Hindawi
Mobile Information Systems
Volume 2023, Article ID 4200153, 15 pages
https://doi.org/10.1155/2023/4200153

Research Article

Remote Sensing Image Change Detection Network Based on Twin High-Resolution Representation

Xiaosuo Wu,[1,2,3] Yaya Ma,[1] Haowen Yan,[1,2] Ze Qiao,[1] Chaoyang Wu,[1] Cunge Guo,[4] Shuang Yao,[1] and Yufeng Fan[1]

[1]*School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China*
[2]*Academician Expert Workstation of Gansu Dayu Jiuzhou SpaceInformation Technology Co. Ltd., Lanzhou 730070, China*
[3]*Institute of Sensor Technology, Gansu Academy of Science, Lanzhou 730070, China*
[4]*Lanzhou Institute of Technology, Lanzhou 730070, China*

Correspondence should be addressed to Chaoyang Wu; wuxs_laser@lzjtu.edu.cn

Received 25 October 2022; Revised 8 December 2022; Accepted 20 January 2023; Published 22 February 2023

Academic Editor: A. H. Alamoodi

With the increase of spatial resolution of remote sensing images, features of feature imaging become more and more complex, and the change detection methods based on techniques such as texture representation and local semantics are difficult to meet the demand. Most change detection methods usually focus on extracting semantic features and ignore the importance of high-resolution shallow information and fine-grained features, which often lead to uncertainty in edge detection and small target detection. For single-input networks when two temporal images are connected, the shallow layer of the network cannot provide the information of the individual original image to the deep layer features to help reconstruct the image, and therefore, the change detection results may be missing in detail and feature compactness. For this purpose, a twins context aggregation network (TCANet) is proposed to perform change detection on remote sensing images. In order to reduce the loss of spatial accuracy of remote sensing images and maintain high-resolution representation, we introduce HRNet as our backbone network to initially extract the features of interest. Our proposed context aggregation module (CAM) can amplify the convolutional neural network receptive field to obtain more detailed contextual information without significantly increasing the computational effort. The side output embedding module (SOEM) is proposed to improve the accuracy of small volume target change detection as well as to shorten the training process and speed up the detection while ensuring the performance. The method has experimented on the publicly available CDD dataset, the SYSU-CD dataset, and a challenging DSIFN dataset. With significant improvements in precision, recall, $F1$ score, and overall accuracy, the method outperforms the five methods mentioned in the literature.

1. Introduction

Change detection of acquired remote sensing images of the same geographical area, at different times, is an important part of many practical applications such as land use, vegetation change detection, ecosystem detection, and damage assessment [1]. The traditionally time-consuming and laborious way of analyzing changes in remotely sensed images based on manual work makes the automation of this process an important and practically needed research area. Automatic implementation of time-series image change detection is of great scientific research and application value, and research in this field has been carried out in the remote sensing community for decades [2–6]. In recent years, artificial intelligence algorithm techniques represented by deep learning have developed rapidly and have been applied in various fields, such as computer vision [7], speech recognition [8], and information retrieval [9], especially in the field of computer vision. J. Bruna et al. proposed a fully convolutional neural network [10] to achieve end-to-end pixel classification of images using convolutional layers instead of fully connected layers. Hughes et al. proposed a pseudo-twin convolutional neural network-based method applied to the detection of changes between SAR images and

optical images [11]. As the research progressed, in 2017, Ashish Vaswani, Niki Parmar, and several other researchers joined together to publish the landmark paper that ushered in the era of large models [12]. In this paper, they proposed the famous transformer architecture. In 2018, a model called BERT blew up the NLP community, setting a new SOTA record of 11 NLP tasks, and it was transformer that was responsible for it. Transformer has been extensively used in the field of remote sensing, especially in the field of semantic segmentation. Gori et al. [13] used RNN to compress node information and learn graph node labels and first proposed the concept of graph neural network (GNN). Graph neural networks (GNNs) are a deep learning-based method for operating in the graph domain. Later, graph convolutional network (GCN) was proposed in the literature [14], which formally used CNNs for modeling graph-structured data. Although both transformer and graph neural networks are popular deep learning methods in recent years, relatively little research has been conducted in dual-time remote sensing image change detection. Therefore, twin convolutional neural networks are still used in this paper.

At present, scholars at home and abroad have proposed various remote sensing image change detection methods based on deep learning. In terms of the framework of the methods, they can be broadly divided into 3 categories. The first type is the method of extracting features first and then detecting them, i.e., feature extraction of the image using a deep network followed by change detection based on the features [15]. The second category is the method of preclassification followed by detection. That is, the images are primarily preclassified using traditional algorithms, and then, the deep network is trained with explicitly changed and unchanged samples. Finally, uncertain samples are fed into the trained network to obtain the result graph [16]. Although both types of methods are based on deep learning and their results also are better than traditional methods, they are still subject to human experience and prone to errors in the steps required for threshold judgment, clustering, and sample selection in the detection process. The third category is the methods based on fully convolutional networks. It is a completely end-to-end learning framework with no human interference in between, and the whole process is more robust and efficient [6]. Depending on the input method of the image, this type of method can be subdivided into networks with a single input and networks with dual input.

