# CBSASNet: A Siamese Network Based on Channel Bias Split Attention for Remote Sensing Change Detection

Naiwei He, Liejun Wang, Panpan Zheng, Cui Zhang, and Lele Li

*Abstract*— **Remote sensing image change detection (CD) is an important technology for monitoring ground object change. Although transformer-based CD methods have been proposed and achieved good results, however, there does exist one open problem: transformer-based methods are weak for localizing information acquisition, easily ignore detailed information, and are of high computational complexity. Also, the variation of target sizes challenges the generalization of networks. To address these issues, we propose a Siamese network named as CBSASNet for remote sensing CD, in which channel bias split attention (CBSA) is employed to recover the information in the change region, and the cross-temporal fusion module (CTFM) is utilized to highlight the information of change regions through the optimized single-temporal image features. The experimental result indicates that CBSASNet does not only outperform 16 state-of-the-art works, but also its modules complement each other in the ablation testing.**

*Index Terms*— **Change detection (CD), channel bias split attention (CBSA), convolutional neural networks (CNNs), fusion, remote sensing image.**

## I. Introduction

REMOTE sensing image change detection (CD) is to generate change maps by comparing and analyzing two-phase or multitemporal images. The pixels in the change graph have only two values of 0 and 1, that is, the change graph is a two-value change graph, where 0 and 1 represent change and invariance, respectively. CD techniques are widely used in urban planning [1], [2], ecosystem detection [3], [4], land cover analysis [5], and natural disaster damage assessment [6]. As a result, it can be seen that the development of these applications relies heavily on the progress of remote sensing image CD technology, which has attracted great attention from related scholars and a large number of CD methods have been proposed. Currently, emerging CD methods can be broadly categorized into two main groups: traditional methods and deep learning-based methods.

Early traditional remote sensing image CD methods relied on manually analyzing the spectral information of the image and manually selecting appropriate thresholds to identify the change regions, such as principal component analysis (PCA) [7], [8], Gabor filter [9], the tasseled cap transformation [10], multivariate alteration detection (MAD) [11], and change vector analysis (CVA) [12], [13]. However, the selection of these methods is affected by seasonal variations, lighting conditions, and satellite sensors, making it difficult to achieve robustness for higher CD accuracy. Therefore, the superiority of machine learning methods comes out, which can learn to summarize the laws through some samples, and then make decisions and predictions, such as random forest [14], support vector machine (SVM) [15], and decision tree [16]. However, the manually extracted features rely heavily on a priori domain knowledge, and, are less capable of representing deep change information, which limits the generalization ability of the methods.

In recent years, convolutional neural networks (CNNs) have achieved great success in various fields such as image classification [17], [18], semantic segmentation [19], and target detection [20], [21], which has led researchers to be interested in introducing CNNs into the field of remote sensing image CD. The CD task can generally be regarded as an image segmentation task, but unlike a general segmentation task, the CD task requires the input to be a set of dual-temporal images or multitemporal images. Based on this situation, Zhan et al. [22] made a breakthrough by introducing a Siamese convolutional network into the CD task, which achieves the goal of reducing the number of network parameters by processing dual-time images in parallel. However, the convolutional operation is limited by the size of the receptive field, making it difficult to effectively model global information. Some scholars have effectively improved the network's ability to capture global information by using the dilated convolution [23] and adding the attention mechanism [23], [24]. Daudt et al. [25] proposed three effective CD architectures by changing the fusion time and fusion method of the dual-temporal images. However, it is difficult to fuse the dual-time images very effectively by simply fusing the dual-temporal images by cascading or differencing.

Manuscript received 25 December 2023; revised 18 April 2024; accepted 30 April 2024. Date of publication 24 May 2024; date of current version 8 July 2024. This work was supported in part by the Scientific and Technological Innovation 2030 Major Project under Grant 2022ZD0115800, in part by the Xinjiang Uygur Autonomous Region Tianshan Talent Program Project under Grant 2022TSYCLJ0036, and in part by the Postgraduate Scientific Research Innovation Project of Xinjiang Uygur Autonomous Region under Grant XJ2023G098. *(Corresponding author: Liejun Wang.)*

The authors are with the School of Computer Science and Technology, Xinjiang University, Ürümqi 830046, China (e-mail: hnw@stu.xju.edu.cn; wljxju@xju.edu.cn; 007652@xju.edu.cn; zhangcui2020@stu.xju.edu.cn; lewis268@stu.xju.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3405025

Concatenated networks extract and refine features of the input data step by step through multiple subnetworks in a sequential manner. This structure allows flexibility in adjusting the number and type of subnetworks according to the needs of a particular task and has therefore been introduced by researchers into the field of CD. Heidary et al. [26] proposed a CD method based on a fully convolutional twin cascade network, which improves prediction accuracy by preprocessing and introducing an attention gate layer in place of Gaussian attention. Du et al. [27] proposed an end-to-end CD method called MTCDN. MTCDN achieves CD by integrating image generation, image recognition, and CD into a GAN, using a cyclic structure to unify optical and SAR images in a single feature domain. In addition, MTCDN achieves CD in heterogeneous images using a Siamese cascade network of UNet++. Huang et al. [28] proposed SEIFNet, which extracts multiscale features through a twin cascade network, highlights the change region using ST-DEM, and decodes it incrementally through ACFM and RM, and finally aggregates the multiscale decoded features to obtain the change results. Heidary et al. [29] proposed a method CTS-Unet for urban CD, which combines CNN with Swin transformer to form a cascade network that uses spatial and contextual relationships to achieve CD.

Currently, vision transformers (ViTs) [30] have become a promising alternative to CNNs for learning visual representations. Leveraging the powerful long-range modeling capabilities of ViTs, scholars have introduced them into the field of remote sensing CD and have proposed many effective methods. Yan et al. [31] proposed FTN, which constructs a pure transformer-based CD network. Zhang et al. [32] proposed SwinSUNet, incorporating the Swin transformer [33] into the CD field. In addition to pure transformer networks, researchers have also attempted to combine transformers with other methods to take advantage of both. BIT [34] and MFAT-Net [35] use a combination of transformer and ResNet-18 to detect change targets. Tang et al. [36] proposed WNet, which employs CNN and transformer to build twin networks for parallel processing of dual-temporal images. Bandara and Patel [37] proposed ChangeFormer, which combines a transformer with MLP for remote sensing CD.

Dual-temporal images or multitemporal images used in remote sensing image CD are images taken at the same location at different points in time, so even if the shooting location is the same, the spectral information of the two images is still very different due to the influence of many objective factors such as seasons, illumination, and sensors, which causes a great deal of trouble for the matching of contextual information of the two images. The vast majority of current CD methods use CNN methods in the encoding stage, which use a fixed-size convolution to downsample the feature information as a way to obtain feature information at different scales. Due to the limited size of the receptive field of a single convolutional kernel, not enough contextual information can be extracted in the feature extraction phase, which reduces the performance of the model. Although it is possible to increase the receptive field of the model by increasing the depth of the model, however, increasing the depth of the model too much is also a great burden on the hardware devices. The existing fusion mainly contains early fusion, cascade, or difference, these fusion methods will cause semantic deviation during feature fusion, which will lead to error detection.

Based on the above observations, we designed a network called CBSASNet in the hope of solving the CD task using a purely convolutional approach. CBSASNet guarantees the performance of the model by balancing the depth of the model with the size of the receptive field using channel bias split attention (CBSA). In addition, the method also changes the fusion method by using a cross-time fusion module to effectively utilize the extracted information to highlight the change region information. The main contributions of this article can be summarized as follows.

1) We designed CBSA to help the network extract detailed information better under the premise of integrating multiple receptive fields by adding a simple channel mapping.
2) We designed a cross-temporal fusion module (CTFM). This module subdivides the cascaded bi-temporal features through simple convolution and reinforces the difference information through a unilateral shortcut connection, emphasizing the changing areas.
3) We have constructed a Siamese network for remote sensing CD (CBSASNet) using the CBSA module and CTFM, and optimized the network during the training phase with a binary cross-entropy loss function. The CBSA module extracts and fuses multiscale information, while the CTFM module emphasizes the different information. Together, they effectively enhance the accuracy of CD.
4) We conducted extensive experiments on three public datasets, WHU-CD, CDD, and LEVIR-CD, to verify the effectiveness and superiority of the proposed method. Compared with the 16 state-of-the-art CD methods, CBSANet improved the $F1$-score on the three datasets by 0.86%, 0.19%, and 0.70%, respectively.

The remainder of the article is organized as follows. Section II briefly describes the work related to the CD task. In Section III, we present the detailed structure of each part of CBSASNet and the overall. Section IV provides experimental results that validate the effectiveness of CBSANet, and discusses and analyzes them. In Section V, we summarize the work done.

## II. Related Work

### A. CNN-Based Networks

CNNs have powerful feature extraction capabilities and have achieved commendable results in many fields, and many previous CD methods have been implemented based on CNNs. It is not difficult to find that researchers usually enhance the performance ability of the network by changing the structure of the backbone network, optimizing the loss function, and adding the attention mechanism. In terms of the backbone network structure, the unique structure of the

Siamese network is well adapted to the dual-time-phase remote sensing image CD-based task, so it is frequently applied to the research of dual-time-phase remote sensing image CD task. Zhan et al. [22] made a breakthrough in introducing Siamese networks to the CD task and proposed a shared weights Siamese network utilizing dual tributaries to process dual time-phase images simultaneously, which has become a benchmark for most of the later researchers on the task. Daudt et al. [25] proposed three fully convolutional-based network architectures that perform end-to-end training, where FC-EF takes the cascaded dual temporal images as the input to the network, while the two networks, FC-Siam-conc and FC-Siam-diff, use the Siamese network architecture, which takes the dual temporal images directly as the input to the network. In order to improve the lack of feature extraction capability in convolutional operations, Zhang et al. [23] used the dilated convolution in order to improve the acceptance domain of the convolutional kernel. Fang et al. [24] used dense hopping connections based on U-Net to maintain feature information and localization information. Similarly, Li et al. [38] proposed the DARNet which also utilized dense skip connections to aggregate features of different scales. Zhong and Wu [39] proposed the T-UNet, which emphasized highlighting the change information between two time-phase images through a three-branch structure to accurately identify the edges of changing objects.

In the remote sensing CD mission, there is a serious imbalance in the dataset data due to the fact that the changing pixels are much smaller than the unchanging pixels in the dual-temporal remote sensing images. This is an inherent characteristic of the dataset itself, and researchers have taken a variety of approaches to reduce or even eliminate the effects of data imbalance, most commonly by improving the loss function and equalizing the proportion of changing and unchanging pixels involved in the loss calculation. Zhan et al. [22] used a weighted contrastive loss to increase the weight of varying pixels in the loss calculation to reduce the effect of data imbalance in training. Chen et al. [40] proposed the weighted double-margin contrastive loss based on the traditional contrast loss, which improves the contribution of the two feature pairs to the value of the loss by decreasing the weight of the invariant feature pair and increasing the weight of the changing feature pair. Chen et al. [34] used minimizing cross-entropy loss to optimize the parameters of the network during the training phase of the network. Fang et al. [24] used a hybrid loss consisting of a weighted cross-entropy loss and a dice coefficient loss to optimize the network parameters. Feng et al. [41] used a deep supervision approach, thus facilitating the network can get more accurate prediction maps. Yan et al. [31] used a hybrid loss function along with deep supervision of different levels of features to calculate the loss values.

