDSCN [48] uses channel attention to enhance the backbone network's attention to feature information.

SNUNet [49] adopts the architecture of UNet++ and also uses the deep supervision mechanism to optimize the training process. An Ensemble Channel Attention Module is proposed to fuse the output features at multiple scales.

MSCANet [33] uses Transformer to parallelly process the feature information obtained by each layer of the backbone network. It adopts a processing method that combines CNN and Transformer.

BIT [54] uses a Transformer encoder to globally associate deep semantic information and reconstructs the difference features through convolutional layers.

Change Former [35] uses a Transformer as the backbone network and enhances the global correlation in semantic information.

D. Comparison Experiment

This section evaluates the performance of the network on three datasets, subjective evaluation of the network through the network output and objective evaluation of the network through the metric scores.

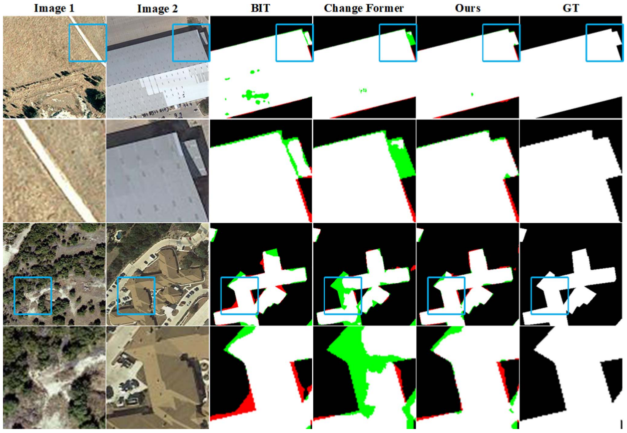


Fig. 8. Detailed comparison of processing results for large targets.

1) LEVIR-CD: A selection of the change maps obtained for each comparison method on the LEVIR dataset is in Fig. 7. We have processed the change maps by means of a visual overlay, where the red part represents the misidentified areas and the green represents the unidentified areas. The samples in the change maps include large targets, complex changing targets, dense targets, and small targets. The next analysis will be carried out from these areas.

Sample A and B both contain large buildings. For sample A, BIT and Change formerly exhibited obvious discontinuities, while DTCDSCN and MSCANet not only show feature discontinuities, but also target missing. In comparison, Ours accurately locates all the change features, although there is still room for improvement in edge judgment for this sample. However, compared to the nine contrastive methods, Ours performs the best. Due to the influence of sunlight, large shadows appear in the middle of the buildings in sample B. The results of BIT and Change Former are obviously affected by the shadow part. FC-EF and FC-Siam-Di show poor feature continuity in their output results due to the limitation of convolution's field of view. Although IFNet exhibits good edge processing results, there are still many noise points within the features. Our method not only shows good ability to counteract the effects of lighting, but also exhibits good feature continuity.

To highlight the ability to handle large buildings, we selected two sets of samples for local feature amplification comparison, as in Fig. 8. In the first sample, the building in the upper right corner is obscured by shadows due to sunlight. From the locally amplified results, Ours can effectively deal with the lighting and shadow issues. However, BIT and Change Former clearly exhibit feature separation. Additionally, there are more noise artifacts in the interior of the buildings in BIT and Change Former compared to Ours. In the second sample, the building has a complex structure and is surrounded by a complex environment. From the processing results, although ours also has some shortcomings in edge detection, compared to BIT and Change Former, ours provides a more complete description of the overall contour of the building.

Both input images in Sample C contain change targets, and the two buildings have different colors and styles. In Image 1, the

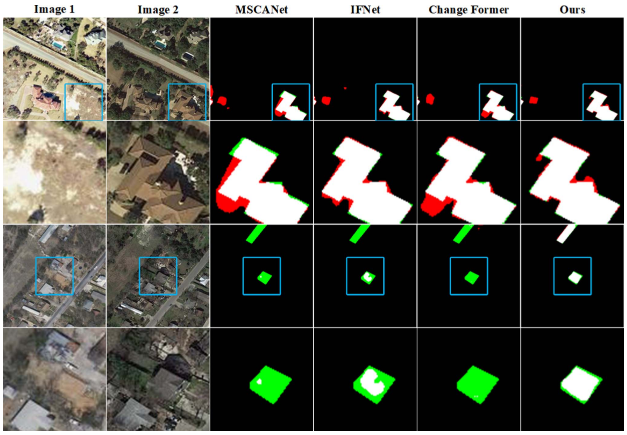


Fig. 9. Detailed comparison of processing results for small and complex targets.

TABLE I
INDICATOR RESULTS OBTAINED FOR EACH METHOD ON THE LEVIR-CD DATASET

Method	OA	F1	IoU	Precision	Recall
FC-EF	98.379	84.02	72.447	84.387	83.660
FC-Siam-Di	98.714	87.169	77.256	88.626	85.759
FC-Siam-Conc	98.688	87.254	77.389	86.374	88.151
DTCDSCN	99.002	90.090	81.967	91.183	89.023
MSCANet	98.859	88.665	79.638	89.786	87.571
IFNet	98.949	89.319	80.699	**92.568**	86.290
SNUNet	98.665	86.317	75.928	90.325	82.650
BIT	99.029	90.342	82.385	91.525	**89.189**
Change Former	**99.039**	**90.400**	**82.482**	92.054	88.805
DAFT	**99.166**	**91.814**	**84.866**	**91.843**	**91.784**

Note: Red for best results, blue for second best results.

two buildings have a similar color to their surroundings, while in Image 2, although the changed building is highlighted due to lighting effects, its color is in the same color range as the surrounding roads. From the processing results, all the comparative methods missed identifying the buildings in Image 1. On the other hand, ours, despite being affected by the environment, has identified and labeled all the buildings in the prediction map, although there are some issues with edge processing of the buildings.