Although the full convolutional network-based approach achieves better change detection performance, there are still some shortcomings. The cascading pooling operation of encoders from high resolution to low resolution will lead to a decrease in spatial resolution, and it is difficult for the decoder to recover its resolution. Direct use of the full twin convolutional change detection network suffers from low detection completeness, easy false detection, and missed detection. This is primarily limited by the lack of network feature extraction capability and the ineffective use of contextual semantic information in the spatial and channel domains. Full convolutional networks do not easily obtain the edge information of images when extracting the deep features of dual-temporal images. To this end, this paper proposes a twin context aggregation network (TCANet) to address the above problem. First, to obtain a high-resolution representation, we use the HRNet network for our backbone network. Second, in order to enhance the network feature extraction capability as well as to effectively utilize the channel domain contextual semantic information, we propose the context aggregation module (CAM). Finally, in the decoding part, we introduce the side output embedding module (SOEM) in order to obtain the edge information and small target information of the image, while suppressing useless information and further improving the accuracy of change detection. Remote sensing image change detection, as a more cutting-edge research direction, is still one of the relatively less studied areas at present, and the best experimental results can be obtained using our proposed three modules. The main contributions are as follows.

(1) We introduce dual HRNet as the backbone of our twin network. This network can maintain high resolution from beginning to end, and information interaction of different branches can supplement information loss caused by the reduced number of channels.

(2) In the coding part. The context aggregation module (CAM) is proposed in order to improve the extraction of network features and to efficiently utilize the channel domain contextual semantic information. The module turns the output feature maps into four 1/4-channel feature maps and then uses dilation convolution with different dilation rates to integrate the multichannel contextual information in parallel.

(3) In the decoding part. We introduce the side output embedding module (SOEM) in order to obtain the edge change information of the dual-time phase images as well as more detailed fine image details and complex texture features of the high-resolution remote sensing images.

(4) Our network achieves impressive results on all three datasets. More specifically, we have obtained 89.87% $F1$ score on the challenging DSIFN test set.

The rest of the paper is organized as follows: Section 2 describes the work related to change detection. Section 3 discusses the proposal of the method. Section 4 presents the experimental data set and evaluation metrics. Section 5 presents the experimental design and results. Finally, Section 6 discusses our work and conclusions.

2. Related Work

In recent years, many neural network techniques and components for scene segmentation have been applied for change detection tasks to extract deeper representations. First U-Net [17] pioneered the benchmark model and then used the Siamese network [18–24] to become the standard method for change detection. To improve the performance of change detection, a lot of work has been conducted on depth feature extraction and refinement.

2.1. Siamese Neural Network. The Siamese neural network is a coupled architecture based on two artificial neural networks. In simple terms, a Siamese neural network composes of two neural networks with the same structure and shared weights spliced together. The "concatenation" of neural networks is achieved by sharing the weight.

Change detection methods based on fully convolutional networks are roughly divided into two categories. One type is the early fusion approach (single-input); the other is the twin network approach (dual-input). A single-input network is a cascade of dual-temporal images into one image before being fed into the network [25, 26]. For example, in the literature [26], dual-temporal image pairs are concatenated as input to an improved UNet++ network, and change maps at different semantic levels are merged to generate the final change map. In contrast to single-input networks, dual-input networks are borrowed from twin networks [18, 22, 27], where the front-end feature extraction part of the fully convolutional network is replaced by two networks with the same structure. For instance, the literature [18] proposed three fully convolutional neural network frameworks for change detection in remote sensing images, one of which is single input and the other two are dual input. The results of many change detection experiments demonstrate that a dual-input network architecture is more suitable for change detection. Many scholars have done research on Siamese network based on change detection methods. Yu et al. proposed the NestNet [28] network, a model that introduces two parallel modules to extract the respective features of diachronic images and then uses absolutely different operations to process the features of the two images. The literature [22] proposes an IFN-based method, which is a fully convolutional network-based method belonging to a dual input. In this method, dual-temporal images are extracted with depth features by a twin network, and the down-sampled change maps are directly fed into the middle layer of the network during training. And finally, the network parameters are updated by calculating the losses independently. Fang et al. proposed a SNUNet-CD network [29]. This network is a modification of UNet++, which differs from UNet++ in that it uses two twin convolutional filters to extract features from two images, and aggregates and refines features from multiple semantic layers by integrating the channel attention module. Finally, the test results were obtained.