The attention mechanism used in computer vision (CV) was proposed by some scholars after being inspired by the visual ability of animals, which mimics the ability of animals to highlight key information when observing things, and focuses on specific parts or features of an image after it is inputted into the network, so that the network can dynamically assign different weights according to the importance of the inputs, inhibit irrelevant information, and emphasize the key information. Chen and Shi [42] designed a pyramid spatial–temporal attention module that can extract multiscale spatio-temporal contextual features to enhance the ability to recognize detailed information. Similarly, Li et al. [38] proposed a hybrid attention module that effectively fuses multiscale feature information through the efficient spatial–temporal attention module and the channel attention module. Li et al. [43] used a dynamic selection mechanism that can adaptively select the sensory field based on input information. Zhang et al. [44] added feature-map split attention in each module of ResNet [17] and selected weights based on global contextual information and proposed ResNeSt. Xu et al. [45] simplified the number of subblocks based on ResNeSt. Chen et al. [40] used a dual-attention mechanism to capture long-range dependencies between pixels. Zhao et al. [46] used channel attention to aggregate the four outputs of U-Net++ to generate a more accurate variation map. Zhong and Wu [39] designed a multibranch spatial-spectral cross-attention module using dual-temporal image features to correct the change information of the difference image. Feng et al. [41] proposed an intertemporal joint-attention module so that dual-temporal image features can guide each other. Zhao et al. [46] proposed using a geospatial position matching mechanism and a geospatial content reasoning mechanism to effectively extract global information and refine features.

## B. Transformer-Based Networks

In the last two years, the transformer [30] method, originally used in NLP tasks, has been widely used in CV tasks. Dosovitskiy et al. [47] proposed ViT, which is the first fully transformer-based network used in the field of CV and achieved results comparable to the results of CNN methods in the field of image classification. The proposal of ViT started a new framework in the field of CV. Zhang et al. [32] introduced Swin transformer into remote sensing CD task and constructed SwinSUNet, which is a pure Swin transformer [33] network with a concatenated U-shaped structure. Swin transformer effectively solves the high computational complexity of the transformer by replacing the standard multihead self-attention module with a window-based multihead self-attention module and also achieves better results in the CD task. Yan et al. [31] designed a pure transformer network FTN to extract features from a global perspective and fused different levels of features using a pyramid approach. Zheng et al. [48] proposed that Changemask can be used for semantic CD. Chen et al. [34] tried to combine CNN and transformer methods to propose the BIT model, which uses CNN to extract feature maps and a transformer to model contextual relationships based on feature maps, which combines the advantages of the two in terms of local feature extraction and global features. Feng et al. [49] proposed ICIF-Net based on BIT to solve the multiscale feature interaction neglected by BIT. However, the difficulty that transformer-based methods need to occupy a large amount of computational resources has been troubling researchers.

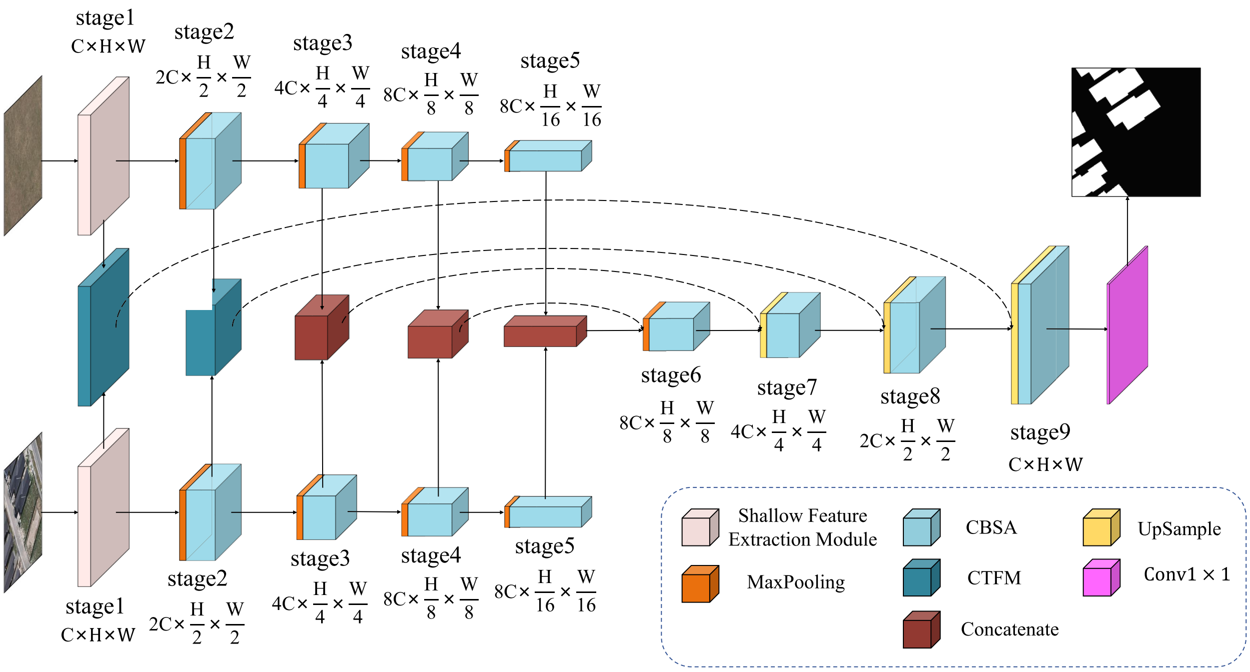


Fig. 1. Architecture of the proposed CBSASNet.

## III. Materials and Methods

This chapter introduces our CD method, CBSASNet, and Fig. 1 presents the complete architecture of CBSASNet. To enhance the network's ability to extract both global information and local detail information, we first use the CBSA module to construct a twin network as an encoder to process dual-temporal images in parallel. This module helps the network to better extract detailed information by integrating multiple receptive fields. Second, we propose the CTFM to fuse features from dual-temporal images, reinforcing the difference information through unilateral shortcut connections, thereby highlighting the areas of change.

### A. CBSA Module

*1) Motivation:* The number of stacked layers can provide information under different receptive fields, thereby enriching the features extracted by the network, such as ResNet [17]. Nevertheless, an excessive number of stacked layers inevitably leads to a surge in the number of parameters, thereby increasing the computational load. By using split-attention blocks, ResNeSt [44] retains the basic structure of ResNet and integrates information from multiscale receptive fields at one level very effectively, achieving better results without additional computational burden. However, there does exist a tradeoff between the captures of global and detailed information. HRViT [50] increases channel mapping to enhance nonlinear capability. Inspired by this, we designed CBSA to help the network extract detailed information better under the premise of integrating multiple receptive fields by adding a simple channel mapping. In short, after the split-attention block integrates the information extracted from multiscale receptive fields through global attention, a channel mapping of the original features is added to the features to enhance the nonlinear mapping capability of the feature channel dimensions, thus increasing the diversity of features and enhancing the network's fitting ability.

*2) CBSA Module:* The framework diagram of CBSA is shown in Fig. 2. We assume that the input received by CBSA is $d_1 \in \mathcal{R}^{C\times H\times W}$. First, $d_1$ is fed into a 1 × 1 convolution. the 1 × 1 convolution is responsible for adjusting the dimensionality (decreasing or increasing $C$). After that, the adjusted $d_1$ is divided equally into $d_1^1 \in \mathcal{R}^{C\times H\times W}$ and $d_1^2 \in \mathcal{R}^{C\times H\times W}$ and fed to the left and right branches, respectively. The left branch undergoes a 3 × 3 convolution and the right branch undergoes two 3 × 3 convolutions, which are described in more detail as shown below

$$d_1^1, d_1^2 = \mathcal{S}(\mathrm{Conv}_{1\times1}(d_1)) \tag{1}$$

$$d_2^1 = \mathrm{Conv}_{3\times3}\left(d_1^1\right) \tag{2}$$

$$d_2^2 = \mathrm{Conv}_{3\times3}\left(d_1^2\right) \tag{3}$$

where $\mathcal{S}(\cdot)$ denotes the average chunking in the channel dimension, and $\mathrm{Conv}_{1\times1}$ and $\mathrm{Conv}_{3\times3}$ denote the vanilla convolution containing a batch normalization and ReLU activation function with convolution kernel sizes of 1 × 1 and 3 × 3, respectively. Then, in order to be able to make the information of the two branches complement each other and not be isolated, $d_2^1$ is summed up with $d_2^2$ element-by-element and then a 3 × 3 convolutional layer is used to refine the information contained in $d_2^1$ and $d_2^2$ to obtain $d_3^1 \in \mathcal{R}^{(c/2)\times H\times W}$ and are then summed up through the element-by-element summation operation to obtain $d_3^3 \in \mathcal{R}^{(c/2)\times H\times W}$, as follows:

$$d_3^1 = \mathrm{Conv}_{3\times3}\left(d_2^1 \oplus d_2^2\right) \tag{4}$$

$$d_3^3 = d_3^1 \oplus d_2^2 \tag{5}$$

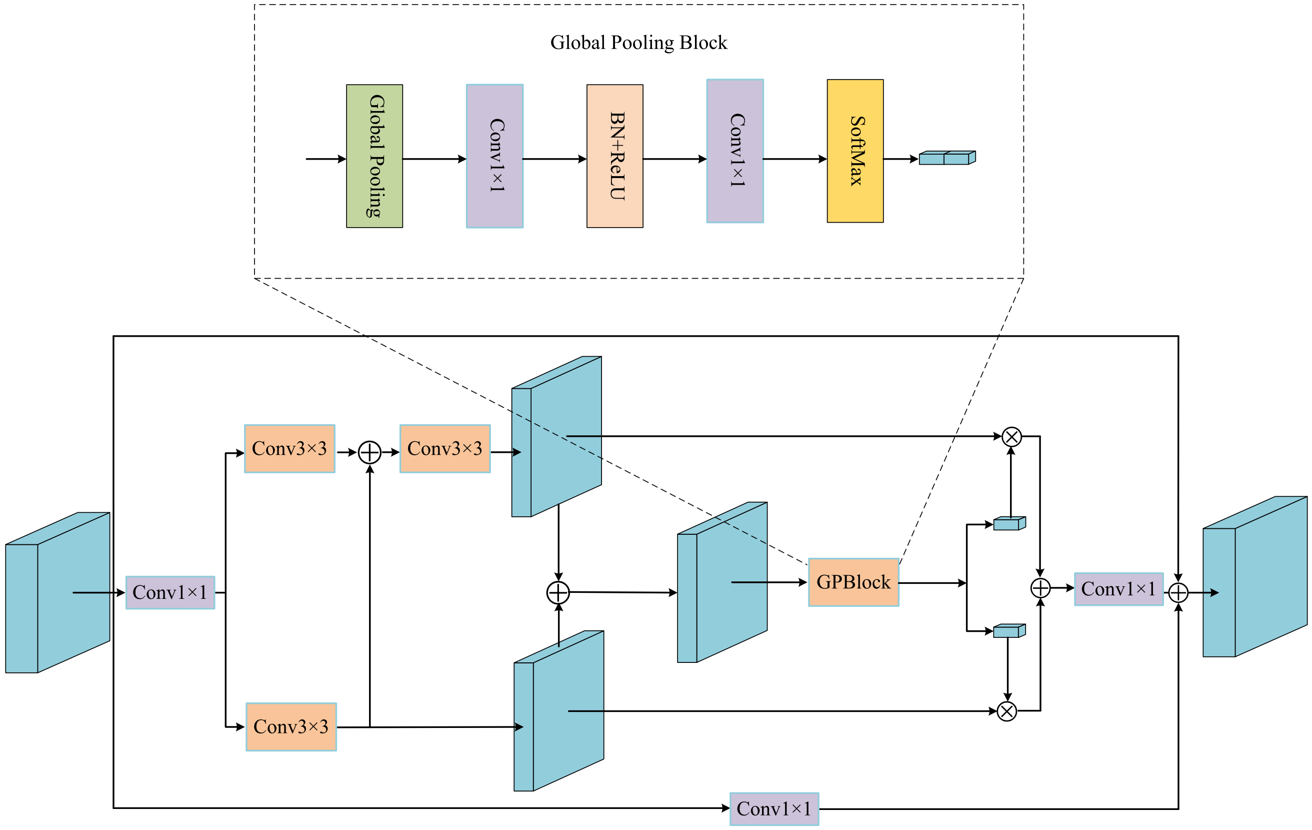


Fig. 2. Architecture of the proposed CBSA.

