





# Axial Cross Attention Meets CNN: Bibranch Fusion Network for Change Detection

Lei Song , Min Xia , *Member, IEEE*, Ligu Weng , Haifeng Lin , Ming Qian , and Binyu Chen

**Abstract**—In the previous years, vision transformer has demonstrated a global information extraction capability in the field of computer vision that convolutional neural network (CNN) lacks. Due to the lack of inductive bias in vision transformer, it requires a large amount of data to support its training. In the field of remote sensing, it costs a lot to obtain a significant number of high-resolution remote sensing images. Most existing change detection networks based on deep learning rely heavily on the CNN, which cannot effectively utilize the long-distance dependence between pixels for difference discrimination. Therefore, this work aims to use a high-performance vision transformer to conduct change detection research with limited data. A bibranch fusion network based on axial cross attention (ACABFNet) is proposed. The network extracts local and global information of images through the CNN branch and transformer branch, respectively, and then, fuses local and global features by the bidirectional fusion approach. In the upsampling stage, similar feature information and difference feature information of the two branches are explicitly generated by feature addition and feature subtraction. Considering that the self-attention mechanism is not efficient enough for global attention over small datasets, we propose the axial cross attention. First, global attention along the height and width dimensions of images is performed respectively, and then cross attention is used to fuse the global feature information along two dimensions. Compared with the original self-attention, the structure is more graphics processing unit friendly and efficient. Experimental results on three datasets reveal that the ACABFNet outperforms existing change detection algorithms.

**Index Terms**—Attention, change detection, remote sensing image, vision transformer.

## I. INTRODUCTION

CHANGE detection of remote sensing images refers to the process of feature recognition in a collection of

multitemporal remote sensing images captured by a satellite or unmanned aerial vehicle (UAV) in the same area at various times, aiming to identify changed and unchanged areas or to identify different types of changed areas [1]. In recent years, change detection has become one of the research hotspots in the field of Earth observation, playing a very important role in urban development planning [2], disaster loss assessment [3], water body change [4], vegetation cover change monitoring [5], and other practical applications.

With the rapid advancement of remote sensing image observation techniques, the Earth observation system of space information presents six characteristics [6], [7]: high spatial resolution, high spectral resolution, high temporal resolution, multiplatform, multisensor, and multiangle, providing reliable data sources for obtaining rich spatial information of the surface and promoting the further development of change detection research. Traditional remote sensing image change detection methods occupied the mainstream prior to the rise of deep learning. The most popular method is the direct comparison method, which includes the difference method [8], the ratio method [9], and the change vector analysis method [10]. The method is straightforward to use. After image preprocessing, it just has to execute pixel-level calculations on the remote sensing images, and then, choose a suitable threshold value to separate the changed area from the unchanged area. However, the performance of the method relies heavily on circumstances, which means that change detection methods should be appropriately chosen for different scenarios. With the rapid growth of change detection technologies, object-level change detection methods [11] began to arise. The method uses geographical objects in remote sensing images as the primary classification unit, classifying them comprehensively using texture, shape, spectrum, and other factors to reduce intraclass variation and eliminate the salt and pepper effect caused by misclassification. Compared with pixel-level change detection methods, object-level change detection methods can obtain richer feature representation, better modeling image context information. However, when processing remote sensing images in different imaging environments, these algorithms cannot effectively extract the rich feature information in images, which makes it difficult to achieve high-precision detection results.

In recent years, deep learning has significantly promoted the development of semantic segmentation. Considering that this article only considers the changed category and the unchanged category, remote sensing image change detection can be regarded as binary semantic segmentation. In comparison with

Manuscript received 8 July 2022; revised 20 August 2022, 8 October 2022, and 4 November 2022; accepted 20 November 2022. Date of publication 23 November 2022; date of current version 7 December 2022. This work was supported by the National Natural Science Foundation of China under Grant 42075130. (*Corresponding author: Min Xia.*)

Lei Song, Min Xia, Ligu Weng, and Binyu Chen are with the Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: 771599073@qq.com; xiamin@nuist.edu.cn; 002311@nuist.edu.cn; icelan\_cby@nuist.edu.cn).

Haifeng Lin is with the College of Information Science and Technology, Nanjing Forestry University, Nanjing 210000, China (e-mail: haifeng.lin@njfu.edu.cn).

Ming Qian is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China (e-mail: mingqian@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3224081

traditional change detection methods, the methods based on deep learning can process remote sensing images with a vast quantity of data [12], [13]. Their capacity to characterize features is far superior to the former, and the step of manually designing feature extraction method is avoided [14]. In 2015, Long et al. [15] proposed fully convolutional networks, removing the final fully connected layer of a standard convolutional neural network (CNN) and allowing the network to output a prediction result the same size as the input. In 2017, Chen et al. [16] introduced the atrous convolution into fully convolutional networks and utilized the conditional random fields for postprocessing, which effectively enlarged the receptive field of the network without increasing parameters and optimized the segmentation boundary. In the same year, Zhao et al. [17] proposed the PSPNet, which took advantage of the pyramid pooling module to aggregate the contextual information from different areas, improving the global expression ability of the network. In 2019, Fu et al. [18] proposed the DANet, a dual-attention scene parsing network, modeling the global semantic interdependencies by self-attention mechanism instead of previous multiscale feature fusion.

In addition, Daudt et al. [19] introduced fully convolutional neural networks into the field of change detection in 2018 and proposed three change detection algorithms (FC-EF, FC-Siam-conc, and FC-Siam-diff). They used the Siamese network structure for change detection for the first time. FC-EF is an early fusion-based model, which connects bitemporal images in the channel dimension and feeds them into the fully convolutional network. FC-Siam-conc and FC-Siam-diff are both Siamese-based structures. Siamese network is used to extract the features of bitemporal images, respectively, and then, concatenation or difference operation is used to obtain the differences between them. This article mainly conducts a series of change detection research based on the early fusion method.

In 2021, Dosovitskiy et al. [20] migrated the transformer [21] from natural language processing to computer vision and proposed vision transformer, introducing a new idea in image processing. Subsequently, networks based on the vision transformer have been constantly emerging in the field of computer vision. Thanks to the local correlation and the translational invariance, the CNN can perform well on small- and medium-sized datasets [22]. The vision transformer lacks aforementioned inductive bias, so it needs to be backed up by a massive amount of data to outperform the CNN. Since the CNN can effectively model the local detailed information, while the vision transformer excels at modeling global image information, combining the CNN and vision transformer becomes a viable option. In 2021, Peng et al. [23] proposed the conformer, which used the parallel structure of the CNN and transformer, and the feature coupling unit to fuse local information and global information of images. In the same year, Guo et al. [24] proposed a hybrid series structure of the CNN and transformer, replacing the multi-layer perceptron in the transformer with convolution, achieving a balance between speed and accuracy. In addition, Srinivas et al. [25] proposed the bottleneck Transformer, which replaced the  $3 \times 3$  convolution in the bottleneck layer with multihead self-attention, significantly improving the baseline of downstream

tasks. However, considering that the aforementioned networks are basically carried out under the support of large datasets such as ImageNet [26] and COCO [27], the problem is whether the networks can maintain the same excellent performance under a small amount of data. Besides, the aforementioned methods are basically aimed at improving the convolution structure or the connection mode between the CNN and transformer, but ignore the important impact of the huge calculation amount of self-attention on the results.

In the field of remote sensing, because of the high cost of acquiring plenty of high-resolution remote sensing images, how to achieve the best performance under the premise of a small amount of data based on the vision transformer is the starting point of this article. In view of this, a bibranch fusion network based on axial cross attention (ACABFNet) is proposed in this article, the structure of which is depicted in Fig. 1. The overall structure of the network is a parallel dual-branch structure of the CNN and Transformer. The CNN is used to extract fine-grained features of images, while the transformer is used to extract global features of images. We fuse the two different features through a bidirectional interactive structure. In the upsampling stage, the overall local features and overall global features are integrated on the two branches, respectively, and the similarity and difference of the two feature information are explicitly modeled by addition and subtraction. It is worth mentioning that, considering the low efficiency of global attention of the original self-attention, we propose the axial cross attention. The axial attention is utilized to pay global attention to the images along the height and width dimensions, respectively, and then, the cross attention is used to fuse global feature information on the two dimensions. The structure is more efficient at extracting global features. In conclusion, our contributions are as follows.