In terms of the recognition of dense and small targets, we analyze it through sample D. Sample D contains a large number of dense small buildings that are highly similar to the surrounding environment. Networks such as IFNet, SNUNet, BIT, and Change Former all have varying degrees of target missing. However, Ours identifies all targets and annotates them in the final prediction map.

The feature information of small targets mainly exists in the feature information obtained by shallow networks. Ours uses a backbone network based on transformer to create an adaptive frequency Ffilter, which not only enhances the acquisition of global feature information but also obtains feature information of different categories of objects through frequency information processing, thus more effectively recognizing small targets.

To highlight the detection of small and complex targets, we select two sets of samples for local zooming comparison, as in Fig. 9. In the first sample, the buildings have complex structures. MSCANet and Change Former do not perform well in edge judgment, while ours shows good results. In the second set of samples, the volume of the buildings is small and the color depth is roughly the same as the surrounding environment, which brings considerable difficulty to the CD task. Compared with MSCANet, Change Former, and IFNet, Ours shows excellent performance.

In the following, we will objectively evaluate various networks based on their performance scores on the LEVIR dataset. As in Table I, we present the metric results of each compared method. Our DAFT achieved the best results in four metrics, namely *OA*, *F1*, *IoU*, and *Recall*, and ranked second in *Precision*. It is worth mentioning that DAFT outperforms the second-best method by 1.141 points in the *F1* metric and 2.384 points in the *IoU* metric.

Based on the comprehensive results and metrics, DAFT performs well on the LEVIR dataset. Especially in small and complex object detection, DAFT outperforms mainstream methods. However, there is still room for improvement in edge processing for some samples.

2) WHU-CD: As in Fig. 10, we also select several representative samples for display. Sample A contains a large building, and the environment in the upper right corner of the building has similar color information as the building. BIT and Change Former do not handle environmental interference as well as Ours. IFNet and SNUNet perform poorly in terms of the continuity of large targets. Although DTCDSCN and MSCANet achieve good processing results, there is still some gap between them and Ours in edge processing. In the processing of sample B, feature missing phenomena appeared in networks such as DTCDSCN, SNUNet, and Change Former. This is mainly because there are many vehicles around the buildings in the bitemporal images. These environmental factors have a certain impact on the feature discrimination of the networks. This also indicates that these networks have shortcomings in feature recognition. On the other hand, ours not only accurately detects the target but also performs well in edge processing compared to FC-Siam-Conc.

In sample C, the protruding part of the building is obscured by shadows due to lighting conditions. Networks such as SNUNet, BIT, and Change Former are clearly affected by the shadows and make recognition errors in the protruding part. In contrast, ours demonstrates excellent ability in dealing with environmental factors.

To highlight the ability to handle large buildings, we selected two additional sets of samples for local detail display, as in Fig. 11. In the first set of samples, both Image1 and Image2 contain the detection target, and the color information of the target is similar to the surrounding environment. FC-Siam-Di and MSCANet have some shortcomings in the ability to

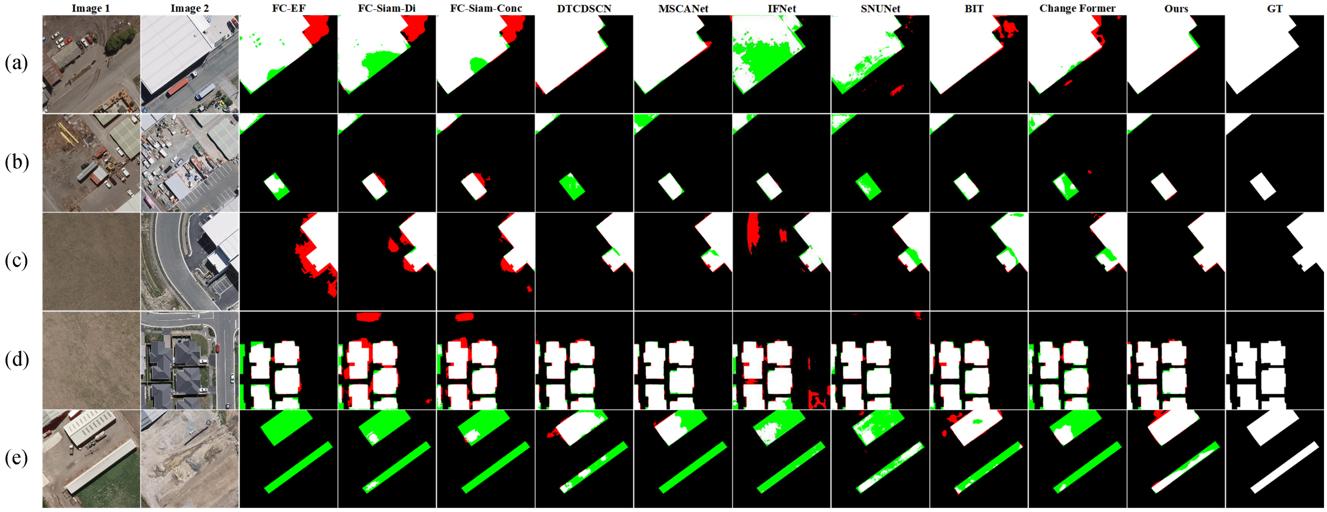


Fig. 10. Processing results of each method on WHU-CD.