where $\oplus$ denotes the element-by-element summation operation. After that $d_4^3 \in \mathcal{R}^{(c/2)\times 1\times 1}$ is obtained from $d_3^3$ using global average pooling operation, $F$ retains the most important information in the feature map, and discards other unimportant information and noise generated during the feature extraction phase. After that, the channel information of $d_4^3$ is adjusted by two layers of $1 \times 1$ convolution to get $d_5^3 \in \mathcal{R}^{C\times 1\times 1}$. After that, the SoftMax function is used to normalize $d_5^3$. Then, the normalized $d_5^3$ is divided equally into $d_6^4$ and $d_6^5$ in the channel dimension and multiplied with $d_3^1$ and $d_2^2$, respectively, and the global information obtained is used to enhance the representation of the feature map with respect to the key information. It is shown as follows:

$$d_4^3 = \mathcal{A}\left(d_3^3\right) \tag{6}$$

$$d_5^3 = \mathrm{ReLU}\left(\mathrm{Conv}_1\left(\mathrm{Conv}_{1\times 1}\left(d_4^3\right)\right)\right) \tag{7}$$

$$d_6^4, d_6^5 = \mathcal{S}\left(\sigma\left(d_5^3\right)\right) \tag{8}$$

$$d_6^6 = \left(d_3^1 \otimes d_6^4\right) \oplus \left(d_2^2 \otimes d_6^5\right) \tag{9}$$

where $\mathcal{A}(\cdot)$ denotes global pooling, $\sigma(\cdot)$ denotes the mapping of feature information to $[0, 1]$ using the SaftMax function in the channel dimension, $\otimes$ denotes the element-by-element multiplication operation, $\mathrm{Conv}_1(\cdot)$ denotes the inclusion of a batch-normalized $1 \times 1$ convolutional layer, and $\mathrm{ReLU}(\cdot)$ denotes the ReLU activation function. Finally, the results of the $1 \times 1$ convolution of the original $d_1$ and $d_1$ are used as residual connections. Then, the results of the $1 \times 1$ convolution of each element with $d_6^6$ are added. This process can be demonstrated as

$$d_1' = \mathrm{ReLU}\left(d_1 \oplus \mathrm{Conv}_{1\times 1}(d_1) \oplus \mathrm{Conv}_{1\times 1}\left(d_6^6\right)\right). \tag{10}$$

### B. Cross-Temporal Fusion Module

*1) Motivation:* The captured bi-temporal images are influenced by the time of shooting, leading to interference from irrelevant changes such as seasons and lighting conditions [41]. In real-world scenarios, there does exist an imbalance between the foreground and background. This kind of imbalance poses a major challenge in CD: a cascading-based approach can ensure the completeness of bi-temporal image information during the decoding process [24], [25] while it is difficult to highlight the differences between nonchanging and changing areas. It is observed that the fusion of bi-temporal image features can effectively highlight or amplify the differential signals of changing regions. To this end, we designed a CTFM. This module subdivides the cascaded bi-temporal features through simple convolution and reinforces the difference information through a unilateral shortcut connection, emphasizing the changing areas.

*2) Cross-Temporal Fusion Module:* As shown in Fig. 3. The CTFM takes $f_i^1 \in \mathcal{R}^{C\times H\times W}$ and $f_i^2 \in \mathcal{R}^{C\times H\times W}$ as inputs, and $f_i^1$ and $f_i^2$ are each passed through a $3 \times 3$ convolutional layer before doing preliminary fusion using a channel dimension cascade to obtain $f_i^{31}$. After that, the fused result $f_i^{31}$ is passed through two $3 \times 3$ convolutional layers to obtain $f_i^{32} \in \mathcal{R}^{C\times H\times W}$, which is used to smooth out the difference information in the initial fused features. This process can be

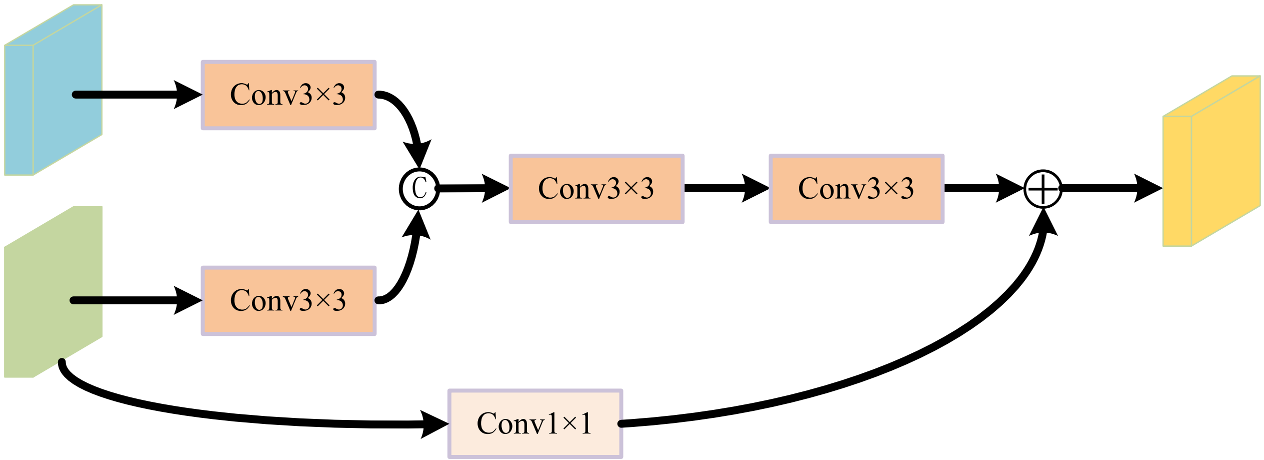


Fig. 3. Diagram of the proposed CTFM.

represented as follows:

$$f_i^{31} = \text{Cat}\left(\text{Conv}_{3\times3}\left(f_i^1\right), \text{Conv}_{3\times3}\left(f_i^2\right)\right) \tag{11}$$

$$f_i^{32} = \text{Conv}_{3\times3}\left(\text{Conv}_{3\times3}\left(f_i^3\right)\right) \tag{12}$$

where Cat$(\cdot, \cdot)$ denotes the superposition of two features in the channel dimension. Afterward, in order to be more prominent in highlighting the difference information between the dual-temporal images, $f_i^1$ is added element-by-element with $f_i^{32}$ through a 1 × 1 convolutional layer to obtain the final result $f_i^{33}$. We believe that the element-by-element summation of $f_i^1$ and $f_i^{33}$ can both dilute out the noise introduced due to the initial fusion and prevent some of the effective difference information from being lost and also alleviate the later decoder stage of decoding the information to fuse the dual branch information roughness. Mathematically it can be expressed as

$$f_i^{33} = \text{ReLU}\left(\text{Conv}_1\left(f_i^1\right) \oplus f_i^{32}\right). \tag{13}$$

### C. Construction of CBSASNet

*1) Encoder:* We use the Siamese network to extract image features from dual-temporal remote sensing images. Specifically, the network of CBSASNet uses a multiscale shallow feature extraction module to roughly extract the shallow features of the image. As shown in Fig. 1, the shallow feature extraction module mainly consists of a 7 × 7 vanilla convolution, a 7 × 7 depthwise convolution, and a 1 × 1 pointwise convolution, and applies a shortcut connection strategy to cope with possible gradient vanishing. Since the remote sensing CD task needs to detect targets of different sizes, four feature extraction layers are used after the shallow feature extraction module, and in order to balance the number of parameters and performance as much as possible, each feature extraction layer consists of only two CBSAs, which can be used to extract cross-channel information at different scales of receptive fields, highlighting the features of targets of different sizes, and better identifying the changing targets.

*2) Fusion:* The role of fusion is to fuse the dual-temporal phase image features and highlight the different information in the dual-temporal phase image, while by fusing the dual-temporal phase image features directly through the cascade, it is easy to cause semantic bias, which leads to the omission of the change target or wrong detection. The first two phases of the CBSASNet use the fuser, which can not only dilute out the noise introduced due to the initial fusion by reinforcing the unilateral information but also can prevent the loss of some of the effective difference information and highlight the differences in the dual-time phase image features. In the deeper layers of the network, semantic information represents more ambiguous meanings and allows for more information bias, and it will be easier and faster to fuse features by cascading.

*3) Decoder:* In order to recover the image features obtained by downsampling from the encoder and thus obtain the final prediction map, we construct a decoder that is essentially symmetric with the encoder using CBSA. The decoder is divided into five stages, each of the first four stages first uses an upsampling layer that receives the features from the previous stage and performs an upsampling operation on it, then cascades them with the two-branch fusion features from the stage corresponding to the encoder, and finally inputs them into the CBSA to recover the image features. In the final stage of the decoder the feature map channel is compressed to 2 using dot convolution and the prediction result is obtained using softmax function.

### D. Loss Function

We use the binary cross-entropy loss to optimize the parameters of the network in the training phase of the network. Mathematically, the binary cross-entropy loss can be expressed as

$$\mathcal{L} = \frac{1}{\mathcal{H}\times\mathcal{W}} \sum_{j=1}^{\mathcal{H}\times\mathcal{W}} \left(\delta_{G_j}\cdot G_j \log(P_j) + \delta_{\hat{G}_j}\cdot \hat{G}_j \log(\hat{P}_j)\right) \tag{14}$$

where $G_j$ denotes the label of the $j$th pixel in the ground truth map, and if $G_j$ is 0 then $\overline{G}_j$ is 1, at which point $G_j$ indicates that the semantics of the corresponding dual-temporal phase image at that location changes, and $\overline{G}_j$ is the opposite.

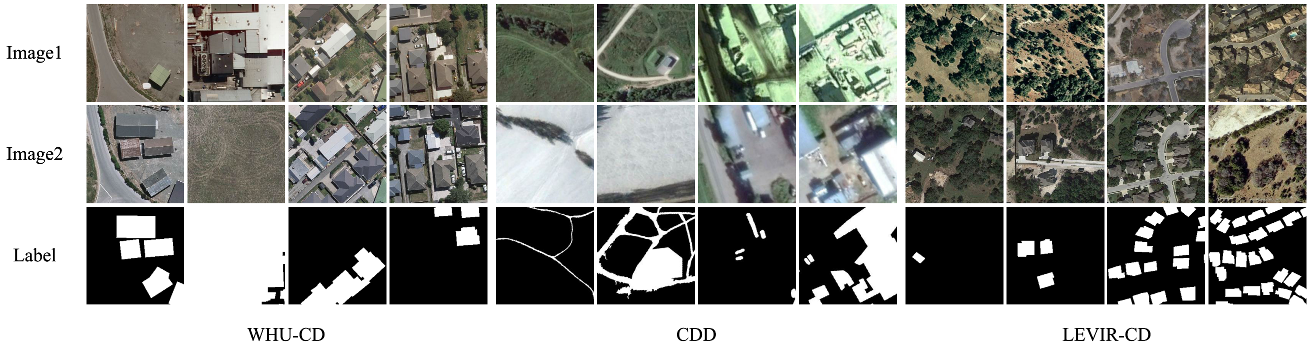


Fig. 4. Sample dataset presentation. (Top to bottom) T1 images, T2 images, and ground truth. Samples from one dataset per four columns. (Left to right) WHU-CD, CDD, and LEVIR-CD.