2.2. Contextual Information Aggregation. Since each pixel point in an image cannot be isolated, one pixel must be related to the surrounding pixels in some way. The interconnection of a large number of pixels is what produces the various objects in an image, so the contextual features of an image are of great importance. Not having sufficient access to rich contextual information during the change detection task can have an impact on our detection results. Many approaches add modules on top of the encoded network to expand their effective receptive field and integrate more contextual information. In the paper [30], global pooling operations are introduced to learn the scenario-level global context, and the importance of the receptive field is discussed. PSPNet [31] extends the application of global pooling to image subregions and proposes a parallel spatial pooling design that aggregates multiscale contextual information. Dilated convolution is another design that can amplify the receptive field of CNNs without significantly increasing the computational effort [32, 33]. Combining the atrous convolution and multilevel pooling design in PSPNet, the atrous space pyramid pooling (ASPP) module was proposed in [34] and improved in [35–37]. The attention mechanism [38–40] uses the sigmoid function to generate "attention" descriptors after global pooling operations, which are another contextual aggregation design. In order to better serve the change detection of remote sensing images, this paper uses multiple parallel inflated convolutions to obtain global and local contextual information.

3. The Proposed Methodology

In this section, we describe in detail the proposed twin context aggregation network (TCANet) for remote sensing image change detection. First, the general structure we proposed will be outlined. After this, we give an illustration of the HRNet (our baseline network) architecture. Finally, the design of each module is presented, including the CAM module and the SOEM module.

3.1. Overview of the Network. As shown in Figure 1, a twin context aggregation network (TCANet) is designed in this paper. The model feeds dual-temporal images into two networks with shared parameters to extract features separately. First, dual-temporal images are input into the backbone network HRNet to obtain four change feature maps with different sizes and a different number of channels. The structure preserves spatial detail information but does not take full advantage of contextual information. Therefore, we integrate more contextual information to improve network performance by introducing a scale-dependent contextual aggregation module in four different branches. Since the four parallel outputs generated by HRNet are information-dispersed, we embed different levels of local contextual information from the context aggregation module (CAM) into these features to make the outputs informative. Then, the two feature values obtained from the feature encoding stage are differenced, and the absolute values are taken to obtain dual-temporal feature fusion information at different scales. Finally, we input the fused feature maps to the side output embedding module to facilitate the detection of edges and small targets. The detailed process of the network is as follows.

As shown in Figure 1, B1, B2, B3, and B4 are the four parallel branches generated by HRNet. CAM module processing is performed first, then the output of the CAM module is up-sampled (the output of the first CAM module is not up-sampled), and then the associated feature maps are concatenated.

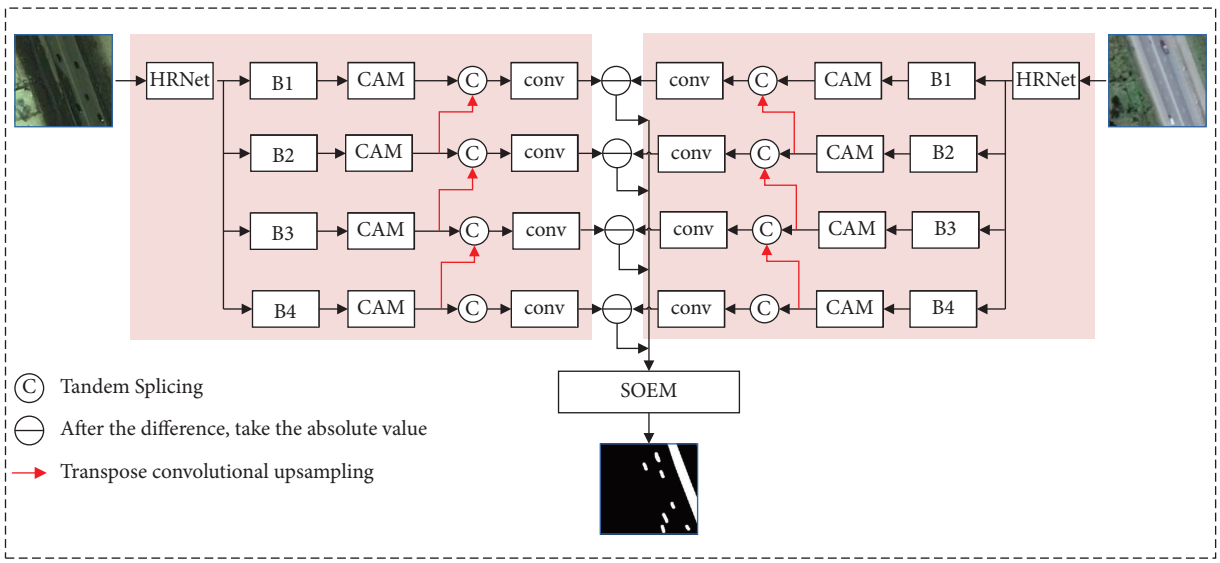

Figure 1: Overall architecture.