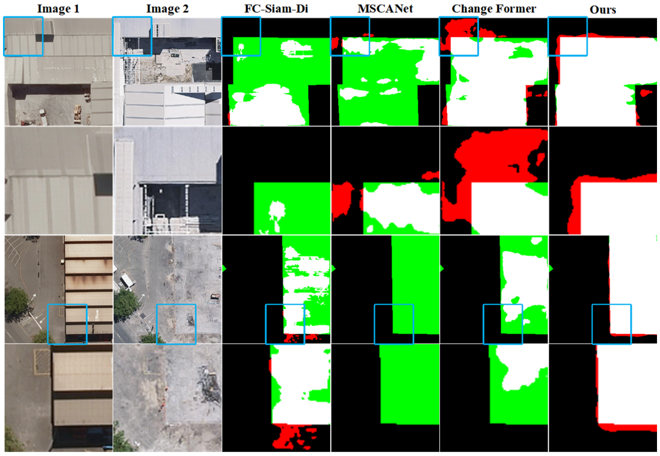


Fig. 11. Detailed comparison of processing results for large targets.

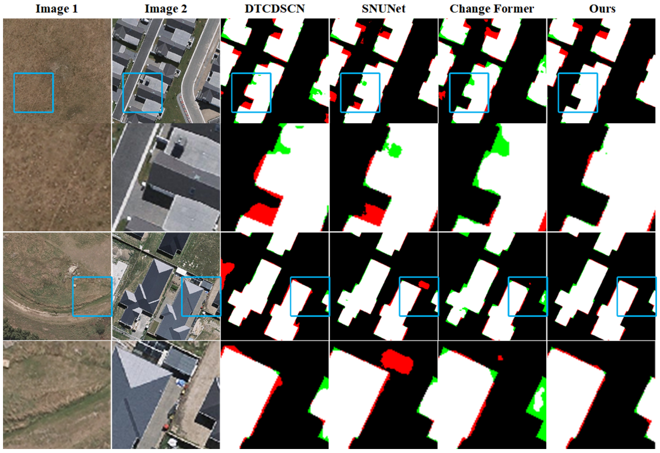


Fig. 12. Detailed comparison of treatment results for dense buildings.

discriminate features. Although Change Former has shown some effect, it is still inferior to ours in edge feature processing. In the second set of samples, Ours still performs better than the other three comparison methods. However, there is still room for improvement in the processing of edge features for ours, which will be our future direction of efforts.

In sample D, there are densely packed buildings, and due to image cropping, only a part of the building on the left is visible. FC-EF, FC-Siam-Conc, Change Former, and other networks show obvious deficiencies in recognizing the building on the left. In addition, in terms of edge feature processing, IFNet and BIT show unsatisfactory results. In contrast, ours shows good performance in both building recognition and edge feature processing, achieving results closest to the GT.

To highlight the ability to handle dense buildings, we select two groups of samples for local detail enlargement, as in Fig. 12. In the first sample, the building is shaded due to the influence of sunlight. Ours shows good ability in handling environmental factors. In contrast, DTCDSCN, SNUNet, and Change Former are all affected to varying degrees. In the second sample, Ours performs significantly better in handling edge features than SNUNet and Change Former.

Table II shows that DAFT achieves the best results in all four metrics. DAFT outperforms the second-best method by 2.054 in *F1* and 3.46 in *IoU*. Overall, DAFT demonstrates good processing capability, particularly in handling shadows, on the WHU-CD. Furthermore, our network also shows better performance in feature continuity than mainstream networks. However, DAFT still exhibits weaknesses in feature extraction for certain samples, such as in the processing of long rectangular buildings in sample E of Fig. 10.

3) GZ-CD: Fig. 13 displays partial experimental results on GZ-CD. In the processing of sample A, DTCDSCN, MSCANet, SNUNet, BIT, and ours all show good performance. Change Former and IFNet do not perform satisfactorily. For sample B, FC-Siam-Di and SNUNet exhibit less feature continuity than Ours, due to the small size of the building. Although BIT and Change Former complete the building detection more comprehensively, they do not handle edge features as well as ours due to environmental interference.