TABLE I
MAIN PARAMETERS OF THE DATASET

| Dataset | Patch Size | Number of Pixels Changed | Number of Pixels Unchanged | Ratio | Train | Val | Test |
|---|---|---|---|---|---|---|---|
| WHU-CD | 256×256 | 21,352,815 | 477,759,663 | 1:22.37 | 5,948 | 743 | 743 |
| CDD | 256×256 | 134,068,750 | 914,376,178 | 1:6.82 | 10,000 | 3,000 | 3,000 |
| LEVIR-CD | 256×256 | 30,913,975 | 637,028,937 | 1:20.61 | 7,120 | 1,024 | 2,048 |

$P_j$ denotes the probability that the predicted image element represents the corresponding location of the dual-temporal phase image that changes, and $\overline{P}_j$ denotes the probability that the pixel at that location does not change. $\delta_{G_j}$ and $\delta_{\hat{G}_j}$ denote the weights of changed and unchanged pixels, respectively, to adjust the relative importance of each category.

## IV. EXPERIMENTS

### A. Datasets

*1) WHU-CD [51]:* The imagery in this dataset uses two aerial images from Christchurch, New Zealand, taken at different times, one in 2012 and the other in 2016. The dataset images were taken at a location that was struck by a magnitude 6.3 earthquake in February 2011, after which the area was rebuilt. The aerial imagery covers a 450-km$^2$ area of Christchurch, New Zealand, with a spatial resolution of 0.075 m. The image contains a large number of buildings, and it is these buildings that we detect changes. We cropped the original image of 32 507 × 15 354 pixels without overlapping into blocks of 256 × 256 pixels to obtain a total of 7434 pairs and randomly assigned all the image blocks into a training set, validation set and test set with an allocation ratio of 8:1:1.

*2) CDD [52]:* The dataset images are seasonally changing images of the same area obtained from Google Earth and contain seven pairs of 4725 × 2700 pixels and four pairs of 1900 × 1000 pixels images with spatial resolutions ranging from 3 to 100 cm/px, and the images contain automobiles and large buildings that can provide objects of varying sizes for the CD. The images are cropped into 256 × 256 sized fragments by random rotation to get a total of 16 000 image pairs, and finally, the image set is divided into 10 000 training sets, and 3000 test and validation sets.

*3) LEVIR-CD [42]:* This dataset consists of 637 image pairs collected from Google Earth taken in 20 different areas of several cities in Texas, USA, from 2002 to 2018, with a spatial resolution of 0.5 m/pixel for the acquired pairs, and a size of 1024 × 1024 pixels for each pair. The dataset covers a wide range of building types, including a variety of homes, garages, and large warehouses. All images were cropped into nonoverlapping 256 × 256-pixel image blocks and divided in a 7:1:2 ratio to obtain 7120 (training), 1024 (validation), and 2048 (test) pairs of patches.

Fig. 4 gives a presentation of some of the images in the three datasets. It can be seen from Fig. 4 that the three datasets have different focuses, where the WHU-CD dataset focuses on sparse and large building changes, the CDD dataset focuses on changes caused by buildings, thin and irregular roads, and traffic vehicles, and the LEVIR-CD dataset focuses on changes caused by sparse or dense small buildings. Detailed information on the three datasets is provided in Table I, and the severe imbalance between changing and unchanging pixels in the three datasets is also a major challenge in the CD task.

### B. Evaluation Indicators

In order to fully evaluate our proposed CBSASNet, we use six evaluation metrics, precision (Pre), recall (Rec), $F1$-score ($F1$), IoU, overall accuracy (OA), and kappa, which are all in the interval [0, 1], when the value is closer to 1, it means that the network is more effective. The expressions for these six evaluation metrics are as follows:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{15}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{16}$$

$$F1 = 2/(\text{Recall}^{-1} + \text{Precision}^{-1}) \tag{17}$$

$$\text{IoU} = \text{TP}/(\text{TP} + \text{FP} + \text{FN}) \tag{18}$$

$$\text{OA} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{19}$$

$$\text{Kappa} = (\text{OA} - P)/(1 - P) \tag{20}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. $P$ in kappa is the intermediate variable in the computation, which denotes the hypothesized probability of chance consistency between the target and the prediction, and it can be expressed as

$$P = \frac{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})+(\mathrm{FN}+\mathrm{TN})(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN})^2}. \quad (21)$$

### C. Experimental Details

All of our experiments were implemented in the Pytorch deep learning framework. Experiments were conducted using a model Nvidia Titan RTX (24G) GPU with batch size set to 8. The Adam network optimizer was used with the two decay factors set to 0.9 and 0.999 and the learning rate set to 0.0001. Our network is trained for 300 epochs on all datasets, and in each epoch we sequentially perform one training and one validation, keeping the model parameters with the highest $F1$-scores. The models with $F1$-scores in the validation set will be used to evaluate the test set.

### D. Comparative Experiments

*1) Comparison of Methods:* We compared CBSASNet with several classical and excellent methods for CD, such as FC-Siam-conc [25], FC-Siam-diff [25], FC-EF [25], BIT [34], IFN [53], SNUNet-CD [24], ICIF-Net [49], PA-Former [54], RDP-Net [55], DMINet [41], GeSANet [46], AERNet [56], HANet [57], WNet [36], ScratchFormer [58], and SEIFNet [28].

1) *FC-Siam-Conc [25]:* It has a similar structure to FC-EF, also using a U-shaped structure, the difference is that FC-Siam-conc uses the encoder of FC-EF twice to form two encoder branches to process the bi-phasic images separately, and they share weights between them, and the results obtained are cascaded in the channel dimension.
2) *FC-Siam-Diff [25]:* Also a variant of FC-EF, the difference with FC-Siam-conc is that the shortcut connection of FC-Siam-diff uses the absolute value of the difference between the two encoded branches.
3) *FC-EF [25]:* A classical fully convolutional CD network with a U-shaped structure, the network uses a dual-temporal image fused image as input and four hop connections between the encoder and decoder.
4) *BIT [34]:* This network first introduces a transformer into the field of CD and adopts a hybrid approach of CNN and transformer to construct the network. The long-range relationship between image features is effectively constructed by the transformer.
5) *IFN [53]:* This network uses VGG16 as a backbone network for DFEN to extract bi-phasic image features, and finally the prediction maps are generated by the DDN. The feature maps of the layers extracted from DFEN are shortcuts to DDN layers with the same dimensions.
6) *SNUNet-CD [24]:* This network uses a densely connected Siamese network as the backbone network to reduce the loss of depth localization information, and the four-level depth features obtained from the backbone network are transmitted to the channel attention module for feature refinement. In addition, we use the SNUNet-CD model with an initial number of channels of 32, which is the most cost-effective model for this network.
7) *ICIF-Net [54]:* The network uses CNN and transformer to extract local and global features, a linearized conv attention module for feature information interaction at the same resolution, and an interscale feature fusion module to collect feature information at different resolutions.
8) *PA-Former [55]:* This network uses deep features extracted using ResNet18 and spatio-temporal information obtained using the transformer module that obtains a priori features and integrates them into the deep features.
9) *RDP-Net [49]:* Similar to FC-EF, this network also takes as input the fused image after cascading in the channel dimension. The input is segmented into image chunks by a region partitioning layer, the local information in the image chunks and the global information of the whole image are explored using ConvMixer, and the prediction map is obtained by fusing the multilayer depth output using the depth attention module.
10) *DMINet [41]:* It uses two ResNet18 networks to do the initial feature extraction, uses an attention module that unifies the self-attention mechanism and the cross-attention mechanism in a single module to guide the global feature distribution of the two-branch feature map, and employs subtraction and cascading to aggregate the two-branch image features.
11) *GeSANet [46]:* ResNet18 is used as the backbone network, which uses a geospatial position matching mechanism and a geospatial content reasoning mechanism to filter pseudo-change information.
12) *AERNet [56]:* The method uses ResNet34 to construct the Siamese network and uses a global contextual feature aggregation module to aggregate multilayer contextual feature focus information. Channel and positional associations between features are captured using an enhanced coordinate attention-guided attention decoding block, and the network's ability to perceive and refine the edges of changing regions is enhanced using an edge refinement module. Meanwhile, an adaptive weighted binary cross-entropy loss function combined with a deep supervision strategy is used to enhance the feature learning ability of the network in the presence of dataset imbalance.
13) *HANet [57]:* The method proposes a progressive foreground balanced sampling-based approach to solve the sample imbalance problem by gradually adding background images. In addition, a discriminative Siamese network is designed using the HAN module, in which the HAN module is able to capture the long-range relations efficiently.
14) *WNet [36]:* The method merges the twin CNN and twin transformer into the encoder to extract local fine-grained information and global remote context information while introducing the deformability idea to enhance the network's understanding of irregular regions. A differential enhancement module is embedded

TABLE II

COMPARATIVE RESULTS OF PRECISION, RECALL, $F1$-SCORE, IOU, OA, AND KAPPA FOR ALL METHODS ON THE WHU-CD DATASET, WITH THE OPTIMAL RESULTS SHOWN IN BOLD FONT

| Methods | Years | IoU(%) | Pre(%) | Rec(%) | F1(%) | OA(%) | Kappa(%) |
|---|---|---|---|---|---|---|---|
| FC-Siam-conc [25] | 2018 | 63.01 | 72.06 | 83.39 | 77.31 | 97.72 | 76.11 |
| FC-Siam-diff [25] | 2018 | 67.18 | 83.83 | 77.18 | 80.37 | 98.24 | 79.45 |
| FC-EF [25] | 2018 | 67.64 | 85.49 | 76.42 | 80.70 | 98.29 | 79.81 |
| BIT [34] | 2020 | 72.12 | 81.32 | 86.44 | 83.80 | 98.44 | 82.98 |
| IFN [53] | 2020 | 80.40 | **97.81** | 81.87 | 89.13 | 99.07 | 88.65 |
| SNUNet-CD [24] | 2021 | 76.97 | 89.65 | 84.47 | 86.98 | 98.82 | 86.37 |
| ICIF-Net [49] | 2022 | 82.33 | 94.60 | 86.39 | 90.31 | 99.13 | 89.86 |
| PA-Former [54] | 2022 | 74.05 | 85.04 | 85.15 | 85.09 | 98.61 | 84.36 |
| RDP-Net [55] | 2022 | 77.54 | 89.45 | 85.35 | 87.35 | 98.85 | 86.75 |
| DMINet [41] | 2023 | 74.64 | 84.86 | 86.11 | 85.48 | 98.63 | 84.76 |
| GeSANet [46] | 2023 | 84.61 | 95.05 | 88.51 | 91.66 | 99.25 | 91.27 |
| AERNet [56] | 2023 | 84.44 | 90.27 | 92.89 | 91.56 | 99.20 | 91.14 |
| HANet [57] | 2023 | 79.41 | 94.31 | 83.41 | 88.52 | 98.99 | 88.00 |
| WNet [36] | 2023 | 77.39 | 88.79 | 85.77 | 87.25 | 98.83 | 86.64 |
| ScratchFormer [58] | 2024 | 78.22 | 93.70 | 82.56 | 87.78 | 98.93 | 87.22 |
| SEIFNet [28] | 2024 | 83.84 | 89.91 | 92.55 | 91.21 | 99.17 | 90.77 |
| CBSASNet(ours) | 2024 | **86.08** | 93.93 | 91.15 | **92.52** | **99.31** | **92.16** |

in the encoder to obtain multilevel differential feature mapping by cascading and subtraction. The multilevel differential feature mappings are gradually fused by the CNN-transformer fusion module.