$$U_i = \{\mathrm{CAM}(B_i), \mathrm{up}[\mathrm{CAM}(B_{i+1})]\}\,(\mathrm{i} = 1, 2, 3),$$
$$U_4 = \{\mathrm{CAM}(B_4)\}, \tag{1}$$

where CAM(•) denotes the output feature map after CAM processing, up[•] denotes the up-sampling operation, and {•} represents the tandem stitching operation. Immediately afterwards, U_1, U_2, U_3, and U_4 are channel compressed using 1×1 convolution so that the four output feature maps have the same number of channels. Then, a difference absolute value operation is performed with the output of the other encoder to fuse the features of the two images. That is,

$$F_i = \mathrm{sub}[\mathrm{conv}(U_i), \mathrm{conv}(U_{\mathrm{i}}^*)]\,(\mathrm{i} = 1, 2, 3, 4), \tag{2}$$

where conv(•) represents the 1×1 convolution operation to achieve channel compression. U_i is the feature map extracted from the image at the moment of $T1$ by the encoding structure. U_i^* is the feature map extracted from the image at the moment of $T2$ by the encoding structure, and sub[•] represents the feature fusion of the two images after the difference in absolute value processing. Finally, the obtained F_1, F_2, F_3, and F_4 are input to the side output embedding module (SOEM) to obtain the final predicted map.

3.2. Baseline: HRNet. Most existing encoder methods perform cascading pooling operations (down-sampling) from high to low resolution to obtain. But the cascade pooling operation leads to a loss of spatial accuracy that is difficult to recover by the decoder. To overcome this limitation, HRNet [41, 42] introduced a multiscale parallel design. It maintains high-resolution output throughout and fuses multiscale information so that the network can initially extract as many image features as possible.

As shown in Figure 2, HRNet consists of parallel subnetworks from high resolution to low resolution, with repeated information exchange across multiresolution subnetworks (multiscale fusion). Specifically, the four parallel subnetworks are B1, B2, B3, and B4, where B1 always maintains a high-resolution representation. At the same time, the feature map after each convolution block is convolved 3×3 with a step of 2 (down-sampling) to reduce the space size of the feature map and up-sampling the feature maps after each convolution block to connect to different branches for multiscale fusion (the first convolution block of B1 does not need to be up-sampled). Finally, the network generates four sets of feature maps with different resolutions. They are first up-sampled to recover to the same size as branch B1 and then fused, and the fused feature maps can be used to generate segmentation results, which, of course, we do not need to generate here. The four branches of HRNet are equivalent to 1/4, 1/8, 1/16, and 1/32 of the original input size.

3.3. Contextual Aggregation Module (CAM). Contextual information is essential to determine the two categories of objects changed and unchanged, and many approaches add modules to the top of the encoding network to expand their effective receptive fields and integrate more contextual information. Atrous convolution is one design that can amplify the receptive fields of a convolutional neural network without significantly increasing the computational effort [32, 33] since the dilation convolution structure is simple, easy to understand, and used directly or indirectly by most papers. Therefore, this paper proposes a contextual aggregation module that not only obtains global information but also provides more detailed local information. And it can be used together with the other two modules proposed in this paper for better performance.

The detailed design of the CAM is shown in Figure 3. Given an input feature map size of $T = C \times H \times W$, the number of channels of the input feature map T is reduced to $C = C/4$ by 1×1 convolution. After that, four parallel dilation convolutions with dilation rate (dilation rate) of [1, 2, 4, 8] are used to

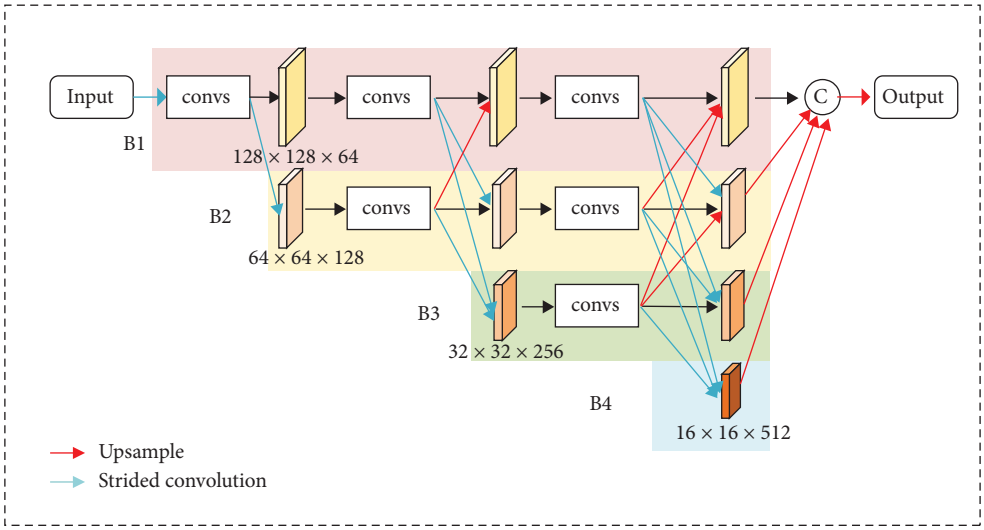

Figure 2: Backbone network.