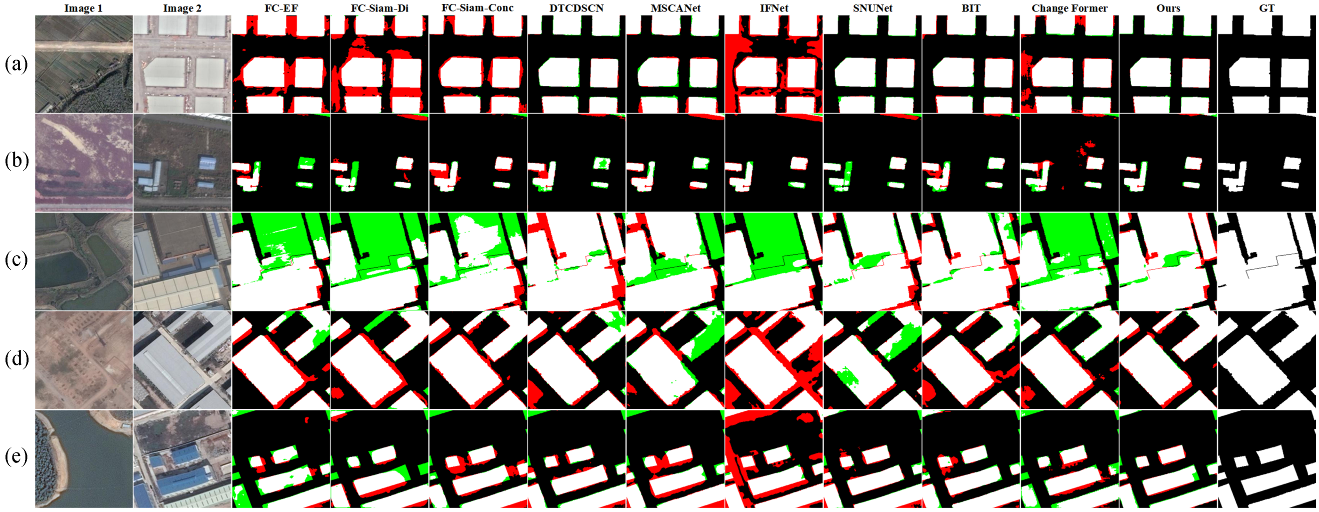


Fig. 13. Processing results of each method on GZ-CD. (a), (b), (c), (d), (e), respectively represent five different sets of bi-temporal images selected from the GZ test set to compare our method with other excellent methods.

TABLE II
Indicator Results Obtained for Each Method on the WHU-CD Dataset

Method	OA	F1	IoU	Precision	Recall
FC-EF	98.015	77.642	63.452	84.608	71.742
FC-Siam-Di	98.060	79.973	66.627	79.783	80.165
FC-Siam-Conc	98.454	84.125	72.596	83.070	85.194
DTCDSCN	99.054	90.031	81.870	91.459	88.548
MSCANet	98.954	88.619	79.564	93.164	84.497
IFNet	98.832	83.405	71.525	96.914	73.196
SNUNet	98.912	88.339	79.114	91.340	85.530
BIT	98.809	87.471	77.732	88.707	86.269
Change Former	98.758	86.882	76.806	88.495	85.326
DAFT	99.248	92.085	85.33	93.414	90.793

Note: Red for best results, blue for second best results.

In sample C, there are buildings with different colors, which requires high feature recognition ability from the network. IFNet and Change Former have missing recognition for the brown building at the top of the sample, while DTCDSCN, SNUNet, BIT, and other networks have completed the detection of all buildings, but still not as good as ours in edge feature processing.

To highlight the ability to detect large buildings, two sets of samples are reselected for local zoom-in comparison. As in Fig. 14, both sets of samples have large buildings and the buildings are closely spaced. The MSCANet, IFNet, and Change Former networks showed coarser results in terms of edge features, while ours' results are the closest to GT and there is no adhesion between the buildings.

Samples D and E both contain a large number of buildings, and there are large shadows around the buildings due to lighting conditions. DTCDSCN and Change Former have identified the buildings more completely, but both have rough handling of edge features. Although ours also has room for improvement in edge processing, it is closer to GT compared to other methods.

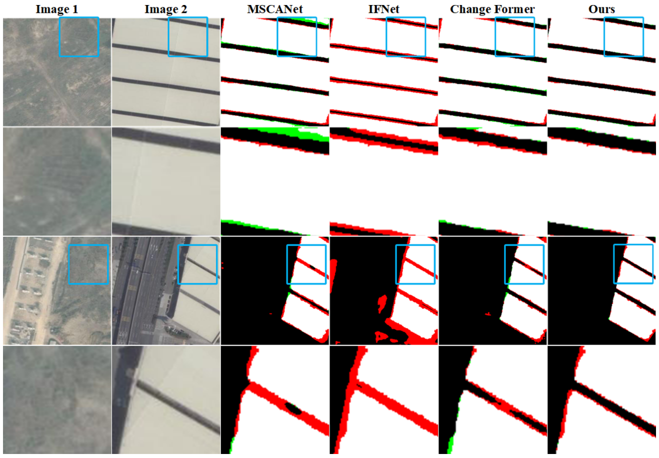


Fig. 14. Detailed comparison of processing results for large targets.

As in Fig. 15, we select samples with edge blurring and small targets for local magnification comparison. In the first sample, due to the blurry edge information of the target building, FC-Siam-Di, BIT, and Change Former all show varying degrees of feature misjudgment. Although ours is not detailed enough in edge labeling, the basic contour is roughly the same as the target. In the second sample, the target building is small in size. BIT and Change Former both miss the detection of the building, while FC-Siam-Di and Oours show better performance.

As in Table III, we show the metric scores achieved by each comparison method on the GZ-CD dataset. Although DAFT does not achieve the best score on Precision, it achieves the best on the other four metrics. DAFT is 1.815 higher than the second place on *F1*, 1.726 higher than the second place on *IoU*, and 1.774 higher than the second place on *Recall*.