15) *ScratchFormer [58]:* The method utilizes hybrid sparse attention to construct a Siamese network that captures information in the change region, ameliorating the pitfalls of traditional self-attention mechanisms that have difficulty in capturing generalization bias when trained from scratch. In addition, a change-enhanced feature fusion module is introduced to fuse features of input image pairs by performing per-channel reweighting to enhance relevant semantic changes while suppressing noisy information.

16) *SEIFNet [28]:* The method first obtains a multilevel feature map from the Siamese hierarchical backbone network and introduces a spatio-temporal disparity enhancement module to capture the global information combined with local information in the dual-temporal-phase feature maps at each level. An adaptive context fusion module is designed and a progressive decoder is constructed using the adaptive context fusion module and the refinement module for integrating the features among the layers.

In order to make a fair comparison of all algorithms, all experiments were performed with the same experimental platform, the same data preprocessing method, and hyperparameters.

*2) Analysis of Experimental Results:* In order to be able to clearly compare the experimental result data, we present all the experimental data in the form of a table. At the same time, in order to be able to present our experimental results very graphically, we selected several pictures from each dataset to visualize the method results using different methods and compare them with labels. Wherein, the black area indicating true negative indicates that the pixel at the location is correctly predicted as a pixel that has not changed; the white area indicating true positive indicates that the pixel at the location is correctly predicted as a pixel that has changed; the green area indicating false negative indicates that the pixel at the location is incorrectly predicted as a pixel that has not changed; and the red area indicating false positive indicates that the pixel at the location is incorrectly predicted as a pixel that has changed.

*a) Experimental results on the WHU-CD dataset:* To evaluate the various methods, we trained and tested them on the WHU-CD dataset, and the results of the experiments are shown in Table II. Among the 17 methods, only five methods have $F1$-scores over 90%, which are ICIF-Net, GeSANet, AERNet, SEIFNet, and CBSASNet. CBSASNet outperforms all the other methods in four metrics, where it outperforms the second-ranked GeSANet in the IoU, $F1$, and kappa metrics by 1.47%, respectively, 0.86%, and 0.89%, respectively. Although CBSASNet is lower than IFN in the Pre metric, CBSASNet's Pre and Rec are closer to each other, indicating that compared to IFN, CBSASNet predicts a relatively balanced mix of changing pixels and unchanging pixels without being too biased toward one class, thus improving the overall prediction accuracy.

In order to show the experimental results more clearly, we randomly select several images from the WHU-CD dataset and use them as visualization results to show the results, and the visualization results are shown in Fig. 5. From Fig. 5, it can be observed that when detecting relatively large change targets, CBSASNet can detect the edge information of the change region more accurately and have fewer false detections and missed detections.

*b) Experimental results on CDD dataset:* Table III shows the experimental results of various methods on the CDD dataset. Among them, three methods IFN, WNet, and CBSASNet perform better with an $F1$-score of more than 97%. Among all the methods CBSASNet has the highest IoU, Rec, $F1$-score, OA, and kappa, especially Rec exceeds the second-ranked WNet by 0.35%.

In order to present the experimental results more clearly, we randomly selected several pictures from the CDD dataset and used them as visualization results to show the visualization

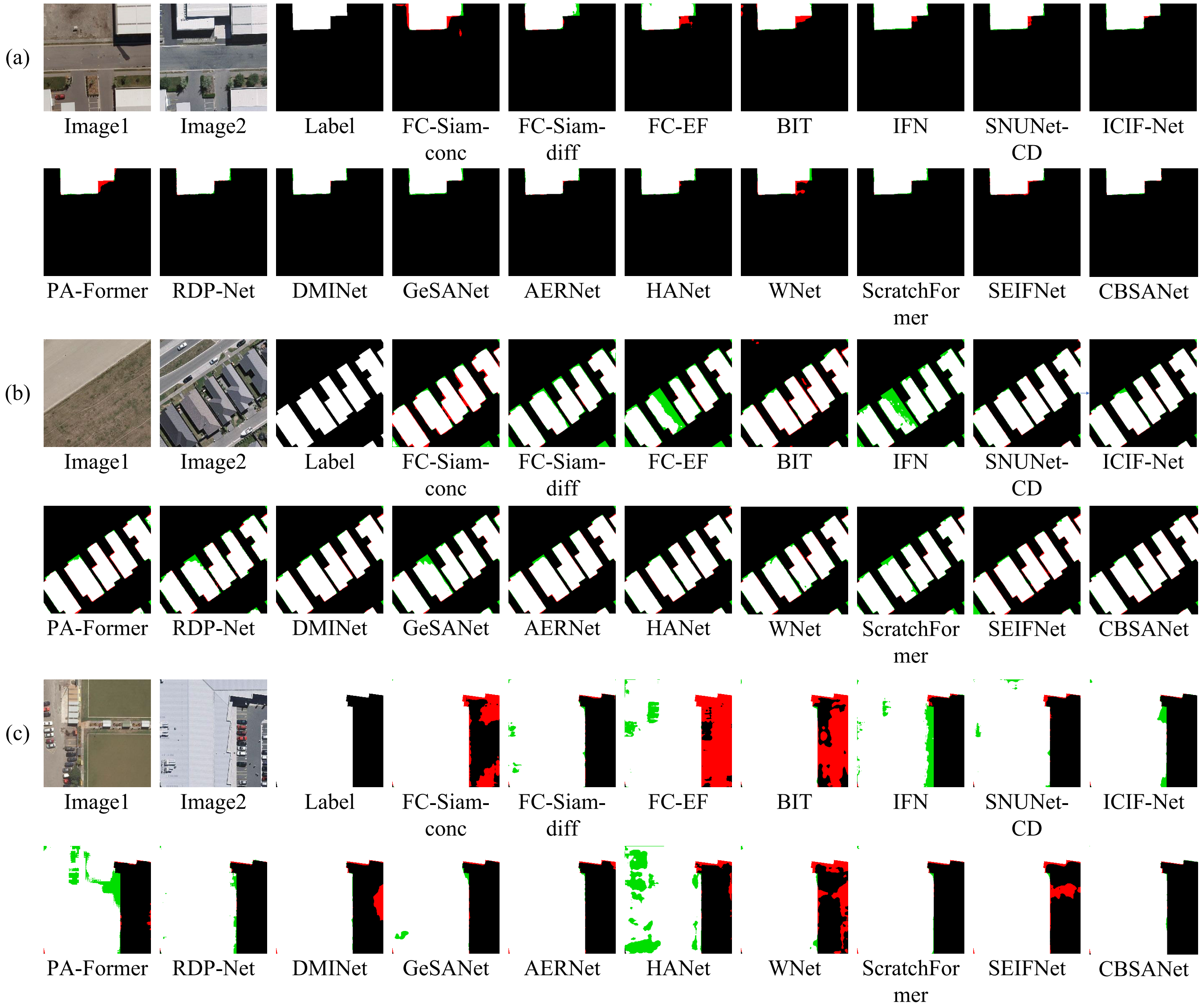


Fig. 5. Visualization results of the 16 CD methods on the WHU-CD dataset. Image rendering colors: white in true positive, red in false positive, black in true negative, and green in false negative. (a)–(c) Represent three randomly selected sample images from that dataset.

TABLE III

Comparative Results of Precision, Recall, $F1$-Score, IoU, OA, and Kappa for All Methods on the CDD Dataset, With the Optimal Results Shown in Bold Font

| **Methods** | **Years** | **IoU(%)** | **Pre(%)** | **Rec(%)** | **F1(%)** | **OA(%)** | **Kappa(%)** |
|---|---|---|---|---|---|---|---|
| FC-Siam-conc [25] | 2018 | 67.94 | 89.59 | 73.76 | 80.91 | 95.89 | 78.63 |
| FC-Siam-diff [25] | 2018 | 64.10 | 89.92 | 69.06 | 78.12 | 95.44 | 75.62 |
| FC-EF [25] | 2018 | 59.49 | 88.21 | 64.63 | 74.60 | 94.81 | 71.79 |
| BIT [34] | 2020 | 89.98 | 95.49 | 93.97 | 94.72 | 98.76 | 94.03 |
| IFN [53] | 2020 | 94.21 | **97.36** | 96.69 | 97.02 | 99.30 | 96.62 |
| SNUNet-CD [24] | 2021 | 93.72 | 96.74 | 96.78 | 96.76 | 99.23 | 96.32 |
| ICIF-Net [49] | 2022 | 93.25 | 96.31 | 96.70 | 96.50 | 99.17 | 96.04 |
| PA-Former [54] | 2022 | 89.54 | 95.24 | 93.73 | 94.48 | 98.71 | 93.75 |
| RDP-Net [55] | 2022 | 93.33 | 96.37 | 96.73 | 96.55 | 99.18 | 96.09 |
| DMINet [41] | 2023 | 91.99 | 95.78 | 95.88 | 95.83 | 99.02 | 95.27 |
| GeSANet [46] | 2023 | 91.92 | 95.97 | 95.61 | 95.79 | 99.01 | 95.23 |
| AERNet [56] | 2023 | 91.98 | 96.03 | 95.61 | 95.82 | 99.02 | 95.26 |
| HANet [57] | 2023 | 89.93 | 95.27 | 94.14 | 94.70 | 98.76 | 93.99 |
| WNet [36] | 2023 | 94.42 | 97.17 | 97.09 | 97.13 | 99.32 | 96.75 |
| ScratchFormer [58] | 2024 | 92.54 | 96.62 | 95.64 | 96.12 | 99.09 | 95.61 |
| SEIFNet [28] | 2024 | 92.75 | 96.68 | 95.81 | 96.24 | 99.12 | 95.74 |
| CBSASNet(ours) | 2024 | **94.58** | 96.99 | **97.44** | **97.21** | **99.34** | **96.84** |