integrate more contextual information. This method can increase the receptive field of the convolution kernel to obtain a larger range of information while keeping the number of parameters constant and can ensure that the output feature map size remains unchanged. Finally, the convolved feature map is connected to T to obtain $X = 2C \times H \times W$, and then, a 1×1 convolution is used to compress X to $C \times H \times W$. The receptive field is calculated by the following formula:

$$T = 2 \times (r - 1) \times (k - 1) + k, \tag{3}$$

where r denotes the dilation rate, and k is the original convolutional kernel size. In this paper, a parallel dilation convolution method is used to sample features with different dilation rates to obtain feature maps with different receptive fields and then fuse these features with channels to obtain information from different channels. If the size of our input image is 512×512, the output sizes of the four branches of HRNet are $128 \times 128 \times 64$, $64 \times 64 \times 128$, $32 \times 32 \times 256$, and $16 \times 16 \times 512$. First, based on experience, we set the dilation rate of the dilation convolution to 1, 2, 4, and 8 (the dilation convolution with rate = 1 is equivalent to the ordinary convolution), and the receptive fields are calculated to be 3×3, 7×7, 15×15, and 31×31, respectively. Since the output feature maps of the latter two branches of HRNet are 32×32 and 16×16, respectively, they basically cover their main areas and achieve global awareness. Although our four CAM modules are designed the same, the four branches of HRNet have different output scales, and we can also get multiscale information. Finally, a feature map that incorporates these different receptive fields will further improve the performance of the network.

3.4. Side Output Embedding Module (SOEM). Side output embedding module (SOEM) proposed in this paper is a combination of feature pyramid network (FPN) and intermediate supervision. Feature pyramid networks can fuse shallow and deep features and can improve the accuracy of small-volume targets and edge detection. Intermediate supervision allows shallow layers to be trained more fully to avoid gradient disappearance and slow convergence. Therefore, this module allows us to improve our network in both speed and accuracy.

As shown in Figure 4, we use four (F_1, F_2, F_3, F_4) feature maps with different sizes after dual-temporal feature fusion to replace the top-down part of the pyramid network. First, the 1×1 convolved feature map is summed with the up-sampled feature map to obtain a feature map with both coarse-grained features and fine-grained features. The obtained small-size feature maps are then compressed to 2 dimensions by 1×1 convolution and up-sampled to a size of 128×128 (large-size feature maps are not up-sampled). Although obtained four groups of feature maps are of the same size, the semantic levels are different, and the spatial location representation is also different. Finally, these four feature maps are concatenated, compressed to 2 dimensions using 1×1 convolution, and up-sampled to a size of 512×512. And using the truth map (ground truth) and the obtained feature map to calculate the loss, the final prediction map is supervised to be generated. The process of obtaining the forecast map is as follows:

$$\begin{cases} D_{i-1} = \text{sum}\{\text{up}[\text{conv}(F_j)], \text{conv}(F_{j-1})\} i = 4, 3, 2, j = 4, 3, 2, \\ D_4 = \text{conv}(F_4), \end{cases} \text{Map} = \text{sup}\,(\text{fus}\{\text{up}[\text{conv}(D_n)]\}) n = 1, 2, 3, 4, \tag{4}$$