Taking into account both subjective evaluation and objective results, DAFT demonstrates good performance in edge feature processing. It can accurately locate and predict targets with

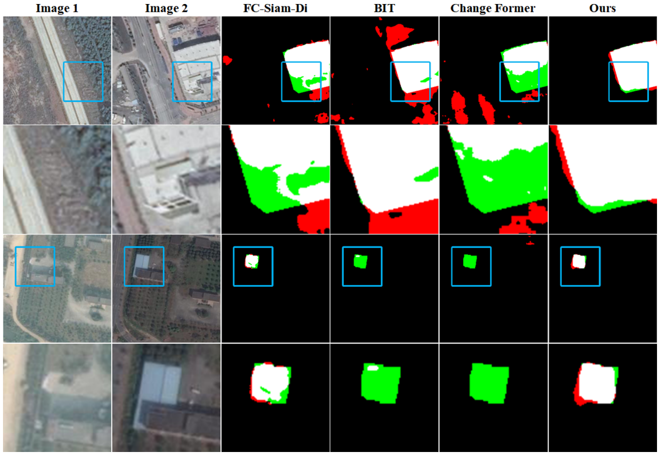


Fig. 15. Detailed comparison of processing results for small target samples and fuzzy samples.

TABLE III
Indicator Results Obtained for Each Method on the GZ-CD Dataset

Method	OA	F1	IoU	Precision	Recall
FC-EF	94.888	71.129	55.194	77.488	65.735
FC-Siam-Di	94.014	65.456	48.650	73.177	59.208
FC-Siam-Conc	95.682	76.244	61.608	80.588	72.344
DTCDSCN	96.833	82.719	70.531	**86.649**	79.131
MSCANet	96.398	80.663	67.593	83.037	78.421
IFNet	96.917	82.152	69.711	**92.194**	74.083
SNUNet	**97.069**	**84.250**	**72.786**	**86.824**	**81.824**
BIT	96.310	80.233	66.992	82.401	78.177
Change Former	95.531	73.657	58.300	84.591	65.227
DAFT	**97.188**	**86.065**	**74.512**	86.584	**83.598**

Note: Red for best results, blue for second best results.

different colors. However, DAFT still exhibits some instability in handling some complex samples in this dataset. Therefore, in future work, we will continue to explore feature extraction for complex samples.

4) Parametric Quantities and Floating Point Calculations: As in Table IV, we present the computational and model complexities of all methods participating in the comparative experiment. In terms of FLOPs, DAFT is lower than three attention-based methods and three Transformer-based methods. In addition, DAFT achieves lower parameter counts than Change Former and MSCANet, but there is still room for improvement compared to BIT. In the future, optimizing parameter counts will also be a key focus of our research.

E. Ablation Experiments

In this section, we will provide a detailed description of the ablation experiments. As in Table V, we present the specific details of each ablation module.

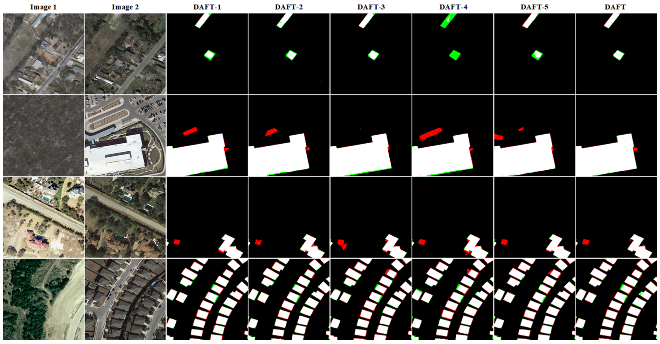


Fig. 16. Results of ablation experiments on LEVIR-CD.

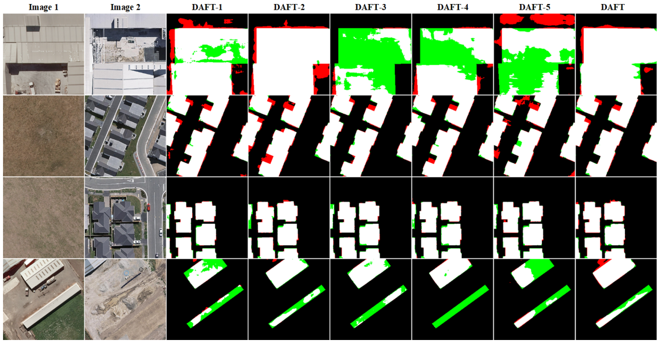


Fig. 17. Results of ablation experiments on WHU-CD.

In DAFT-1, we remove the deep supervision mechanism. In DAFT-2, AFFormer is replaced with ResNet34. Regarding DFEM, during the experiment, we propose two variants of the DFEM module, namely DFEM-1 (DAFT-3) and DFEM-2 (DAFT-4). In DFEM-1, we remove the use of the vision expansion unit. In DFEM-2, the bitemporal features from the backbone network are first fused along the channel dimension, then enter the CoordAttention, and finally complete feature enhancement through vision expansion. DAFT-5 removes DO-Conv and uses traditional convolution.

1) LEVIR-CD: Fig. 16 shows the experimental results of the five ablation models and the original model on the LEVIR dataset. In the first sample, the performance of DAFT-4 is significantly lower than that of the original model. Although the other four ablation models have detected buildings in the sample, they do not perform as well as the original model in terms of edge features. In the second sample, only DAFT-3 performs the same as the original model, while the other four ablation models have varying degrees of misidentification.

Table VI shows the performance metrics of the five ablation models and the original model on the LEVIR dataset. All five ablation models have some differences in performance metrics compared to the original model. DFEM-1 and DFEM-2, corresponding to DAFT-3 and DAFT-4, have a larger performance gap with the original model, indicating the importance and rationality of the DFEM structure.