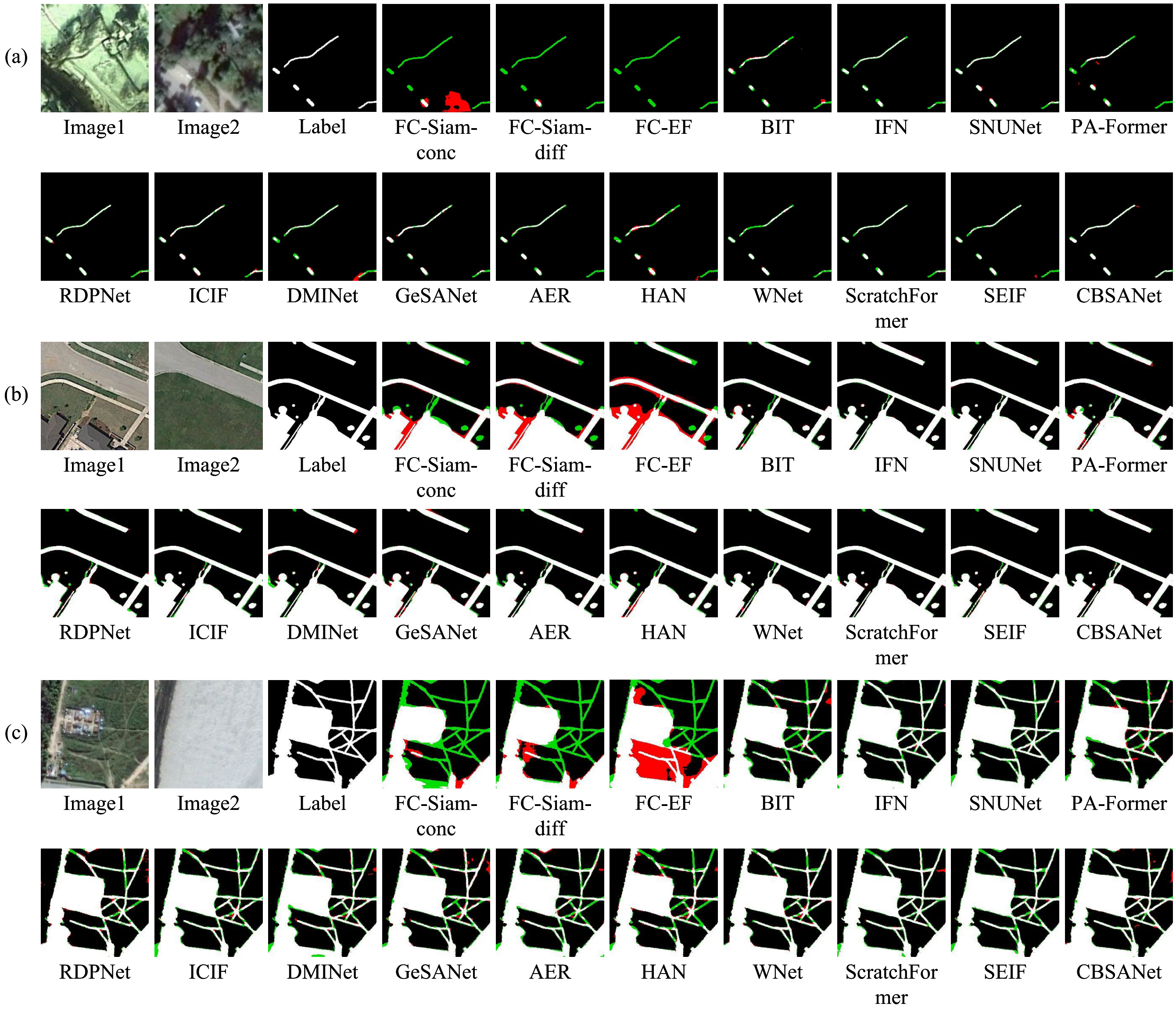


Fig. 6. Visualization results of the 16 CD methods on the CDD dataset. Image rendering colors: white in true positive, red in false positive, black in true negative, and green in false negative. (a)–(c) Represent three randomly selected sample images from that dataset.

results, which are shown in Fig. 6. The selected images contain three types of change regions, which are point-like, elongated, and large change regions. From Fig. 6, it can be seen that CBSASNet has fewer misdetections and omissions, and is able to detect the change regions more completely.

*c) Experimental results on LEVIR-CD dataset:* Table IV shows the experimental results of various methods on the LEVIR-CD dataset. The change regions in the LEVIR-CD dataset are mostly dense and small, which makes it more difficult for the detection of target boundaries. From Table IV, it can be seen that the IoU, $F1$-score, and kappa of CBSASNet outperform the second-ranked AERNet by 1.17%, 0.70%, and 0.73%, respectively.

The visualization results of all methods are presented in Fig. 7. The change areas in the LEVIR-CD dataset are mainly small buildings, with a few larger building changes, and some of the change areas are surrounded by complex environments, as in Fig. 7(a). The complex environment leads to difficulties in detection. From Fig. 7, it is obvious that CBSASNet has fewer misdetections and omissions than the other methods and is able to determine the edges of the change areas more accurately. Visually, CBSASNet has a better performance.

## V. Ablation Experiments

In order to be able to adequately assess the impact of the proposed methodology as well as the improved modules on the model performance, we designed a significant number of experiments to validate our proposed methodology.

### *A. Overall Effectiveness of CTFM and CBSA in CBSASNet*

In order to verify the validity and overall contribution of the proposed CTFM with CBSA to CBSASNet, we provide three variants to compare with the proposed CBSASNet. This includes a) baseline: the encoder and decoder use the base bottleneck to extract picture features, and the encoder fuses

TABLE IV

COMPARATIVE RESULTS OF PRECISION, RECALL, $F1$-SCORE, IOU, OA, AND KAPPA FOR ALL METHODS ON THE LEVIR-CD DATASET, WITH THE OPTIMAL RESULTS SHOWN IN BOLD FONT

| **Methods** | **Years** | **IoU(%)** | **Pre(%)** | **Rec(%)** | **F1(%)** | **OA(%)** | **Kappa(%)** |
|---|---|---|---|---|---|---|---|
| FC-Siam-conc [25] | 2018 | 80.17 | 91.14 | 86.94 | 88.99 | 98.90 | 88.42 |
| FC-Siam-diff [25] | 2018 | 79.92 | 91.68 | 86.18 | 88.84 | 98.90 | 88.26 |
| FC-EF [25] | 2018 | 76.77 | 87.81 | 85.93 | 86.86 | 98.68 | 86.16 |
| BIT [34] | 2020 | 80.58 | 91.19 | 87.38 | 89.24 | 98.93 | 88.68 |
| IFN [53] | 2020 | 82.05 | **95.61** | 85.26 | 90.14 | 99.05 | 89.64 |
| SNUNet-CD [24] | 2021 | 82.30 | 91.90 | 88.74 | 90.29 | 99.03 | 89.78 |
| ICIF-Net [49] | 2022 | 82.12 | 91.96 | 88.47 | 90.18 | 99.02 | 89.67 |
| PA-Former [54] | 2022 | 80.56 | 90.48 | 88.02 | 89.23 | 98.92 | 88.66 |
| RDP-Net [55] | 2022 | 77.92 | 89.21 | 86.03 | 87.59 | 98.76 | 86.94 |
| DMINet [41] | 2023 | 81.63 | 91.01 | 88.79 | 89.88 | 98.98 | 89.35 |
| GeSANet [46] | 2023 | 82.00 | 91.65 | 88.62 | 90.11 | 99.01 | 89.59 |
| AERNet [56] | 2023 | 82.97 | 91.94 | 89.48 | 90.69 | 99.06 | 90.20 |
| HANet [57] | 2023 | 82.43 | 91.42 | 89.35 | 90.37 | 99.03 | 89.86 |
| WNet [36] | 2023 | 82.40 | 91.31 | 89.41 | 90.35 | 99.03 | 89.84 |
| ScratchFormer [58] | 2024 | 80.08 | 91.46 | 86.55 | 88.94 | 98.90 | 88.36 |
| SEIFNet [28] | 2024 | 82.70 | 90.58 | 90.48 | 90.53 | 99.04 | 90.02 |
| CBSASNet(ours) | 2024 | **84.14** | 92.47 | 90.33 | **91.39** | **99.13** | **90.93** |

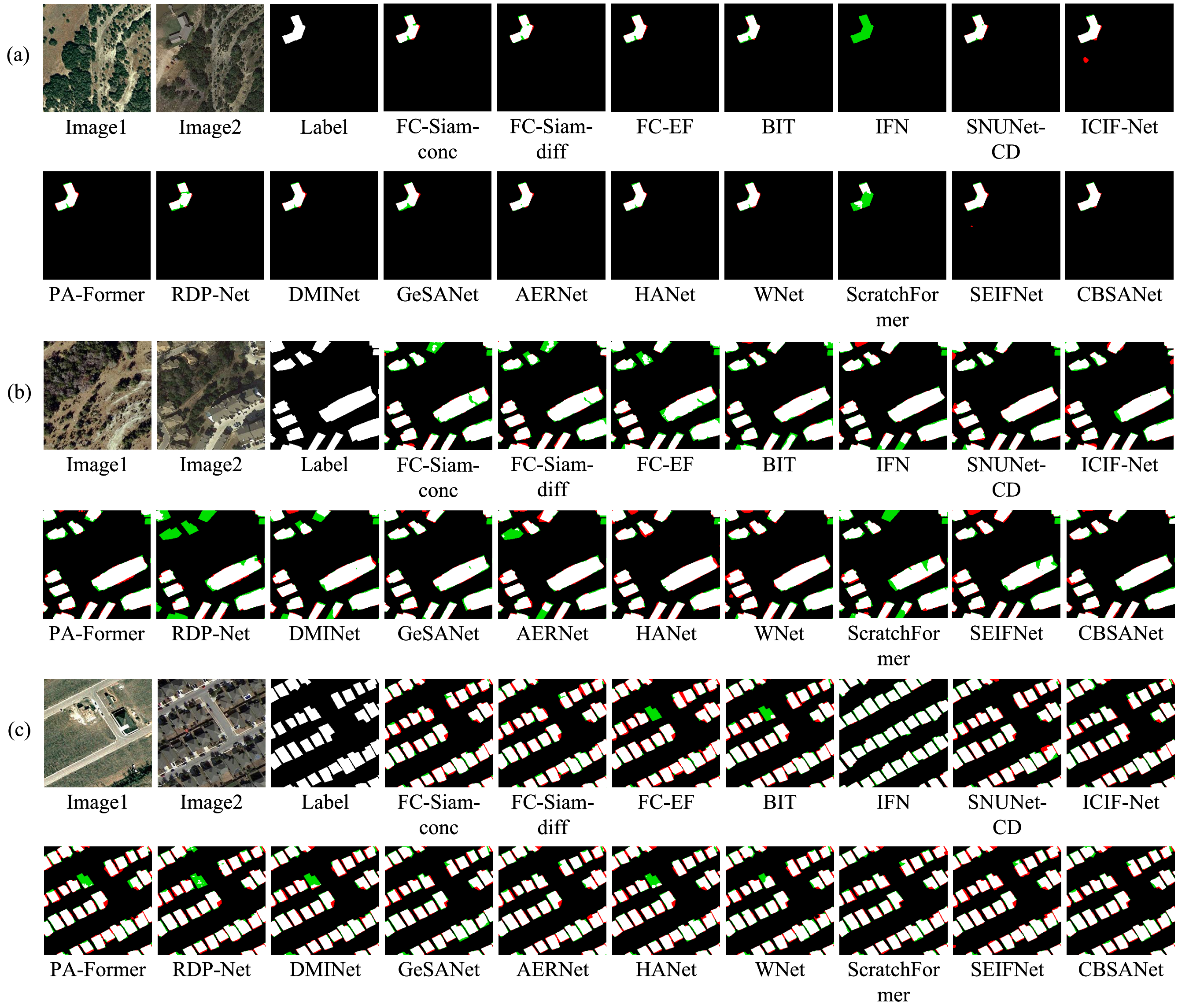


Fig. 7. Visualization results of the 16 CD methods on the LEVIR-CD dataset. Image rendering colors: white in true positive, red in false positive, black in true negative, and green in false negative. (a)–(c) Represent three randomly selected sample images from that dataset.