- 1) A ACABFNet) is proposed. Different from existing classification-based algorithms, the network is designed for semantic segmentation and change detection tasks, fully exploiting the fine-grained and global representation characteristics of images during the downsampling and upsampling stages.
- 2) Axial cross attention is proposed. Axial attention is used to model global representation along the height and width dimensions, respectively, and cross attention is used to fuse the global feature information in the two directions. Compared with the original self-attention, the structure has higher efficiency and accuracy in capturing global feature representation and is more graphics processing unit (GPU) friendly.
- 3) Experimental results on three remote sensing image change detection datasets reveal that the ACABFNet is superior to existing change detection algorithms based on semantic segmentation.

## II. BIBRANCH FUSION NETWORK BASED ON AXIAL CROSS ATTENTION (ACABFNET)

At present, the CNN and vision transformer are two mainstream directions in the field of computer vision. Owing to the inherent inductive bias, the CNN can extract local

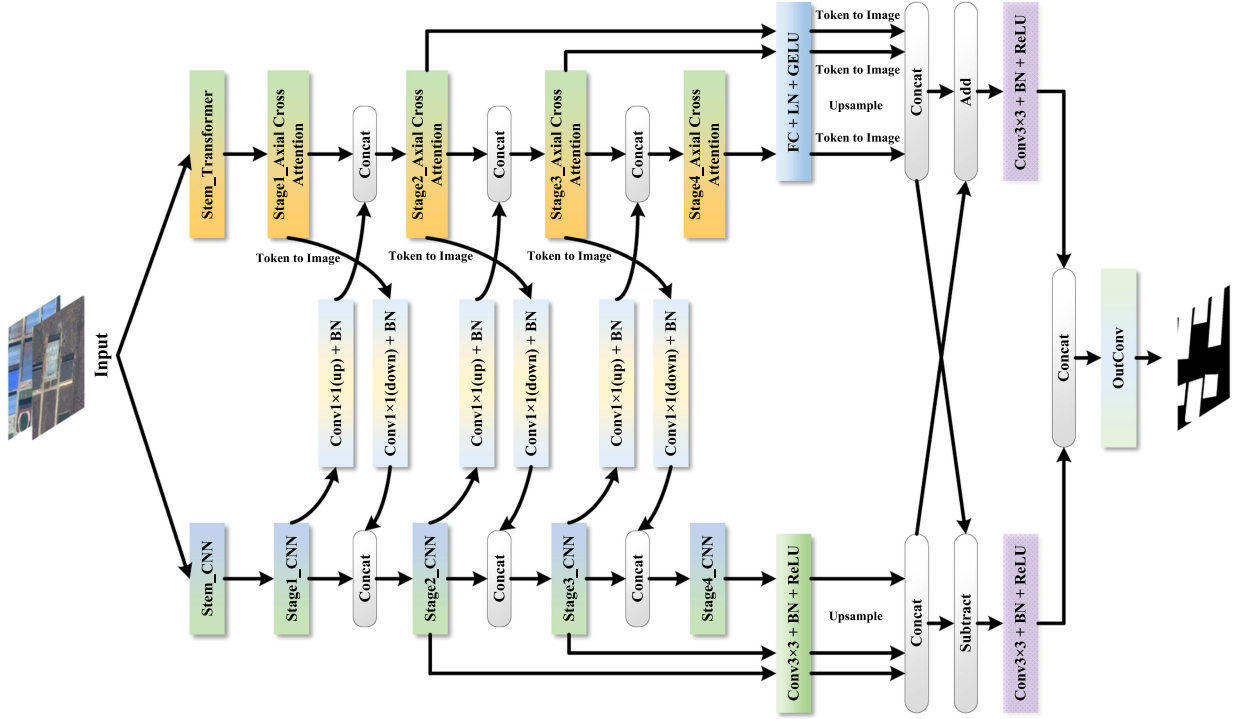


Fig. 1. Structure diagram of the ACABFNet.

TABLE I  
DETAILS OF THE ACABFNET

-	Stage	CNN Output	CNN Branch	-	Transformer Branch	Transformer Output
<b>Downsampling</b>	D1	1/4,64	[CNN Stem]	-	[Transformer Stem]	1/2,64
	D2	1/4,64	[Residual Block,64]×3	$[1 \times 1, 128] \Rightarrow$ $\Leftarrow [1 \times 1, 64]$	[Axial Cross Attention, Head=4,128]×3	1/4,128
	D3	1/8,128	[Residual Block,128]×4	$[1 \times 1, 256] \Rightarrow$ $\Leftarrow [1 \times 1, 128]$	[Axial Cross Attention, Head=8,256]×4	1/8,256
	D4	1/16,256	[Residual Block,256]×6	$[1 \times 1, 512] \Rightarrow$ $\Leftarrow [1 \times 1, 256]$	[Axial Cross Attention, Head=16,512]×6	1/16,512
	D5	1/32,512	[Residual Block,512]×3	$[1 \times 1, 1024] \Rightarrow$ $\Leftarrow [1 \times 1, 512]$	[Axial Cross Attention, Head=32,1024]×3	1/32,1024
<b>Upsampling</b>	U1	1/8,384	[3×3,384]	Add $\Rightarrow$ $\Leftarrow$ Subtract	[MLP,384]	1/8,384
	U2	1/8,128	[3×3,128]	-	[3×3,128]	1/8,128

neighborhood features of images layer by layer using convolution [28]. A transformer can pay global attention to the image patches through the self-attention mechanism. Local representation and global representation are complementary. Based on the idea, we propose the ACABFNet, the details of which are shown in Table I. The ACABFNet is made up of the CNN branch and transformer branch running in parallel. The CNN is used for local refinement, while the transformer is used for global generalization. Feature fusion is carried out by  $1 \times 1$  convolution with bidirectional intersection. In the upsampling stage, the local features and global features are fused by a  $3 \times 3$  convolution and a multilayer perceptron (MLP), respectively, on two branches. The two features are then added and subtracted, allowing the network to explicitly model the similarity and

difference between local and global features, as well as filter out redundant features.

#### A. CNN Branch

A lot of existing work has demonstrated that ResNet [29] is a deep model with excellent performance thanks to the residual connection structure, so we adopt ResNet as a CNN branch. Since the transformer utilizes nonoverlapping image patches for global attention and MLP for global information fusion, it will inevitably lead to the loss of local details. Thanks to the local correlation and translational invariance, the CNN can effectively use the priori information to model local fine-grained information [30], which makes up for the deficiency of the

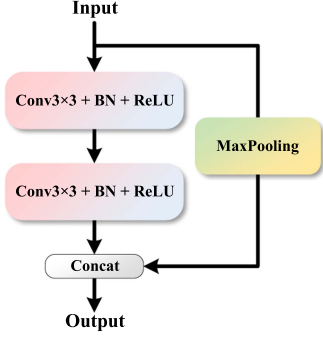


Fig. 2. Structure diagram of Stem in the transformer branch.

transformer on the aforementioned information. To be specific, the CNN branch consists of five stages. First, The CNN stem layer is used to rapidly downsample the input image to obtain a feature map with 1/4 the size of the input. Then, four successive residual layers are followed, with 3, 4, 6, and 3 residual blocks, respectively. As shown in Table I, the feature map is downsampled layer by layer from D2 to D5 stage, finally yielding a feature map with 1/32 the size of the input. Through the CNN Branch, the network can obtain the local feature representation of the input image.