2) WHU-CD: Fig. 17 shows partial results of the five ablation models on the WHU dataset. In the first sample, DAFT-3,

TABLE IV
PARAMETRIC QUANTITIES AND FLOATING POINT CALCULATIONS

Method	Complexity		LEVIR-CD		WHU-CD		GZ-CD	
	Params(M)	FLOPs(G)	F1	IoU	F1	IoU	F1	IoU
FC-EF	1.35	3.57	84.022	72.447	77.642	63.452	71.129	55.194
FC-Siam-Di	1.35	4.72	87.169	77.256	79.973	66.627	65.456	48.650
FC-Siam-Conc	1.55	5.32	87.254	77.389	84.125	72.596	76.244	61.608
DTCDSCN	41.07	13.21	90.090	81.967	90.031	81.870	82.719	70.531
MSCANet	16.59	14.70	88.665	79.638	88.619	79.564	80.663	67.593
IFNet	50.71	82.35	89.319	80.699	83.405	71.525	82.152	69.711
SNUNet	12.03	54.88	86.317	75.928	88.339	79.114	84.250	72.786
BIT	3.55	10.59	90.342	82.385	87.471	77.732	80.233	66.992
Change Former	41.03	202.87	90.400	82.482	86.882	76.806	73.657	58.300
DAFT	17.28	9.42	91.814	84.866	92.085	85.33	86.065	74.512

TABLE V
ABLATION EXPERIMENTAL MODEL TABLE

Method	Deep supervision	AFFormer	DFEM-1	DFEM-2	DO-Conv
DAFT-1	×	✓	×	×	✓
DAFT-2	✓	×	×	×	✓
DAFT-3	✓	✓	✓	×	✓
DAFT-4	✓	✓	×	✓	✓
DAFT-5	✓	✓	×	×	×
DAFT	✓	✓	×	×	✓

TABLE VI
ABLATION EXPERIMENTAL METRICS ON THE LEVIR-CD DATASET

Method	OA	F1	IoU	Precision	Recall
DAFT-1	99.125	91.323	84.032	92.284	92.382
DAFT-2	99.112	91.176	83.783	92.306	90.074
DAFT-3	99.142	91.479	84.295	92.58	90.403
DAFT-4	99.067	90.847	83.229	90.846	90.847
DAFT-5	99.170	91.662	84.785	92.729	90.824
DAFT	99.166	91.814	84.866	91.843	91.784

DAFT-4, and DAFT-5 perform poorly in terms of feature continuity. In the second sample, the buildings have a high similarity with the surrounding environment, and are affected by lighting, causing large shadows around the buildings, which leads to misidentification of buildings by DAFT-2 and DAFT-5.

Table VII presents the performance of the ablation models on the WHU-CD dataset. Based on the analysis of the metrics and change maps, it can be concluded that the deep supervision mechanism, AFFormer, DFEM, and DO-Conv all contribute significantly to the performance of the network.

3) GZ-CD: Fig. 18 shows change maps of the ablation experiments carried out on the GZ-CD. From the results of the whole ablation experiment, the original model is the optimal structure

TABLE VII
ABLATION EXPERIMENTAL METRICS ON THE WHU-CD DATASET

Method	OA	F1	IoU	Precision	Recall
DAFT-1	99.202	91.377	84.123	95.389	87.688
DAFT-2	99.211	91.620	84.536	93.886	89.461
DAFT-3	99.247	91.859	85.288	92.662	90.511
DAFT-4	99.154	90.971	83.438	93.669	88.424
DAFT-5	99.130	91.472	82.935	93.763	87.778
DAFT	99.248	92.085	85.33	93.414	90.793

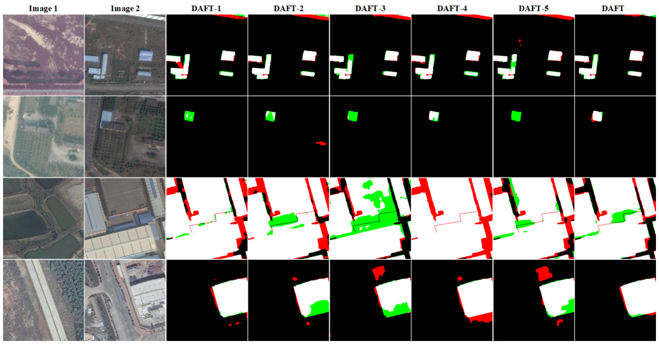


Fig. 18. Results of ablation experiments on GZ-CD.

TABLE VIII
ABLATION EXPERIMENTAL METRICS ON THE GZ-CD DATASET

Method	OA	F1	IoU	Precision	Recall
DAFT-1	97.218	85.180	74.185	86.964	83.467
DAFT-2	97.196	84.352	72.939	90.607	78.905
DAFT-3	96.905	82.191	69.766	91.574	74.552
DAFT-4	97.34	85.867	75.234	87.418	84.371
DAFT-5	96.978	83.274	71.341	88.613	78.542
DAFT	97.188	86.065	74.512	86.584	83.598

TABLE IX
PARAMETRIC QUANTITIES AND FLOATING POINT CALCULATIONS FOR ABLATION EXPERIMENT

Method	**Complexity**		**LEVIR-CD**		**WHU-CD**		**GZ-CD**	
	Params(M)	FLOPs(GMac)	F1	IoU	F1	IoU	F1	IoU
DAFT-1	17.28	9.42	91.323	84.032	91.377	84.123	85.180	74.185
DAFT-2	25.8	150.12	91.176	83.783	91.620	84.536	84.352	72.939
DAFT-3	12.15	9.36	91.479	84.295	91.859	85.288	82.191	69.766
DAFT-4	13.45	9.45	90.847	83.229	90.971	83.438	85.867	75.234
DAFT-5	11.7	30.91	91.662	84.785	91.472	82.935	83.274	71.341
DAFT	17.28	9.42	91.814	84.866	92.085	85.33	86.065	74.512

for this architecture. However, the DAFT still has room for improvement in issues such as fuzzy edge determination. As shown in Table VIII, the metrics achieved by the five ablation modules are somewhat different from the original model, which indirectly proves the reasonableness of the original model structure.