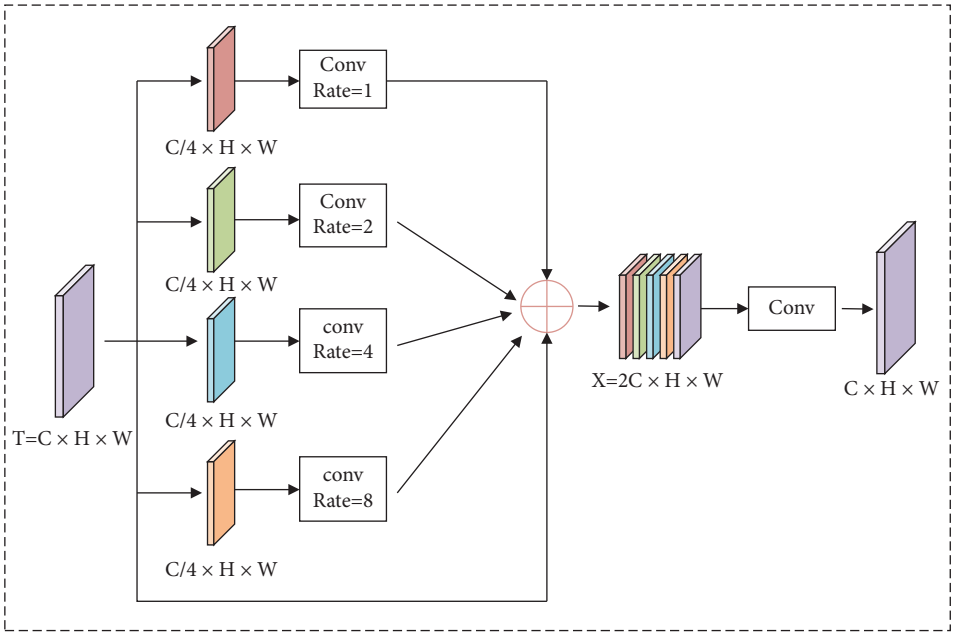

Figure 3: Contextual aggregation module.

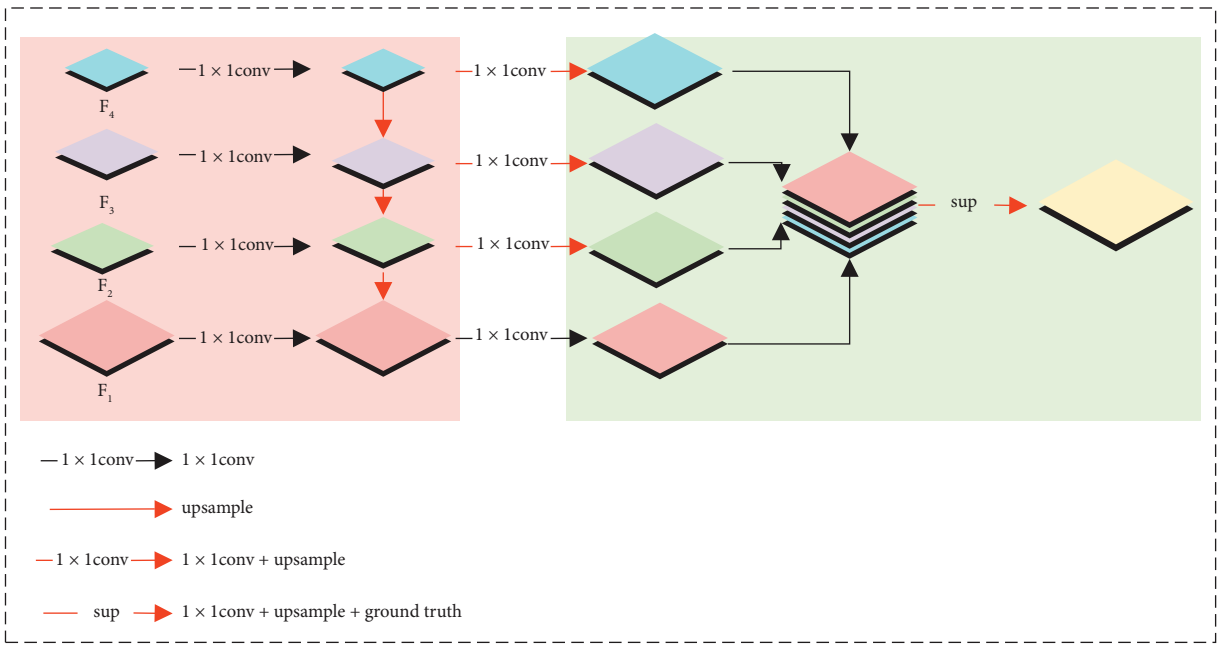

Figure 4: Side output embedding module.

where D_n denotes the feature map obtained by the feature pyramid network, sum{•} denotes the summation of the horizontal and vertical results, up[•] denotes the up-sampling operation, and conv(·) denotes the 1×1 convolution operation, sup(•) denotes compressing the feature map to 2 dimensions and up-sampling to a size of 512×512, and generating the final prediction map using the ground truth with the obtained feature map for loss supervision.

3.5. Loss Function. Due to the problems of significant imbalance of sample categories between changing and non-changing regions, changing targets show diverse scale characteristics and small relative background occupancy. We use a loss function that is a combination of balanced binary cross-entropy and dice coefficient loss that is valid for sample equilibrium, and the loss function L [43] is a weighted sum of the two. The formula is

$$L = L_{\text{bce}} + \lambda L_{\text{dice}},$$
$$L_{\text{bce}} = -\eta \sum_{i \in X} \log \delta(y_i = 1) - (1 - \eta) \sum_{i \in Y} \log \delta(y_i = 0),$$
$$L_{\text{dice}} = 1 - \frac{2Y_{\text{true}}Y_{\text{pred}}}{Y_{\text{true}} + Y_{\text{pred}}}, \tag{5}$$

where L_{bce} is the balanced binary cross-entropy loss; L_{dice} is the dice coefficient loss; λ is the weighting factor, which takes the value of 0.5. Where $\eta = (Y/(X+Y))$, and $1 - \eta = (X/(X+Y))$. X and Y represent the numbers of changed and unchanged pixels in the ground truth label images, respectively. $\delta(\bullet)$ is the sigmoid output at pixel i.