### B. Transformer Branch

Given the excellent performance of the feature pyramid, we adopt the same structure when designing the transformer branch. In addition, the existing work shows that the performance can be effectively improved by using the CNN for downsampling in the initial stage of the transformer network. Therefore, the input first passes through the CNN-based stem layer, which conducts downsampling on the input image using convolution and maxpooling to obtain a feature representation with 1/2 the size of the input, capturing the shallow information of the image quickly and effectively. The structure of stem in the transformer branch is shown in Fig. 2. Then, four consecutive axial cross attention layers are used to capture the long-term dependencies of images. Each layer contains 3, 4, 6, and 3 axial cross attention blocks with 4, 8, 16, and 32 heads, respectively. The feature map is downsampled layer by layer to obtain the feature hierarchical representation structure similar to the CNN branch. The design is based on the following two considerations.

- 1) The hierarchical representation structure is more flexible in model design.
- 2) It is convenient for bidirectional fusion with the CNN Branch.

Self-attention is strong at modeling the long-term dependencies of images [31], but it has to pay attention to all patches of images, which causes feature extraction to be inefficient [32]. There are many background interference factors in remote sensing images. If the global attention is paid to all patches directly and roughly, the phenomenon of feature redundancy will occur. The feature will not only affect the classification accuracy of the model, but also significantly occupy GPU during training, which is extremely unfriendly to hardware devices. Therefore,

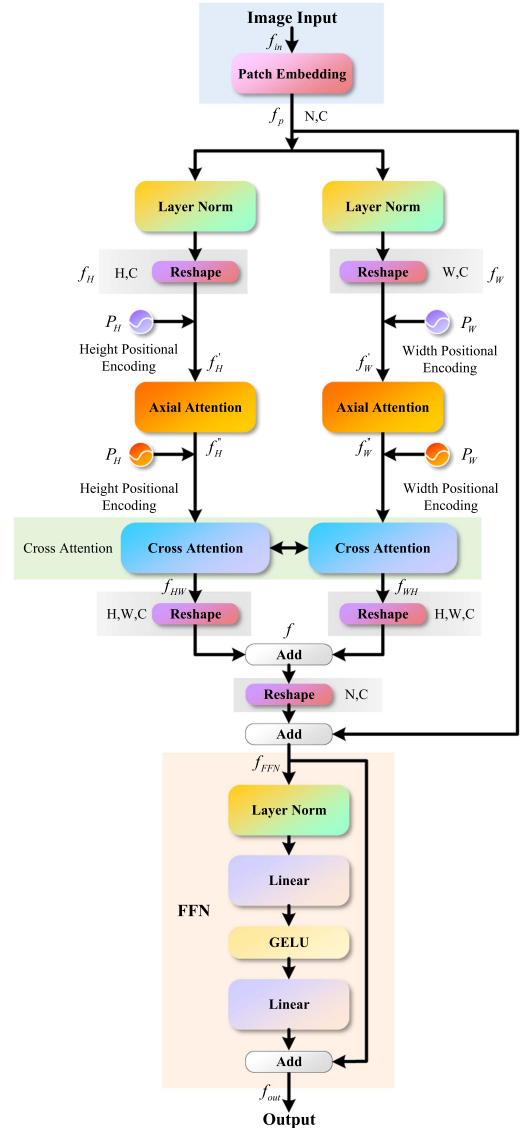


Fig. 3. Structure diagram of axial cross attention.

we propose the axial cross attention, which consists of axial attention, cross attention, and feed forward network. The structure of the axial cross attention is shown in Fig. 3.

Patch embedding (PE) is first applied to transform the input  $f_{in} \in R^{C \times H \times W}$  into a sequence  $f_p \in R^{(H \times W) \times C}$ .  $f_p$  is then reshaped into  $f_{H/W} \in R^{H/W \times C}$  after Layer Norm (LN) in preparation for the following global attention to the image along the height and width, respectively. Since the transformer captures the global semantic information while ignoring the positional information, positional embedding is introduced before axial attention and cross attention, respectively.  $P_H$  represents the positional embedding along the height dimension, while  $P_W$  represents the positional embedding along the width dimension. The aforementioned process is mathematically expressed as follows:

$$f'_{H/W} = \text{Reshape}(\text{LN}(\text{PE}(f_{in}))) + P_{H/W}. \quad (1)$$



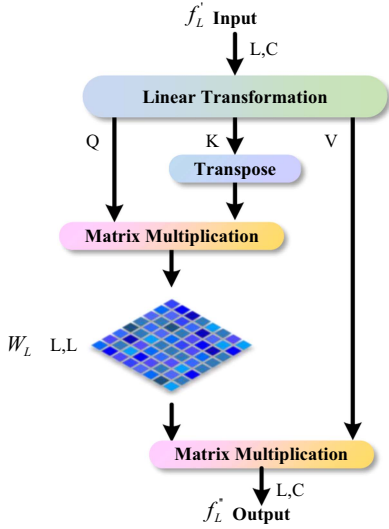


Fig. 4. Structure diagram of axial attention.

Axial attention only carries out global semantic modeling along the height or width dimensions of the image. Compared with the original self-attention, the structure is more efficient in global attention and more GPU friendly. As shown in Fig. 4, the input feature map  $f_{H/W}^i \in R^{H/W \times C}$  first passes through a linear transformation  $\phi$  to obtain the Query ( $Q$ ), Key ( $K$ ), and Value ( $V$ ). The feature relationship matrix  $W_{H/W} \in R^{H \times H/W \times W}$  of the height or width dimension is obtained by the matrix multiplication of  $Q$  and  $K$  and activated by the Softmax function. Finally, the matrix multiplication is applied to  $W_{H/W}$  and  $V$  to obtain the output  $f_{H/W}^o \in R^{H/W \times C}$ . Because it is modeled only along a single dimension, axial attention is a computationally efficient structure. The mathematical expression of the aforementioned process is as follows:

$$Q, K, V = \phi(f_{H/W}^i) \quad (2)$$

$$W_{H/W} = \text{Softmax}(Q \times K) \quad (3)$$

$$f_{H/W}^o = W_{H/W} \times V. \quad (4)$$

Given that a single axial attention loses information from the other dimension, we introduce cross attention to fuse the global semantic information from two different dimensions, the structure of which is shown in Fig. 5. Different from axial attention, the query, key and value of cross attention are derived from two different inputs in order to obtain the feature relationship between the height and width of images. Specifically, two input features  $f_H^i \in R^{H \times C}$  and  $f_W^i \in R^{W \times C}$  first pass through linear transformations  $\phi_H$  and  $\phi_W$ , respectively, to obtain two groups of query ( $Q_H, Q_W$ ), key ( $K_H, K_W$ ), and value ( $V_H, V_W$ ). Then, two feature relationship matrices  $W_{HW} \in R^{H \times W}$  and  $W_{WH} \in R^{W \times H}$  are obtained by matrix multiplication of  $Q_H$  and  $K_W$ ,  $Q_W$  and  $K_H$ , respectively, and activated by the Softmax function. Finally, the matrix multiplication of  $W_{HW}$  and  $V_W$ , and  $W_{WH}$  and  $V_H$ , respectively, gives two outputs  $f_{HW} \in R^{H \times C}$  and  $f_{WH} \in R^{W \times C}$ .  $f_{HW}$  and  $f_{WH}$  fully integrate the global information of the height and width dimensions

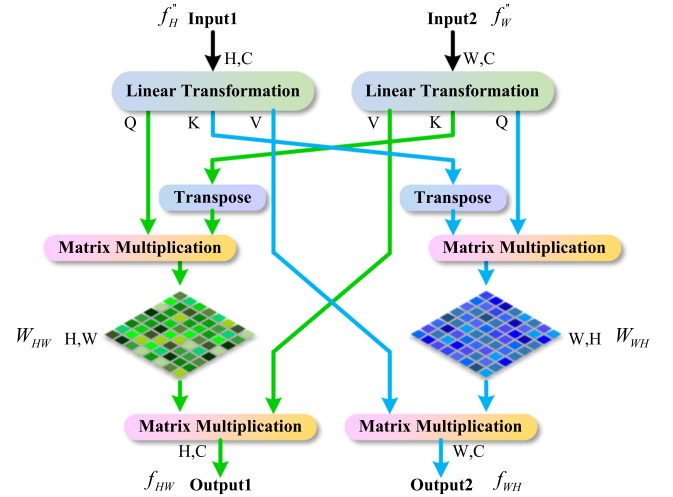


Fig. 5. Structure diagram of cross attention.