TABLE V

ABLATION STUDY OF THREE VARIANTS WITH CBSASNET ON CDD DATASET

| **Methods** | **IoU(%)** | **Pre(%)** | **Rec(%)** | **F1(%)** | **OA(%)** | **Kappa(%)** |
|---|---|---|---|---|---|---|
| Baseline | 92.62 | 96.66 | 95.69 | 96.17 | 99.10 | 95.66 |
| Baseline+CBSA | 94.42 | 97.17 | 97.09 | 97.13 | 99.32 | 96.74 |
| Baseline+CTFM | 93.23 | 96.31 | 96.68 | 96.50 | 99.17 | 96.03 |
| CBSASNet(ours) | 94.58 | 96.99 | 97.44 | 97.21 | 99.34 | 96.84 |

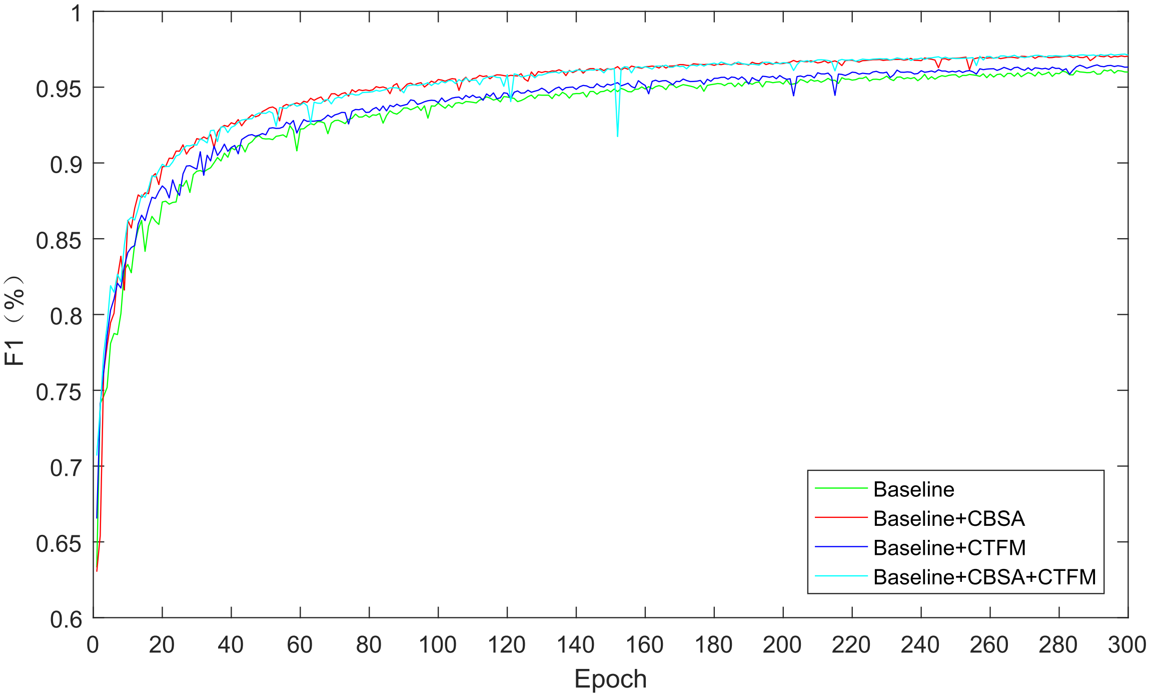


Fig. 8. Line graph of ablation experiments on the CDD dataset.

the dual-branching features using a cascade; b) baseline + CBSA: CBSA is used instead of bottleneck to extract picture features; c) baseline + CTFM: CTFM is used to fuse the dual-branching features; and d) baseline + CTFM + CTFM: The base network plus the CBSA and the CTFM, i.e., our complete network. Table V shows the results of the experiments, and in order to be able to visualize the training process of the models more, a line graph of the $F1$ evaluation metrics of the validation set during the training process of the four models is shown in Fig. 8. From the results in Table V, it can be seen that both CBSA and CTFM contribute to the improvement of the performance of the underlying network, and when both CBSA and CTFM are applied simultaneously, the performance of the network reaches its optimum. From Fig. 8, we can see that the initial performance of CBSASNet is higher compared to the other variants, and the parameters can be better optimized in the subsequent training. At the epoch, greater than 200, the performance of CBSASNet is relatively more stable.

### B. CTFM and the Impact of the Accession Stage

As the extracted features go from shallow to deep, the information represented by the feature map becomes more and more abstract. In order to verify the generalization ability of CTFM to fuse shallow network features and deep network features, we have done a study on the fusion stage of CTFM, and the results of the experiments are shown in Table VI. CTFM12 in the table indicates that CTFM is used instead of the cascade approach to fuse the dual branching features in the first and second stages. From Table VI, it can be seen that the more stages of using CTFM to fuse the dual-branching features, the less semantic deviation occurs after the fusion of the dual-branching features, and the more accurate location information is provided to the decoder. In the experimental stage, in order to improve the performance of the model while being able to improve the model's advantage in training and inference speed, reduce the model's dependence on high-performance hardware, and improve the model's deployment effectiveness and efficiency, we integrated Param and Flops to select CTFM12 as the object of our experiment.

### C. Effectiveness of CBSA

The use of the split-attention module is proposed in ResNeSt [21] to enable cross-channel attention. On this basis, we propose CBSA, the yet difference between the two feature extraction methods is the addition of a 1 × 1 convolutional branch, which can highlight the information of the main region more. In order to verify the effectiveness of the CBSA module with the addition of a 1 × 1 convolutional branch, we conducted experiments comparing it with the original

TABLE VI
ABLATION EXPERIMENTS ON THE NUMBER OF CTFMS ON THE CDD DATASET

| Methods | Flops(G) | Para(M) | IoU(%) | Pre(%) | Rec(%) | F1(%) | OA(%) | Kappa(%) |
|---|---|---|---|---|---|---|---|---|
| w/o CTFM | 17.61 | 5.22 | 94.42 | 97.17 | 97.09 | 97.13 | 99.32 | 96.74 |
| CTFM1 | 26.63 | 5.33 | 94.51 | 96.89 | 97.46 | 97.18 | 99.33 | 96.80 |
| CTFM12 | 31.64 | 5.76 | 94.58 | 96.99 | 97.44 | 97.21 | 99.34 | 96.84 |
| CTFM123 | 38.63 | 7.46 | 94.71 | 97.27 | 97.29 | 97.28 | 99.36 | 96.92 |
| CTFM1234 | 45.62 | 14.28 | 94.88 | 97.22 | 97.52 | 97.37 | 99.38 | 97.02 |
| CTFM12345 | 47.36 | 21.10 | 94.96 | 97.29 | 97.55 | 97.42 | 99.39 | 97.07 |

TABLE VII
ABLATION STUDY OF DIFFERENT SPLIT BLOCKS ON THE WHU-CD DATASET

| Methods | IoU(%) | Pre(%) | Rec(%) | F1(%) | OA(%) | Kappa(%) |
|---|---|---|---|---|---|---|
| Splat-Attention | 84.01 | 92.49 | 90.16 | 91.31 | 99.20 | 90.89 |
| CBSA | 84.84 | 92.65 | 90.96 | 91.80 | 99.24 | 91.40 |

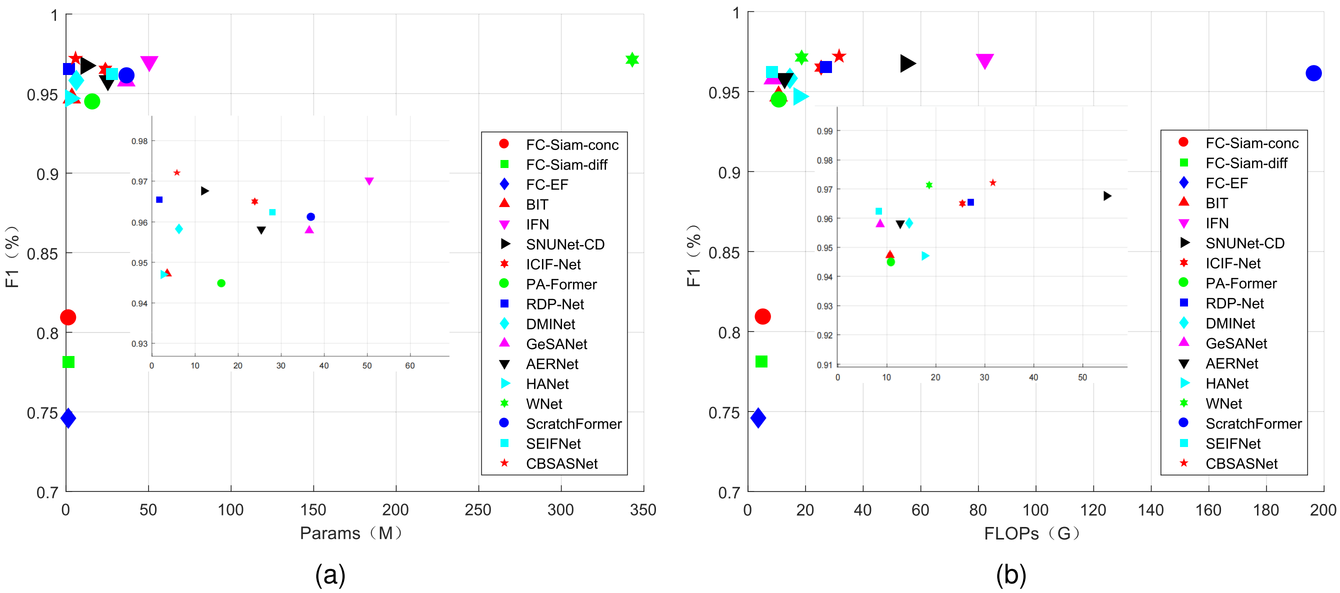


Fig. 9. Comparison of the number and complexity of model parameters of different methods. (a) Comparison of the number of model parameters and performance ($F1$) of different methods. (b) Comparison of the computational complexity and performance ($F1$) of models of different methods.

split-attention module, and the experimental results are shown in Table VII. The experimental results illustrate that the CBSA module is more effective in extracting image features.

### *D. Model Size and Computational Complexity*

We tested the number of parameters and the computational effort of all the methods, which were used to compare the model size as well as the computational complexity of the different methods. The obtained results are displayed in Table IX. The obtained results are displayed in Table VIII. Although CBSASNet does not reach SOTA in terms of the number of parameters and floating-point operations, CBSASNet outperforms GeSANet on the WHU-CD dataset using about 16% of the number of parameters, while WNet, which has a similar performance to CBSASNet on the CDD dataset, uses close to 60 times the number of parameters of CBSASNet. On the LEVIR-CD dataset, CBSASNet uses about one-fifth the number of parameters to outperform AERNet. We show the results of our experiments more graphically in Fig. 9 Compared to other models with a similar number of parameters, CBSASNet has a better performance than them. Thus, this scheme of improving model performance by increasing the computational cost of a specific part has utility and practical value.