Through the ablation experiments conducted on three datasets, we have demonstrated the effectiveness of our model architecture. The application of AFFormer as the backbone network in CD has shown significant improvements in small object detection and adversarial environmental interference. The deep supervision mechanism can noticeably optimize the training process of the network. DFEM is effective in enhancing the difference features in the bitemporal images. Replacing traditional convolution with DO-Conv can improve the feature extraction ability of the CNN architecture.

4) Parametric Quantities and Floating Point Calculations: Table IX displays the parameter and computational cost of the five ablation models and the original model, as well as their performance on the three datasets. DAFT-1 has the same parameter and computational cost as the original model, but its performance is significantly reduced on all three datasets due to the cancellation of the deep supervision mechanism. DAFT-2, which uses ResNet as the backbone network, has significantly higher parameter and computational costs than the original model and other ablation models, and its performance on all datasets is significantly lower than the original model. Although DAFT-3 and DAFT-4 have significantly lower parameter costs than the original model, their results on all three datasets are not ideal. In contrast, DAFT-5 demonstrates that DO-Conv can significantly reduce model computational cost. Although this results in a slight increase in parameter cost, it also achieves better training performance.

V. LIMITATIONS AND FUTURE WORK

This article demonstrates the good performance of DAFT in feature continuity and target localization through comparative experiments. The effectiveness of the model architecture and module application was verified through ablation experiments. However, DAFT still has some shortcomings, especially in edge detection for some samples, indicating that there is still room for improvement in the robustness of DAFT. The detailed information of the changing target mainly exists in the features obtained by the shallow network, and we will focus on exploring the capturing ability of DAFT for detailed information in the future. We will consider further enhancing the features passed by the skip connections to improve the weight of weak features in the feature information.

Although DAFT achieves better results than mainstream methods on the three datasets, it still has a higher parameter and computation cost than lightweight networks such as BIT, indicating that DAFT can still be simplified in terms of its structure. In the future, we will conduct more extensive verification of the DAFT architecture to obtain the contribution of each module to feature extraction and improve performance. In terms of application, DAFT will be tested on more CD datasets to demonstrate its effectiveness and adaptability.

VI. SUMMARY

In this article, we propose DAFT, a CNN-Transformer architecture network for remote sensing CD. In the network, we apply AFFormer as the backbone network in the field of CD. AFFormer adopts a parallel architecture, abandons the “encoder–decoder” structure, creates a linear complexity variant of the Transformer, and uses it to create an AFF. Through the AFF, AFFormer obtains rich frequency information and better completes the feature extraction of bitemporal images. We propose DFEM to receive the bitemporal feature information transmitted by the backbone network and extract the difference features from it. To allow convolution calculations to focus on richer pixel information, we replace all traditional convolutions with DO-Conv. DAFT uses a deep supervision mechanism to optimize the training process, resulting in better performance. DAFT performs well in detecting ground information changes in Texas, USA; can effectively avoid the impact of environmental factors and complete the detection of ground buildings in Christchurch, New Zealand; and performs better than mainstream methods in ground information detection in Guangzhou, China. In the future, we will continue to develop more effective architectures based on DAFT to promote the development of remote sensing CD tasks.

REFERENCES

[1] L. Zhu et al., “A review: Remote sensing sensors,” in *Multi-Purposeful Application of Geospatial Data*. London: U.K.: IntechOpen, 2018, pp. 19–42.

[2] C. Ji, X. Li, H. Wei, and S. Li, “Comparison of different multispectral sensors for photosynthetic and non-photosynthetic vegetation-fraction retrieval,” *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 115.

[3] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogrammetry Remote Sens.*, vol. 115, pp. 22–36, 2016.

[4] Z. Wei, X. Gu, Q. Sun, X. Hu, and Y. Gao, "Analysis of the spatial and temporal pattern of changes in abandoned farmland based on long time series of remote sensing data," *Remote Sens.*, vol. 13, no. 13, 2021, Art. no. 2549.

[5] S. I. Elmahdy and M. M. Mohamed, "Monitoring and analysing the Emirate of Dubai's land use/land cover changes: An integrated, low-cost remote sensing approach," *Int. J. Digit. Earth*, vol. 11, no. 11, pp. 1132–1150, 2018.

[6] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5407–5423, Sep. 2017.

[7] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sens. Environ.*, vol. 265, 2021, Art. no. 112636.

[8] S. Li, "Change detection: How has urban expansion in buenos aires metropolitan region affected croplands," *Int. J. Digit. Earth*, vol. 11, no. 2, pp. 195–211, 2018.

[9] S. Z. Shahraki, D. Sauri, P. Serra, S. Modugno, F. Seifolddini, and A. Pourahmad, "Urban sprawl pattern and land-use change detection in Yazd, Iran," *Habitat Int.*, vol. 35, no. 4, pp. 521–528, 2011.