extracted from axial attention, effectively compensating for the information deficiency of axial attention in the other dimension. The mathematical expression of the aforementioned process is as follows:

$$Q_H, K_H, V_H = \phi_H(f_H^i) \quad (5)$$

$$Q_W, K_W, V_W = \phi_W(f_W^i) \quad (6)$$

$$f_{HW} = \text{Softmax}(Q_H \times K_W) \times V_W \quad (7)$$

$$f_{WH} = \text{Softmax}(Q_W \times K_H) \times V_H. \quad (8)$$

Like the original transformer, the feed forward network (FFN) is used to fuse the global features of axial cross attention. As shown in Fig. 3, the input feature  $f_{FFN} \in R^{(H \times W) \times C}$  first goes through the LN, then two linear transformations  $\phi_1$  and  $\phi_2$ . Between  $\phi_1$  and  $\phi_2$ , there exists a GELU function, which is used to activate the features. Finally, the output  $f_{out} \in R^{(H \times W) \times C}$  of axial cross attention is obtained by adding it to the shortcut connection. Through the FFN, the global feature information of the height and width of the images can be effectively fused and enhanced. The mathematical expression of the aforementioned process is as follows:

$$f_{out} = \phi_2(\text{GELU}(\phi_1(\text{LN}(f_{FFN})))) + f_{FFN}. \quad (9)$$

### C. Feature Recovery

In order to efficiently integrate global features and local features in the upsampling stage, we first fuse the features of the last three stages of the transformer branch and CNN branch, respectively, then add and subtract the fusion features of the two branches, the purpose of which is to enable the network to distinguish the similarity and difference between global features and local features explicitly, and to extract discriminant features more effectively. Specifically, for the transformer branch, linear transformation  $\phi^i$  is first performed on each layer's features  $f_T^i \in R^{N^i \times C^i}$  ( $N^i = H^i \times W^i$ ), transforming them into  $f_{TC}^i \in R^{N^i \times C}$ . Then, the LN and GELU are applied to  $f_{TC}^i$ . For the CNN branch,  $3 \times 3$  convolution  $\psi^i$  is used to operate on

TABLE II  
ABLATION EXPERIMENTS OF ACABFNET

Axial Attention	Cross Attention	+/- operations	MIOU(%)	F1(%)
×	×	×	77.51	78.77
✓	×	✓	80.34	81.72
×	✓	✓	80.28	81.71
✓	✓	×	80.49	82.05
✓	✓	✓	<b>80.77</b>	<b>82.39</b>

Bold face numbers indicate optimal accuracy.

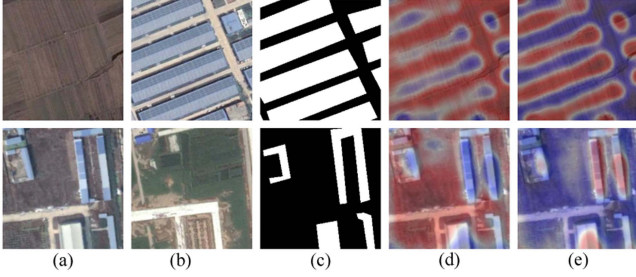


Fig. 6. Visualization of the similarity and difference between global and local features. (a) and (b) Bitemporal remote sensing images. (c) Labels. (d) Heat maps of the similarity features. (e) Heat maps of the difference features.

each layer's features  $f_N^i \in R^{D^i \times H^i \times W^i}$  directly, transforming them into  $f_{NC}^i \in R^{C \times H^i \times W^i}$ . Then  $f_{NC}^i$  goes through batch norm (BN) and ReLU. Through the aforementioned operations, the channels of each layer's features of the two branches are transformed to the same size. Next, in order to concatenate all the features, each layer's features  $f_{TC}^i$  are reshaped into  $f_{TI}^i \in R^{C \times H^i \times W^i}$  and upsampled so that all the features have the same dimension size  $C \times H \times W$ . Both  $H$  and  $W$  are 1/8 the length of the input image's side length. It is worth noting that both bilinear interpolation and transpose convolution are employed for upsampling. The former does not require training, whereas the latter follows the network's training. They can complement each other well. Each layer's features  $f_{NC}^i$  in the CNN branch are also transformed into  $f_{NI}^i \in R^{C \times H \times W}$  by a similar upsampling operation. Finally, concatenation operation is performed on each layer's features of the two branches to obtain the global fusion feature and local fusion feature of the whole network. The mathematical expression of the aforementioned process is as follows:

$$f_{TI}^i = \text{Upsample}(\text{Reshape}(\text{GELU}(\text{LN}(\phi^i(f_T^i)))))) \quad (10)$$

$$f_{NI}^i = \text{Upsample}(\text{ReLU}(\text{BN}(\psi^i(f_N^i))))). \quad (11)$$

Existing semantic segmentation-based change detection algorithms do not take into account the similarity and difference between the global and local feature information when combining the transformer and CNN for feature recognition, resulting in significant feature redundancy. Therefore, at the end of the network, we explicitly model the similarity and difference of global features and local features by addition and subtraction operations, and optimize the features by  $3 \times 3$  convolution block. Fig. 6 shows the heat maps of two discriminant features.

For the change detection task, similarity features mainly focus on the unchanged area of bitemporal remote sensing images, while difference features focus on the changed area, as shown in the red section in the figure.

### III. DATASETS

To comprehensively verify the effectiveness of the ACABFNet proposed in this article, we conduct training and testing of the model on three different remote sensing image change detection datasets, namely BTCDD [1], CDD [33], and LEVIR-CD [34].

#### A. BTCDD

The BTCDD dataset is a remote sensing image change detection dataset that we suggested in our first work. It contains 5281 pairs of high-resolution bitemporal remote sensing images with  $256 \times 256$  pixels each, among which 4224 pairs of images are used as training set and 1057 pairs are used as test set. All of the images were taken in different regions of China from 2010 to 2020. The types of changed areas include factories, farmland, roads, buildings, and mining areas.

#### B. CDD

The CDD dataset consists of seven pairs of bitemporal remote sensing images with  $4725 \times 2700$  pixels each and four pairs with  $1900 \times 1000$  pixels each. Eleven pairs of images are synchronously cropped into 16 000 pairs of image patches with  $256 \times 256$  pixels each, of which 10 000 pairs constitute training set, 3000 pairs constitute validation set, and the rest 3000 pairs constitute test set. Seasonal variations are taken into account to make the trained networks more convincing.

#### C. LEVIR-CD

The LEVIR-CD dataset is composed of 637 pairs of high-resolution Google Earth images with  $1024 \times 1024$  pixels each. All the images were taken in 20 different areas of Texas between 2002 and 2018. The dataset focuses on significant changes in buildings, including villas, apartments, garages, warehouses, and so on. In addition, seasonal variations and illumination variations are taken into consideration, which help to develop high-performance models.

## IV. EXPERIMENTS AND RESULTS

### A. Evaluation Indicators

For assessing the performance of the ACABFNet in the change detection task, we adopt four evaluation indicators, namely Precision, Recall, MIOU, and F1 score. The mathematical expression of the evaluation indicators is as follows:

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

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

$$\text{MIOU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (14)$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (15)$$

where TP, FP, and FN denote True Positives, False Positives, and False Negatives, respectively.

### B. Experimental Details

In this article, all the experiments are implemented on a GeForce RTX 2080Ti GPU based on Pytorch. BCEWithLogitsLoss is utilized as the loss function for training models. Adam is used as the optimizer. The batch size is set to 4. The initial learning rate ( $lr$ ) is set to 0.0001. On BTCDD, the maximum number of epoches ( $\text{max\_epoch}$ ) is set to 250. On CDD and LEVIR-CD,  $\text{max\_epoch}$  is set to 200. In view of the effectiveness of dynamic adjustment of learning rate, we adopt the Poly learning rate reduction strategy, and the learning rate of each epoch is  $lr \times (1 - \frac{\text{epoch}}{\text{max\_epoch}})$ .