### *E. Coefficients of the Loss Function*

In order to assess the impact of the loss function, we investigate by varying the weight coefficients of the binary cross-entropy loss function. The weights of each category are taken in the range of $[0, 1]$, starting from 0.1 with an interval of 0.2. The results are shown in Fig. 9. From the value of $F1$, the fluctuation range of the performance index of CBSASNet under different weights is below 3%. It can be observed through the Pre and Rec metrics that the networks learned under the loss function with different weights have different possibilities of wrong detection when detecting changing regions versus unchanging regions. Specifically, when increasing the weight of a few samples, Pre is smaller than Rec, which means that the probability of misdetecting a changing pixel decreases, while the probability of misdetecting an unchanged pixel increases. As the minority sample

TABLE VIII

QUANTITATIVELY COMPARE THE PERFORMANCE, PARAM, AND FLOPS OF DIFFERENT METHODS. PERFORMANCE IS REFERENCED BY $F1$-SCORE

| **Methods** | **Flops(G)** | **Para(M)** | **WHU-CD(%)** | **CDD(%)** | **LEVIR-CD(%)** |
|---|---|---|---|---|---|
| FC-Siam-conc [25] | 5.33 | 1.55 | 77.31 | 80.91 | 88.99 |
| FC-Siam-diff [25] | 4.73 | 1.35 | 80.37 | 78.12 | 88.84 |
| FC-EF [25] | 3.58 | 1.35 | 80.70 | 74.60 | 86.86 |
| BIT [34] | 10.63 | 3.5 | 83.80 | 94.72 | 89.24 |
| IFN [53] | 82.26 | 50.44 | 89.13 | 97.02 | 90.14 |
| SNUNet-CD [24] | 54.83 | 12.03 | 86.98 | 96.76 | 90.29 |
| ICIF-Net [49] | 25.41 | 23.84 | 90.31 | 96.50 | 90.18 |
| PA-Former [54] | 10.86 | 16.13 | 85.09 | 94.48 | 89.23 |
| RDP-Net [55] | 27.12 | 1.69 | 87.35 | 96.55 | 87.59 |
| DMINet [41] | 14.55 | 6.24 | 85.48 | 95.83 | 89.88 |
| GeSANet [46] | 8.62 | 36.5 | 91.66 | 95.79 | 90.11 |
| AERNet [56] | 12.72 | 25.36 | 91.56 | 95.82 | 90.69 |
| HANet [57] | 17.67 | 2.61 | 88.52 | 94.70 | 90.37 |
| WNet [36] | 18.63 | 342.99 | 87.25 | 97.13 | 90.35 |
| ScratchFormer [58] | 196.59 | 36.92 | 87.78 | 96.12 | 88.94 |
| SEIFNet [28] | 8.37 | 27.9 | 91.21 | 96.24 | 90.53 |
| CBSASNet(ours) | 31.64 | 5.76 | 92.52 | 97.21 | 91.39 |

TABLE IX

COEFFICIENTS OF THE LOSS FUNCTION

| weight | WHU-CD | | | CDD | | | LEVIR-CD | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pre | Rec | F1 | Pre | Rec | F1 | Pre | Rec | F1 |
| [0.1, 0.9] | 87.09 | 93.05 | 89.97 | 91.87 | 99.25 | 95.42 | 83.97 | 94.72 | 89.02 |
| [0.3, 0.7] | 92.82 | 91.41 | 92.11 | 95.69 | 98.39 | 97.02 | 90.56 | 91.54 | 91.05 |
| [0.5, 0.5] | 93.93 | 91.15 | 92.52 | 96.99 | 97.44 | 97.21 | 92.47 | 90.33 | 91.39 |
| [0.7, 0.3] | 93.74 | 90.17 | 91.92 | 98.10 | 95.57 | 96.82 | 95.06 | 86.63 | 90.65 |
| [0.9, 0.1] | 95.83 | 86.43 | 90.89 | 99.05 | 91.86 | 95.32 | 96.92 | 82.03 | 88.86 |

weight decreases gradually, the misjudgment probability will be reversed. Moreover, the degree of fluctuation varies across datasets, with relatively more fluctuation in the LEVIR-CD and WHU-CD datasets, which can be attributed here to the difference in the degree of imbalance across datasets. All in all, the loss function weights affect the learning of the network on a few samples, but their overall performance fluctuates little, which also indicates the stability of our network. In addition, it can also be seen from the table that the network has optimal results on the three datasets when the loss function weights are [0.5, 0.5], and for this reason, we recommend setting the loss function weights to [0.5, 0.5].

## VI. CONCLUSION

In this article, we introduce CBSANet, a Siamese network based on CBSA, for remote sensing image CD in detail. We develop two key modules, the CBSA module and the CTFM, to improve the performance of CD. Specifically, CBSA integrates the image channel information and extracts finer dual-time-phase image features by changing the shortcut connections. On the other hand, CTFM highlights the different information of the dual-temporal phase features more smoothly by enhancing the unilateral features while integrating the dual-temporal phase image features. We have done a lot of experiments on three public datasets, WHU-CD, CDD, and LEVIR-CD, and through experimental validation, it is shown by experimental comparison with other deep learning-based remote sensing CD methods that CBSASNet has good performance in locating spatial information, highlighting change regions and improving network stability.

Our method still has room for improvement. While network performance is improved, the computational cost of CBSASNet also increases. It is observed that, as in the deep layers of the network, the increase of computational costs (parameter scalability) far exceeds its performance improvement effects due to the large feature dimensions. Therefore, the lightweight would be our future research direction based on the existing framework.

## REFERENCES

[1] B. Demir, F. Bovolo, and L. Bruzzone, “Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 300–312, Jan. 2013.

[2] X. Huang, L. Zhang, and T. Zhu, “Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, pp. 105–115, 2013.

[3] C. Tarantino, M. Adamo, R. Lucas, and P. Blonda, “Change detection in (semi-) natural grassland ecosystems for biodiversity monitoring using open data,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 8981–8984.

[4] C.-F. Chen et al., “Multi-decadal mangrove forest change detection and prediction in honduras, central America, with Landsat imagery and a Markov chain model,” *Remote Sens.*, vol. 5, no. 12, pp. 6408–6426, Nov. 2013.

[5] K. Patel, M. Jain, M. I. Patel, and R. Gajjar, “A novel approach for change detection analysis of land cover from multispectral FCC optical image using machine learning,” in *Proc. 2nd Int. Conf. Range Technol. (ICORT)*, Aug. 2021, pp. 1–6.

[6] Y. Yin, “Research on natural disaster target change detection method based on deep learning,” in *Proc. IEEE 3rd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Jan. 2023, pp. 946–950.

[7] T. Celik, “Unsupervised change detection in satellite images using principal component analysis and $k$-means clustering,” *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[8] M. H. Kesikoğlu, Ü. H. Atasever, and C. Özkan, "Unsupervised change detection in satellite images using fuzzy c-means clustering and principal component analysis," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XL-7/W2, pp. 129–132, Oct. 2013.

[9] Z. Li, W. Shi, H. Zhang, and M. Hao, "Change detection based on Gabor wavelet features for very high resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 783–787, May 2017.

[10] S. Jin and S. A. Sader, "Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances," *Remote Sens. Environ.*, vol. 94, no. 3, pp. 364–372, Feb. 2005.

[11] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, 1998.

[12] E. F. Lambin and A. H. Strahlers, "Change-vector analysis in multitemporal space: A tool to detect and categorize land-cover change processes using high temporal-resolution satellite data," *Remote Sens. Environ.*, vol. 48, no. 2, pp. 231–244, May 1994.

[13] P. Du, X. Wang, D. Chen, S. Liu, C. Lin, and Y. Meng, "An improved change detection approach using tri-temporal logic-verified change vector analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 278–293, Mar. 2020.

[14] K. Wessels et al., "Rapid land cover map updates using change detection and robust random forest classifiers," *Remote Sens.*, vol. 8, no. 11, p. 888, Oct. 2016.

[15] H. Nemmour and Y. Chibani, "Multiple support vector machines for land cover change detection: An application for mapping urban extensions," *ISPRS J. Photogramm. Remote Sens.*, vol. 61, no. 2, pp. 125–133, Nov. 2006.

[16] J. Im and J. Jensen, "A change detection model based on neighborhood correlation image analysis and decision tree classification," *Remote Sens. Environ.*, vol. 99, no. 3, pp. 326–340, Nov. 2005.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit.*, Sep. 2017, pp. 4700–4708.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[22] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Oct. 2017.

[23] 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.

[24] 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, pp. 1–5, 2021.

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

[26] F. Heidary, M. Yazdi, M. Dehghani, and P. Setoodeh, "Urban change detection by fully convolutional Siamese concatenate network with attention," 2021, *arXiv:2102.00501*.

[27] Z. Du, X. Li, J. Miao, Y. Huang, H. Shen, and L. Zhang, "Concatenated deep learning framework for multi-task change detection of optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 719–731, 2023.

[28] Y. Huang, X. Li, Z. Du, and H. Shen, "Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5609414.

[29] F. Heidary, M. Yazdi, P. Setoodeh, and M. Dehghani, "CTS-UNet: Urban change detection by convolutional Siamese concatenate network with Swin transformer," *Adv. Space Res.*, vol. 72, no. 10, pp. 4272–4281, 2023.

[30] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[31] T. Yan, Z. Wan, and P. Zhang, "Fully transformer network for change detection of remote sensing images," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1691–1708.

[32] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[33] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[34] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 2503305.

[35] Z. Mao, X. Tong, Z. Luo, and H. Zhang, "MFATNet: Multi-scale feature aggregation via transformer for remote sensing image change detection," *Remote Sens.*, vol. 14, no. 21, p. 5379, Oct. 2022.

[36] X. Tang, T. Zhang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "WNet: W-shaped hierarchical network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615814.

[37] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Kuala Lumpur, Malaysia, 2022, pp. 207–210, doi: 10.1109/IGARSS46834.2022.9883686.

[38] 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, pp. 1–18, 2022.

[39] H. Zhong and C. Wu, "T-UNet: Triplet UNet for change detection in high-resolution remote sensing images," 2023, *arXiv:2308.02356*.

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

[41] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401015.

[42] 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, p. 1662, 2020.

[43] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[44] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2020, pp. 2736–2746.

[45] Q. Xu, Z. Ma, H. Na, and W. Duan, "DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation," *Comput. Biol. Med.*, vol. 154, Mar. 2023, Art. no. 106626.

[46] X. Zhao, K. Zhao, S. Li, and X. Wang, "GeSANet: Geospatial-awareness network for VHR remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402814.

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

[48] 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. Photogramm. Remote Sens.*, vol. 183, pp. 228–239, Jan. 2022.

[49] 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.

[50] J. Gu et al., "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 12094–12103.

[51] 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.

[52] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 565–571, May 2018.

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

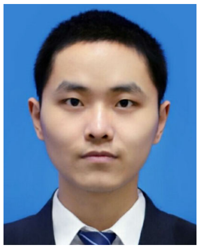
[54] M. Liu, Q. Shi, Z. Chai, and J. Li, "PA-Former: Learning prior-aware transformer for remote sensing building change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[55] H. Chen, F. Pu, R. Yang, R. Tang, and X. Xu, "RDP-Net: Region detail preserving network for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5635010.

[56] J. Zhang et al., "AERNet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617116.

[57] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3867–3878, 2023.

[58] M. Noman et al., "Remote sensing change detection with transformers trained from scratch," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4704214.



**Naiwei He** received the B.E. degree from the Shandong University of Science and Technology, Qingdao, China, in 2019. He is currently pursuing the master's degree with the School of Computer Science and Technology, Xinjiang University, Ürümqi, China.

His research interests include computer vision and remote sensing image change detection.

**Liejun Wang** received the Ph.D. degree from the School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an, China, in 2012.

He is now a Professor with the School of Computer Science and Technology, Xinjiang University, Ürümqi, China. His research include wireless sensor networks, computer vision, and natural language processing.

**Panpan Zheng** received the Ph.D. degree from the University of Arkansas, Fayetteville, AR, USA, in 2020.

He worked as an Applied Scientist with AWS AI Laboratory, Silicon Valley, CA, USA. He is currently working as an Associate Professor with the Department of Computer Science and Technology, Xinjiang University, Ürümqi, China. His research interests include online learning and sequential decision-making, anomaly detection, image change detection, and salient object detection.

**Cui Zhang** received the master's degree from the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China, in 2011. She is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Xinjiang University, Ürümqi, China.

Her research interests include computer vision and remote sensing image change detection.

**Lele Li** received the master's degree in information and communication engineering from the School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree in computer science and technology with the School of Computer Science and Technology, Xinjiang University, Ürümqi, China.

His research interests include deep learning, computer vision, and remote sensing image change detection.