4. Experimental Dataset and Evaluation

To evaluate the effectiveness of the method, we conducted a comprehensive experiment on three datasets, CDD, DSIFN, and SYSU-CD. And precision (P), recall (R), $F1$ score ($F1$), and overall accuracy (OA) were used as evaluation metrics.

4.1. The CDD Dataset. The CDD [44] dataset with real seasonal variation was used for the first experimental data. The dataset contains 7 pairs of images of 4725×2700 pixels in size. To meet the hardware requirements, the original image is sliced into 16,000 sample pairs of size 256×256 pixels. By cropping and rotating 7 pairs of seasonally varying images and dividing them into training set, validation set, and test set in the ratio of 10 : 3:3, the spatial resolution is 3–100 cm.

4.2. The DSIFN Dataset. The second experimental data consist of 6 large, dual-time, high-resolution images covering 6 cities in China (i.e., Beijing, Chengdu, Shenzhen, Chongqing, Wuhan, and Xi'an). The five pairs of dual-time images of Beijing, Chengdu, Shenzhen, Chongqing, and Wuhan are cropped into 394 pairs of subimages of size 512×512. After data enhancement, a collection of 3940 dual-time image pairs was acquired. The Xi'an image pair was cropped into 48 image pairs for model testing. There are 3600 pairs of images in the training dataset, 340 pairs of images in the validation dataset, and 48 pairs of images in the test dataset.

4.3. The SYSU-CD Dataset. The dataset contains 20,000 pairs of aerial images of size 256×256 taken in Hong Kong between 2007 and 2014. The main types of changes in the SYSU-CD dataset include (a) new urban buildings; (b) suburban sprawl; (c) preconstruction foundation works; (d) changes in vegetation; (e) road expansion; and (f) marine construction. The 20,000 pairs of images are divided into a training set, a validation set, and a test set in the ratio of 3 : 1 : 1. There are 12,000 pairs of images in the training data set, 4,000 pairs of images in the validation data set, and 4,000 pairs of images in the test data set.

4.4. Evaluation Metrics. Remote sensing image change detection usually uses precision (P), recall (R), $F1$ score ($F1$), and overall accuracy (OA) as evaluation indexes, as shown in equations (6) to (9). $F1$ score is the summed average of precision and recall, and the higher the $F1$ score, the more robust the model is. In the CD task, a large value of P denotes a small number of false alarms, and a large value of R represents a small number of missed detections. Meanwhile, $F1$ and OA reveal the overall performance, where their larger values will lead to better performance. Four evaluation metrics are described as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{6}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{7}$$
$$F1 = 2\bullet\frac{\text{Precision}\bullet\text{Recall}}{\text{Precision} + \text{Recall}}, \tag{8}$$
$$OA = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{9}$$

where precision represents the precision rate, and recall represents the recall rate. P and N represent the judgment results of the model, T and F are used to evaluate whether the judgment results of the model are correct, FP refers to false positive cases, FN refers to false negative cases, TP refers to true cases, and TN refers to true negative cases.

5. Experimental Design and Results

Our network is implemented by TensorFlow as the backend using the Keras framework. The lab is equipped with a dedicated server for the training of the network, for which we use small batch gradient descent with a batch size of 4. We chose the Adma optimizer to optimize the network, where the initial learning rate for each dataset was set to 0.001, and all experiments were trained for 500 rounds.

5.1. Intermodule Ablation Experiments. In this section, we conduct intermodule ablation experiments on CDD, SYSU-CD, and DSIFN datasets. Table 1 shows the quantitative analysis of the different modules on the two datasets CDD and DSIFN. Table 2 shows the quantitative analysis of the different modules on the SYSU-CD dataset. Figures 5–7 show the qualitative analysis of the two data sets.

5.2. Ablation Study of the Baseline Network. We conduct experiments with vgg_16 and HRNet as the backbone networks, respectively. The experimental results shown in Tables 1 and 2 demonstrate better performance when HRNet is used as the backbone network. Therefore, our benchmark network first uses two twin high-resolution networks HRNet as the feature encoding module for feature extraction.

Table 1: Results of the ablation experiments of the modules on the CDD dataset and DSIFN dataset.

Model	CDD				DSIFN			
	Precision (%)	Recall (%)	*F*1 (%)	OA (%)	Precision (%)	Recall (%)	*F*1 (%)	OA (%)
Baseline (vgg_16)	0.9066	0.7554	0.8083	0.9554	0.8331	0.7766	0.7956	0.9276
Baseline (HRNet)	0.9484	0.8577	0.9007	0.9603	0.8643	0.8355	0.8496	0.9436
HRNet + CAM	0.9555	0.8659	0.9085	0.9673	0.9002	0.8443	0.8714	0.9484
HRNet + CAM + SOME	**0.9670**	**0.8798**	**0.9213**	**0.9788**	**0.9104**	**0.8893**	**0.8987**	**0.9537**

The best ones are marked in bold.