### C. Ablation Experiments

For verifying the effectiveness of the presented ACABFNet, we conduct ablation experiments on the BTCDD dataset. The experimental results are shown in Table II.

- 1) *Transformer branch*: We remove the transformer branch from the overall network structure, so the network degenerates into ResNet34 [29]. It is worth noting that we retain the original prediction head to complete change detection. It can be concluded from Table II that the transformer branch has brought significant improvements to our model, with an increase of 3.26% and 3.62%, respectively, on two evaluation indicators.
- 2) *Axial cross attention*: We conduct ablation experiments on axial attention and cross attention, respectively. From Table II, we can find that after removing the two attention modules, respectively, the model has performance degradation on two important indicators to varying degrees. The combination of the two enables the model to fuse the global feature information on the two spatial dimensions of the images effectively, improving the feature expression performance. Axial attention improves MIOU and F1 score by 0.49% and 0.68%, respectively. Cross attention improves MIOU and F1 score by 0.43% and 0.67%, respectively.

- 3) *+/- operations*: The addition and subtraction operations explicitly distinguish the similarity and difference between the global features of the transformer branch and the local features of the CNN branch, effectively filtering out the redundant feature information. As can be seen from Table II, the operations make the model improve MIOU and F1 score by 0.28% and 0.34%, respectively.

### D. Comparative Experiments

The performance of the ACABFNet is evaluated on three datasets, namely BTCDD, CDD, and LEVIR-CD. A quantity of comparative experiments completely demonstrate the outstanding performance of the ACABFNet. Considering the fairness of the experiment, the super parameter settings of all experiments on a dataset are the same. In addition, for the single-input semantic segmentation-based model (SETR/HRNet/PSPNet, etc.), we directly concatenate the bitemporal images along the channel dimension, and then, send it to the network. For the dual-input change detection model (FC-EF/FC-Siam-conc/FC-Siam-diff), the bitemporal images are sent to the network at the same time.

1) *Comparative Experiments on BTCDD*: We first conduct comparative experiments on the BTCDD dataset. To make the experimental results more convincing, we compare the change detection algorithms based on the CNN and vision transformer at the same time. The experimental results are shown in Table III, where \* represents the transformer-based method. From the table, it can be seen that some CNN-based algorithms are significantly better than transformer-based algorithms, which may be due to the small amount of data in BTCDD. In comparison to existing change detection methods based on deep learning, our suggested ACABFNet achieves the best accuracy on most indicators. Its MIOU and F1 score are 0.76% and 1.07% higher than PSANet's, respectively. In addition, compared with BiSeNet, HRNet, PVT, and SegFormer, our algorithm achieves significant improvement on four indicators on the premise of adding a small amount of calculation. It should be emphasized here that, suppose  $N(N = H \times W)$  represents the length of the sequence, and  $d$  represents the dimension. The computational complexity of self-attention is  $\mathcal{O}(N^2 \cdot d)$ . Our axial cross attention calculates  $H$  and  $W$ , respectively, and its computational complexity is  $\mathcal{O}(L^2 \cdot d)$  when  $L = H = W$ . Compared with the original self-attention,  $\mathcal{O}(L^2)$  times the calculation is saved. Therefore, axial cross attention can effectively obtain the global representations of the images with a relatively low amount of calculation, which is one of the advantages of our algorithm. However, the number of parameters of our model is a little large, and we will solve this problem in future work.

Fig. 7 shows the prediction results of different algorithms. We display two groups of images from 1057 groups of prediction results. As shown in the figure, the performance of the transformer-based models represented by SETR, PVT, and SegFormer is unsatisfactory. Due to the lack of inductive bias, the transformer usually needs training on large datasets to achieve a better performance. Moreover, existing change detection algorithms based on the CNN are difficult to effectively model the global feature information, so the phenomenon of false detection and

TABLE III  
COMPARATIVE EXPERIMENTS ON BTCDD

Algorithm	Precision(%)	Recall(%)	MIOU(%)	F1(%)	Params(M)	FLOPs(G)
FC-Siam-diff [19]	82.91	41.15	61.64	55.00	1.35	4.26
FC-EF [19]	73.75	56.62	66.13	64.06	1.35	3.12
FC-Siam-conc [19]	81.64	52.84	66.77	64.16	1.55	4.87
SETR* [35]	79.00	67.82	72.93	72.98	308.28	91.35
PVT* [36]	79.39	71.61	74.49	75.30	61.18	12.86
DeepLab V3+ [37]	75.20	78.27	75.48	76.70	54.93	45.77
HRNet [38]	80.93	74.61	76.69	77.64	65.86	23.50
BiSeNet [39]	83.11	75.46	78.03	79.10	50.25	11.14
SegFormer* [40]	82.90	77.61	78.73	80.17	81.54	15.00
PSPNet [17]	83.90	76.82	79.00	80.20	67.94	64.03
OCNet [41]	81.73	79.44	79.18	80.57	55.03	57.04
DANet [18]	83.56	78.57	79.47	80.99	66.55	70.80
PSANet [42]	<b>84.76</b>	78.15	80.01	81.32	69.33	71.54
ACABFNet (Ours)	84.47	<b>80.41</b>	<b>80.77</b>	<b>82.39</b>	101.06	23.93

Bold face numbers indicate optimal accuracy

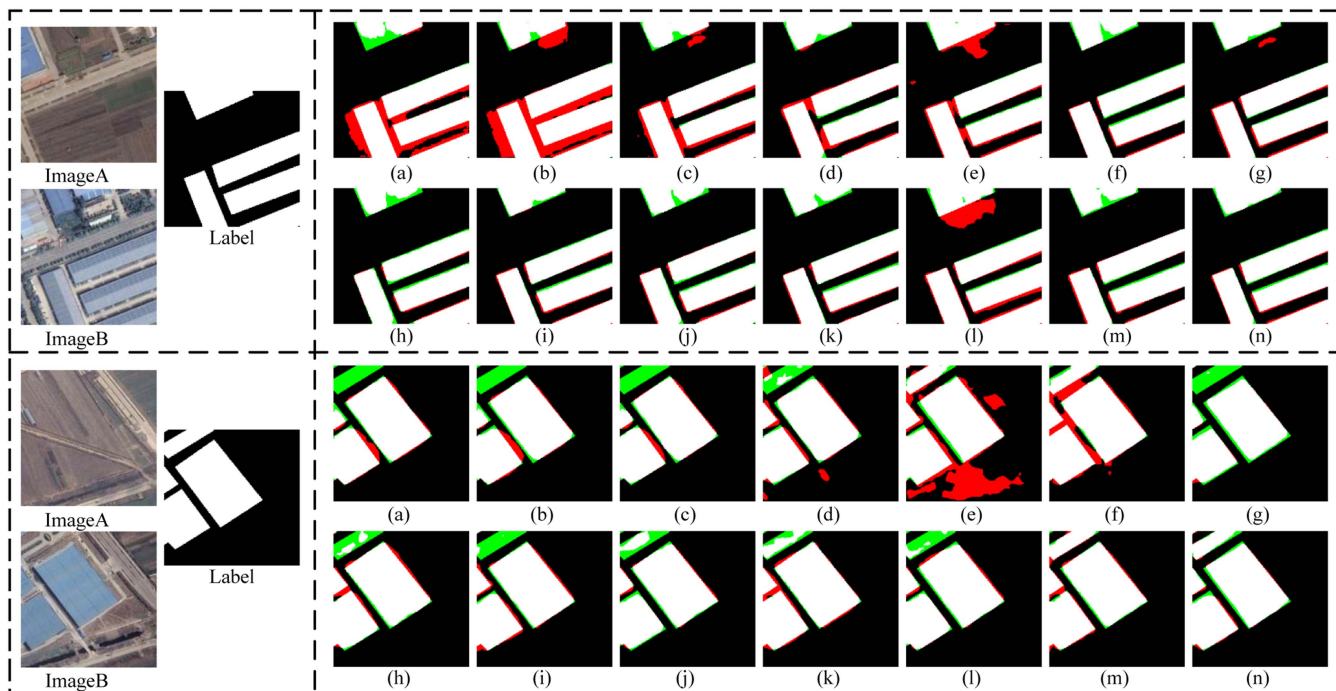


Fig. 7. Schematic diagram of prediction results of different algorithms on BTCDD. (a)–(n) Prediction results of FC-Siam-diff, FC-EF, FC-Siam-conc, SETR, PVT, DeepLab V3+, HRNet, BiSeNet, SegFormer, PSPNet, OCNet, DANet, PSANet, and ACABFNet. The red indicates false alarms. The green indicates omission alarms.