[10] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.

[11] J. Wang et al., "FWENet: A deep convolutional neural network for flood water body extraction based on SAR images," *Int. J. Digit. Earth*, vol. 15, no. 1, pp. 345–361, 2022.

[12] W. G. C. Bandara and V. M. Patel, "Revisiting consistency regularization for semi-supervised change detection in remote sensing images," 2022, *arXiv:2204.08454*.

[13] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, 1997.

[14] Y. Bazi, L. Bruzzone, and F. Melgani, "An unsupervised approach based on the generalized Gaussian model to automatic change detection in multitemporal SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 874–887, Apr. 2005.

[15] O. Hall and G. J. Hay, "A multiscale object-specific approach to digital change detection," *Int. J. Appl. Earth Observation Geoinf.*, vol. 4, no. 4, pp. 311–327, 2003.

[16] 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.

[17] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. Roy. Stat. Soc. Ser. C (Appl. Statist.)*, vol. 28, no. 1, pp. 100–108, 1979.

[18] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, pp. 293–300, 1999.

[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE 7th Int. Conf. Comput. Vis.*, 1999, pp. 1150–1157.

[20] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, pp. 63–86, 2004.

[21] J. Chen et al., "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021.

[22] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 179, pp. 14–34, 2021.

[23] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409818.

[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 pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 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, 2023, Art. no. 5402420.

[27] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Proc. Syst.*, vol. 30, no. 4, pp. 834–848, Apr. 2018.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[29] 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.

[30] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," 2022, *arXiv:2202.09741*.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[32] R. Caye 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.

[33] 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, 2022.

[34] 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.

[35] 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.

[36] B. Dong, P. Wang, and F. Wang, "Head-free lightweight semantic segmentation with linear transformer," 2023, *arXiv:2301.04648*.

[37] J. Cao et al., "Do-conv: Depthwise over-parameterized convolutional layer," *IEEE Trans. Image Process.*, vol. 31, pp. 3726–3736, 2022.

[38] 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.

[39] 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.

[40] G. Cheng, G. Wang, and J. Han, "ISNet: Towards improving separability for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623811.

[41] T. Chen, Z. Lu, Y. Yang, Y. Zhang, B. Du, and A. Plaza, "A Siamese network based U-net for change detection in high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2357–2369, 2022.

[42] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 228–239, 2022.

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

[44] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[45] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1161–1177.

[46] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 2235–2239.

[47] 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.

[48] 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.

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

[50] Z. Huang, J. Li, Z. Hua, and L. Fan, "Underwater image enhancement via adaptive group attention-based multiscale cascade transformer," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5015618.

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

[52] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet: A nested U-net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support: 4th Int. Workshop, DLMIA, 8th Int. Workshop, ML-CDS, Held Conjunction MICCAI* 2018, pp. 3–11.

[53] W. Wang, X. Tan, P. Zhang, and X. Wang, "A CBAM based multiscale transformer fusion approach for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6817–6825, 2022.

[54] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.

[55] 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, 2023.

[56] K. Lu and X. Huang, "RCDT: Relational remote sensing change detection with transformer," 2022, *arXiv:2212.04869*.

[57] 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, 2022, Art. no. 4410213.

[58] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

[59] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4504–4513.

[60] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 86–103.

[61] 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.

[62] 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.

[63] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

Zhaojin Fu received the B.S. degree in computer science and technology in 2021 from the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China, where he is currently working toward the master's degree in electronic information with the School of Information and Electronic Engineering.

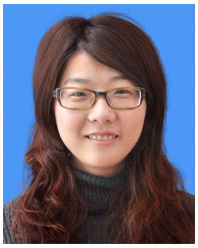
His research interests include computer graphics, computer vision, and image processing.

Jinjiang Li received the B.S. and M.S. degrees in computer science from the Taiyuan University of Technology, Taiyuan, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from Shandong University, Jinan, China, in 2010.

From 2004 to 2006, he was an Assistant Research Fellow with the Institute of Computer Science and Technology, Peking University, Beijing, China. From 2012 to 2014, he was a Postdoctoral Fellow with Tsinghua University, Beijing, China. He is currently a Professor with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China. His research interests include image processing, computer graphics, computer vision, and machine learning.

Zheng Chen received the B.S. degree in information and computational science from Shandong Agricultural University, Tai'an, China, in 2008, the M.S. degree in computer application technology from Shandong Normal University, Jinan, China, in 2015, and the Ph.D. degree in signal and information processing from the Dalian University of Technology, Dalian, China, in 2022.

He is currently a Lecturer with the Shandong Technology and Business University, Yantai, China. His research interests include computer vision, hand pose estimation, and hand shape recovery.



Lu Ren received the Ph.D. degree in signal and information processing from the Dalian University of Technology, Dalian, China, in 2021.

She is currently a Lecturer with Shandong Technology and Business University, Yantai, China. Her current research interests include sentiment analysis and text mining.

Zhen Hua received the B.S. and M.S. degrees in electrical automation from the Taiyuan University of Technology, Taiyuan, China, in 1989 and 1992, respectively, and the Ph.D. degree in electronic information engineering from the China University of Mining and Technology, Beijing, China, in 2008.

She is currently a Professor with Shandong Technology and Business University, Yantai, China. Her research interests include computer aided geometric design, information visualization, virtual reality, and image processing.