missing detection is very serious when predicting the changed areas. As shown in the first group of comparison diagrams in Fig. 7, some CNN-based models incorrectly predict the land around the blue factories as changed areas (indicated in red), resulting in serious false alarms. In addition, look at the second group of images. The CNN-based detection models have an obvious phenomenon of missing detection (indicated in green). The ACABFNet proposed by us combines the advantages of the CNN in local feature extraction and the transformer in global feature extraction to effectively improve the accuracy of change

detection. Compared with existing change detection algorithms, the ACABFNet outperforms them in detecting the boundaries of changed areas. Fig. 8 shows the heat maps of the intermediate layers of the ACABFNet. Red indicates higher attention, and blue indicates lower attention. From the figure, we can see that the CNN branch pays more attention to fine-grained boundary information, and the transformer branch focuses more on the overall representation of the images. The two are complementary to each other, which enables the model to capture more abundant information about the images.



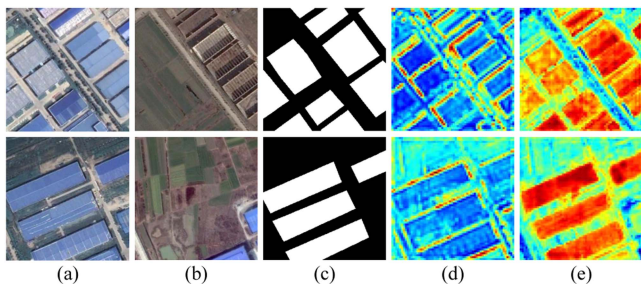


Fig. 8. Visualization of the heat maps of the intermediate layers of the ACABFNet. (a) and (b) Bitemporal remote sensing images. (c) Labels. (d) Heat maps of the features of the CNN Branch. (e) Heat maps of the features of the transformer branch.

TABLE IV  
COMPARATIVE EXPERIMENTS ON CDD

Algorithm	Precision(%)	Recall(%)	MIOU(%)	F1(%)
FC-EF [19]	85.45	55.11	71.76	67.00
FC-Siam-diff [19]	88.74	62.47	76.05	73.32
FC-Siam-conc [19]	92.03	61.02	76.17	73.39
HRNet [38]	93.40	90.64	91.55	92.00
SETR* [35]	95.41	91.97	93.21	93.66
BiSeNet [39]	95.76	94.33	94.62	95.04
PVT* [36]	96.05	94.28	94.74	95.16
PSPNet [17]	96.58	95.48	95.65	96.03
OCNet [41]	96.36	95.71	95.66	96.03
SegFormer* [40]	96.33	96.02	95.81	96.17
PSANet [42]	96.65	95.74	95.83	96.19
DeepLab V3+ [37]	96.25	96.57	96.06	96.41
DANet [18]	96.57	96.38	96.13	96.48
ACABFNet (Ours)	<b>96.98</b>	<b>97.32</b>	<b>96.85</b>	<b>97.15</b>

Bold face numbers indicate optimal accuracy.

Fig. 9 shows the accuracy curves of our ACABFNet and several models on BTCDD. It can be seen from the figure that our ACABFNet is superior to other algorithms on the test set. This means that the ACABFNet has a better generalization performance. In addition, under the same number of iterations during training, our model can achieve the optimal detection accuracy on the test set, which may be attributed to the effective complementarity between the CNN branch and the transformer branch. The simultaneous development of local features and global features makes it easier for the model to detect more complex changed areas.

2) *Comparative Experiments on CDD*: A single dataset is insufficient to comprehensively examine the performance of the model, so we also experiment on CDD dataset. All of the models are retrained and retested on CDD dataset. The experimental results are shown in Table IV. Similarly, the transformer-based algorithms are denoted by \*. Considering that the training set of CDD dataset contains only 10 000 pairs of images, which is a fraction of the size of large datasets such as ImageNet and COCO, it is understandable that transformer-based models perform poorly. Thanks to the high efficiency of axial cross attention, the ACABFNet can quickly converge to the optimal accuracy. As shown in table, the ACABFNet improves MIOU

TABLE V  
COMPARATIVE EXPERIMENTS ON LEVIR-CD

Algorithm	Precision(%)	Recall(%)	MIOU(%)	F1(%)
SETR* [35]	86.99	79.47	84.66	83.06
FC-EF [19]	85.64	81.08	84.82	83.30
FC-Siam-diff [19]	88.91	83.65	87.16	86.20
FC-Siam-conc [19]	87.82	85.68	87.59	86.74
PVT* [36]	90.55	85.47	88.61	87.93
BiSeNet [39]	89.73	86.30	88.64	87.98
PSPNet [17]	90.29	87.01	89.19	88.62
DeepLab V3+ [37]	88.92	88.54	89.27	88.73
HRNet [38]	90.61	87.26	89.43	88.90
SegFormer* [40]	91.01	87.43	89.67	89.18
OCNet [41]	90.47	88.35	89.85	89.40
PSANet [42]	91.14	87.90	89.94	89.49
DANet [18]	91.22	87.95	90.00	89.56
ACABFNet (Ours)	<b>91.40</b>	<b>89.96</b>	<b>90.98</b>	<b>90.68</b>

Boldface numbers indicate optimal accuracy.

and F1 score by 0.72% and 0.67%, respectively, compared with the suboptimal model DANet.

Fig. 10 shows the prediction results of various algorithms on CDD dataset. We select two groups of prediction results from 3000 groups for display. As shown in the figure, most of the existing deep learning-based models can only predict a portion of the track when detecting narrow roads, leading to serious omission alarms (indicated in green). The coherence of the prediction results is poor. Our ACABFNet can detect the entire changed track coherently, and the detection edges are smoother. This is due to our axial cross attention focusing on the image from a global perspective, and the explicit discrimination of similarity and difference between global features and local features.

3) *Comparative Experiments on LEVIR-CD*: We carry out the third group of comparative experiments on LEVIR-CD dataset. Considering the limitation of GPU, the original image pairs with  $1024 \times 1024$  pixels each are synchronously cropped into image patches with  $256 \times 256$  pixels each. The training set includes 7120 pairs of images and the test set includes 2048 pairs of images. Similarly, all of the models are retrained and retested on LEVIR-CD dataset. The experimental results are shown in Table V. \* indicates the vision transformer. Our proposed ACABFNet is superior to the existing deep learning-based models on four indicators. Especially on MIOU and F1 score, the ACABFNet is 1.31% and 1.5% higher than the transformer-based model SegFormer, and 0.98% and 1.12% higher than the suboptimal model DANet, respectively. This adequately proves the effectiveness of our algorithm.

Fig. 11 shows the prediction results of different algorithms on LEVIR-CD dataset. The two groups of prediction images are from 2048 groups of predicted images from the test set. As can be seen from the figure, when detecting the changes of several adjacent buildings, most of the prediction boundaries of existing deep-learning-based models are serrated, and even the phenomenon of adhesion occurs, resulting in false detection and missing detection (indicated in red and green, respectively).

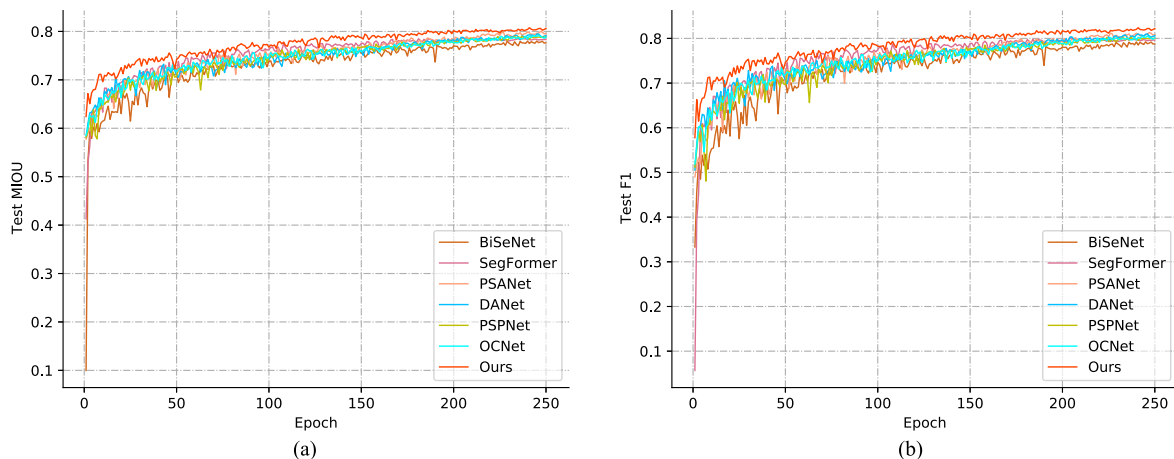


Fig. 9. Accuracy curves on BTCDD. MIOU and F1 are reported. (a) Test MIOU curve. (b) Test F1 curve.

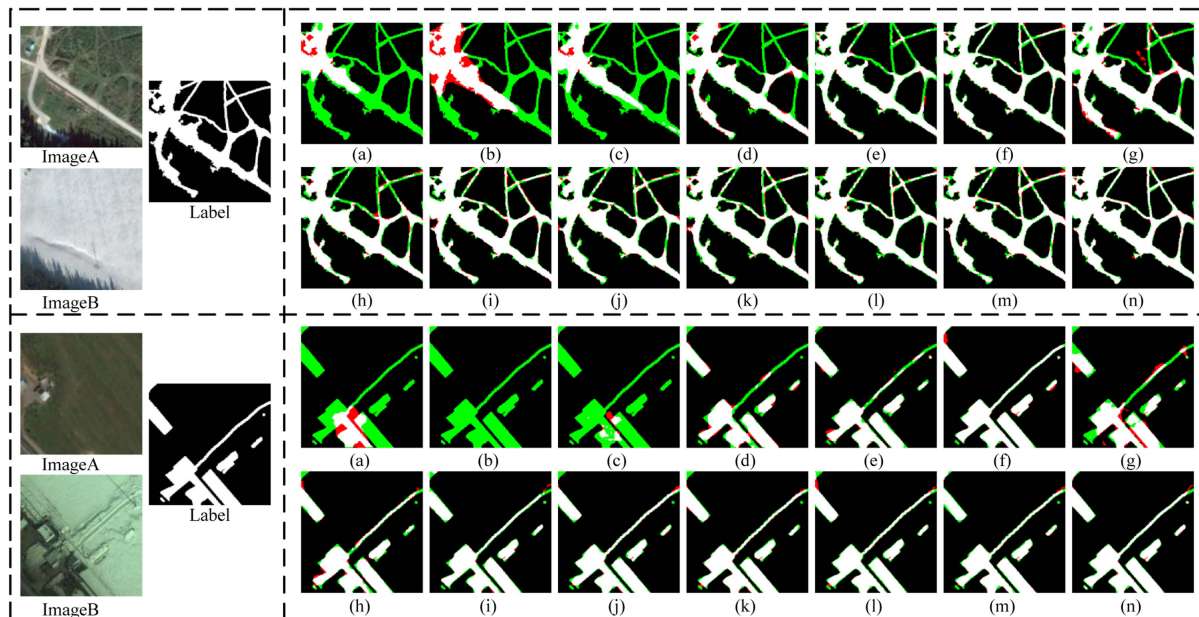


Fig. 10. Schematic diagram of prediction results of different algorithms on CDD. (a)–(n) Prediction results of FC-Siam-diff, FC-EF, FC-Siam-conc, SETR, PVT, DeepLab V3+, HRNet, BiSeNet, SegFormer, PSPNet, OCNet, DANet, PSANet, and ACABFNet. The red indicates false alarms. The green indicates omission alarms.

Our ACABFNet can distinguish each changed building clearly, and the prediction boundaries are smoother, which effectively reduces the occurrence of false alarms and omission alarms.

## V. DISCUSSION

The aforementioned extensive experiments effectively prove the advantages of our algorithm from multiple perspectives. Conventional change detection methods cannot cope with the task of change detection in different imaging environments due to their simple feature extraction ways. As shown in the two visualization maps of PCA-Means in Fig. 7, there are dense areas of false detection and missing detection. Most of the existing learning-based change detection methods rely heavily on the

CNN framework, which is limited by the size of the convolution kernel, and cannot effectively distinguish the differences between two images and model the relationships among changed areas from a global semantic perspective. As shown in Figs. 10 and 11, there are strong semantic associations among the narrow roads and the dense buildings. Ignoring this leads to the serious omission alarms. Some existing transformer-based methods, such as SETR, PVT, and SegFormer, model the long-distance dependence based on self-attention to solve the aforementioned problem. However, it should be noted that only using self-attention to process the images will often lead to the loss of local details, especially the small objects. Different from the existing change detection methods, our proposed ACABFNet utilizes both the local attention capability of the CNN and the global

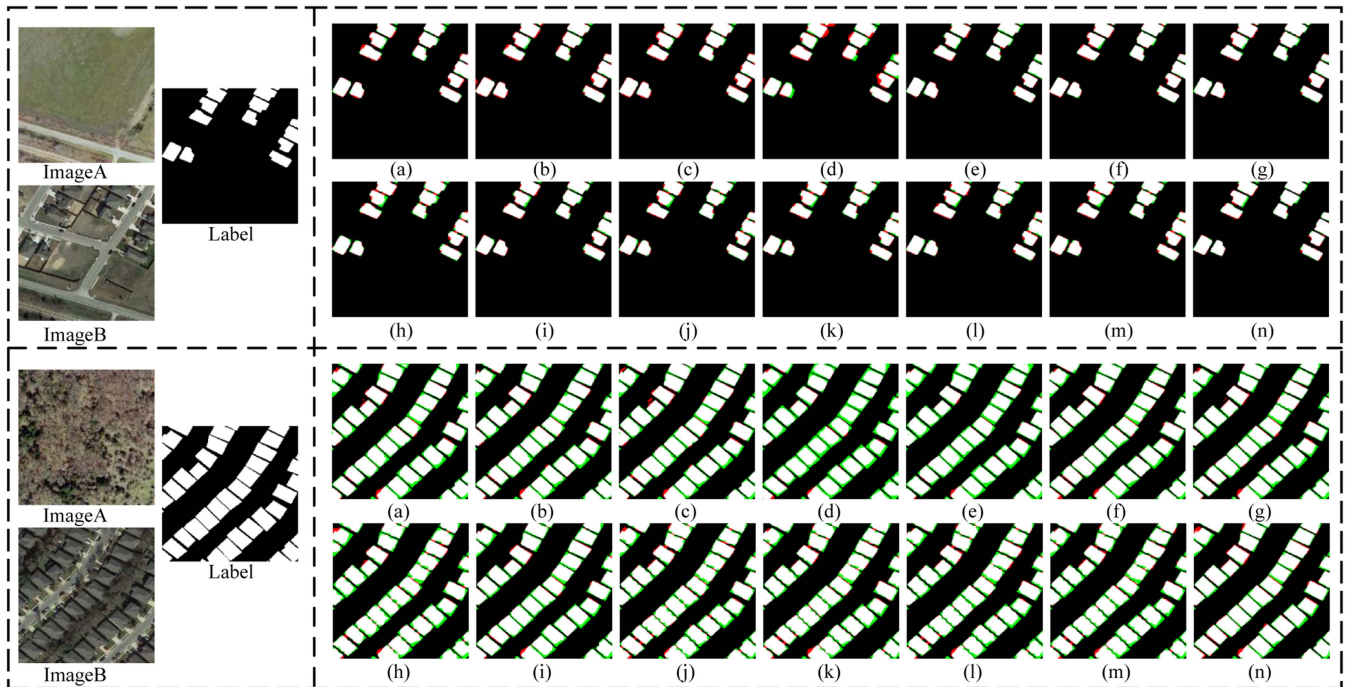


Fig. 11. Schematic diagram of predicted results of different algorithms on LEVIR-CD dataset. (a)–(n) Prediction results of FC-Siam-diff, FC-EF, FC-Siam-conc, SETR, PVT, DeepLab V3+, HRNet, BiSeNet, SegFormer, PSPNet, OCNet, DANet, PSANet, and ACABFNet. The red indicates false alarms. The green indicates omission alarms.

semantic modeling capability of axial cross attention, achieving reliable improvement with regard to detection accuracy with relatively low FLOPs. The curves in Fig. 9 indicate that our model converges faster under the same iteration number and outperforms other algorithms on the test set.

## VI. CONCLUSION

In this article, we present a ACABFNet. The network is composed of CNN branch and transformer branch in parallel, and two-way interaction is carried out through  $1 \times 1$  convolution to fuse local and global feature information. Considering that the efficiency of global attention of the self-attention mechanism is low, we propose an axial cross attention, which first pays global attention to the images along the height and width dimensions, respectively, and then, fuses the global feature information from the two dimensions through cross attention. Compared with the original self-attention, axial cross attention has higher global feature extraction efficiency and is more GPU friendly. Furthermore, at the end of the network, we perform explicit discrimination of similarity and difference between global features and local features by addition and subtraction operations, effectively filtering out the redundant features. Experimental results on BTCDD, CDD, and LEVIR-CD datasets show that the ACABFNet outperforms existing conventional change detection algorithms and semantic segmentation-based algorithms.

### Code availability section

Name of the library: ACABFNet

Hardware requirements: GeForce RTX 2080Ti (12 G)

Software required: python3.8

Packages: Pytorch

The source codes are available for downloading at the link: <https://github.com/SONGLEI-arch/ACABFNet>

## REFERENCES

- [1] L. Song, M. Xia, J. Jin, M. Qian, and Y. Zhang, "SUACDNet: Attentional change detection network based on siamese U-shaped structure," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, 2021, Art. no. 102597.
- [2] M. Malmir, M. M. K. Zarkesh, S. M. Monavari, S. A. Jozi, and E. Sharifi, "Urban development change detection based on multi-temporal satellite images as a fast tracking approach—A case study of Ahwaz county, Southwestern Iran," *Environ. Monit. Assessment*, vol. 187, no. 3, pp. 1–10, 2015.
- [3] P. Washaya, T. Balz, and B. Mohamadi, "Coherence change-detection with sentinel-1 for natural and anthropogenic disaster monitoring in urban areas," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1026.
- [4] K. Rokni, A. Ahmad, A. Selamat, and S. Hazini, "Water feature extraction and change detection using multitemporal landsat imagery," *Remote Sens.*, vol. 6, no. 5, pp. 4173–4189, 2014.
- [5] J. Zhou, B. Yu, and J. Qin, "Multi-level spatial analysis for change detection of urban vegetation at individual tree scale," *Remote Sens.*, vol. 6, no. 9, pp. 9086–9103, 2014.
- [6] C. Lu, M. Xia, and H. Lin, "Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation," *Neural Comput. Appl.*, vol. 34, pp. 6149–6162, 2022.
- [7] Y. Qu, M. Xia, and Y. Zhang, "Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow," *Comput. Geosci.*, vol. 157, 2021, Art. no. 104940.
- [8] R. Weismiller, S. Kristof, D. Scholz, P. Anuta, and S. Momin, "Change detection in coastal zone environments," *Photogrammetric Eng. Remote Sens.*, vol. 43, no. 12, pp. 1533–1539, 1977.
- [9] E. J. Rignot and J. J. Van Zyl, "Change detection techniques for ERS-1 SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 31, no. 4, pp. 896–906, Jul. 1993.
- [10] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 218–236, Jan. 2007.



- [11] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.
- [12] M. Xia, Y. Qu, and H. Lin, "Padanet: Parallel asymmetric double attention network for clouds and its shadow detection," *J. Appl. Remote Sens.*, vol. 15, no. 4, 2021, Art. no. 046512.
- [13] J. Gao, L. Weng, M. Xia, and H. Lin, "MLNet: Multichannel feature fusion lozenge network for land segmentation," *J. Appl. Remote Sens.*, vol. 16, no. 1, pp. 1–19, 2022.
- [14] Z. Wang, M. Xia, M. Lu, L. Pan, and J. Liu, "Parameter identification in power transmission systems based on graph convolution network," *IEEE Trans. Power Del.*, vol. 37, no. 4, pp. 3155–3163, Aug. 2022.
- [15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
- [16] 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," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
- [17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.
- [18] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
- [19] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.
- [20] A. Dosovitskiy et al., "An image is worth  $16 \times 16$  words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [21] A. Vaswani et al., "Attention is all you need," in *31st Int. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
- [22] C. Lu, M. Xia, M. Qian, and B. Chen, "Dual-branch network for cloud and cloud shadow segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5410012.
- [23] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 367–376.
- [24] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12175–12185.
- [25] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16519–16529.
- [26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [27] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [28] K. Pang, L. Weng, Y. Zhang, J. Liu, H. Lin, and M. Xia, "SGBNet: An ultra light-weight network for real-time semantic segmentation of land cover," *Int. J. Remote Sens.*, vol. 43, no. 15–16, pp. 5917–5939, 2022.
- [29] 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.
- [30] M. Xia, X. Zhang, W. Liu, L. Weng, and Y. Xu, "Multi-stage feature constraints learning for age estimation," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, no. 1, pp. 2417–2428, Jan. 2020.
- [31] S. Miao, M. Xia, M. Qian, Y. Zhang, J. Liu, and H. Lin, "Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery," *Int. J. Remote Sens.*, vol. 43, no. 15–16, pp. 5940–5960, 2022.
- [32] B. Chen, M. Xia, M. Qian, and J. Huang, "MANet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images," *Int. J. Remote Sens.*, vol. 43, no. 15–16, pp. 5874–5894, 2022.
- [33] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multi-source 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.
- [34] 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.
- [35] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
- [36] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.
- [37] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
- [38] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
- [39] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.
- [40] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077–12090, 2021.
- [41] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "OCNet: Object context network for scene parsing," 2018, *arXiv:1809.00916*.
- [42] H. Zhao et al., "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 267–283.

**Lei Song** received the master's degree in machine learning and remote sensing image analysis from the Nanjing University of Information Science and Technology, Nanjing, China, in 2022.

His research interests are machine learning and its application.

**Min Xia** (Member, IEEE) received the Ph.D. degree in cybernetics control engineering from Donghua University, Shanghai, China, in 2009.

He was an Assistant Researcher with Hongkong Polytechnic University, Hong Kong, from August 2008 to April 2010. He is currently a Professor with the Nanjing University of Information Science and Technology, Nanjing, China. His main research interests include machine learning theory and remote sensing image analysis.

**Liguo Weng** received the doctorate degree in electrical engineering from North Carolina Agricultural and Technical State University, Greensboro, NC, USA, in 2010.

He is currently an Associate Professor with the Nanjing University of Information Science and Technology, Nanjing, China. His research interests include remote sensing image analysis, artificial intelligence, and deep learning.

**Haifeng Lin** received the doctorate degree in computer science from Nanjing Forestry University, Nanjing, China, in 2012.

He is currently an Associate Professor with the Nanjing Forestry University. His main research interests include machine learning theory and remote sensing image analysis.

**Ming Qian** received the master's degree in electronic information from the Nanjing University of Information Science and Technology, Nanjing, China. He is currently working toward the doctorate degree in computer science with Wuhan University.

His main research interests are machine learning and remote sensing big data analysis.

**Binyu Chen** received the postgraduate degree in machine learning and remote sensing image analysis from the Nanjing University of Information Science and Technology, Nanjing, China, in 2022.

Her research interests are machine learning and its application.