Table 2: Results of the ablation experiments of the modules on the SYSU-CD dataset.

Model	SYSU-CD			
	Precision (%)	Recall (%)	*F*1 (%)	OA (%)
Baseline (vgg_16)	0.8630	0.7681	0.8040	0.9077
Baseline (HRNet)	0.8684	0.7988	0.8455	0.8974
HRNet + CAM	0.8807	0.8197	0.8608	0.9225
HRNet + CAM + SOME	**0.9120**	**0.8393**	**0.8702**	**0.9489**

The best ones are marked in bold.

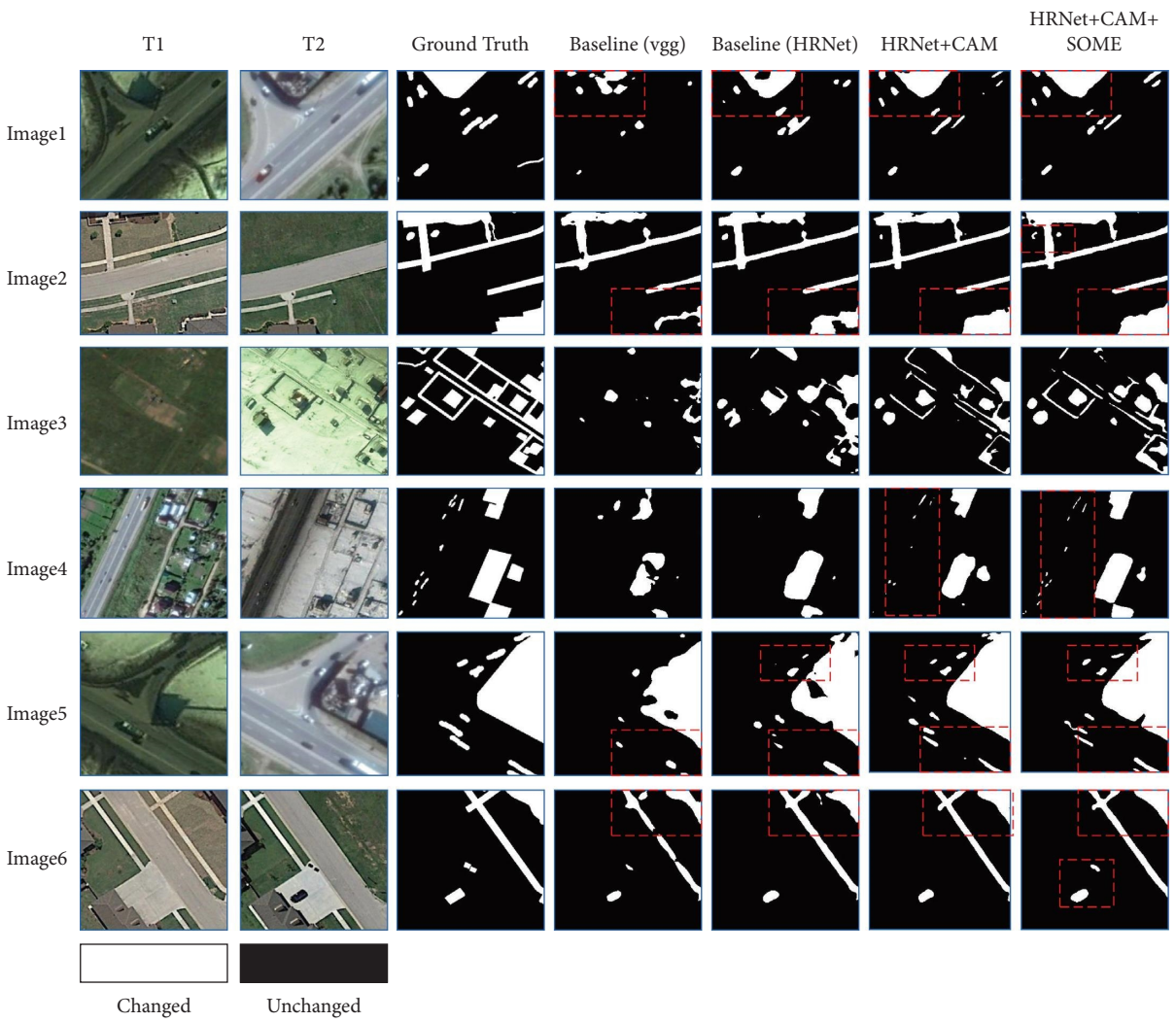

Figure 5: Visualization of the three modular ablation experiments on the CDD dataset. Improved areas are marked with red boxes.

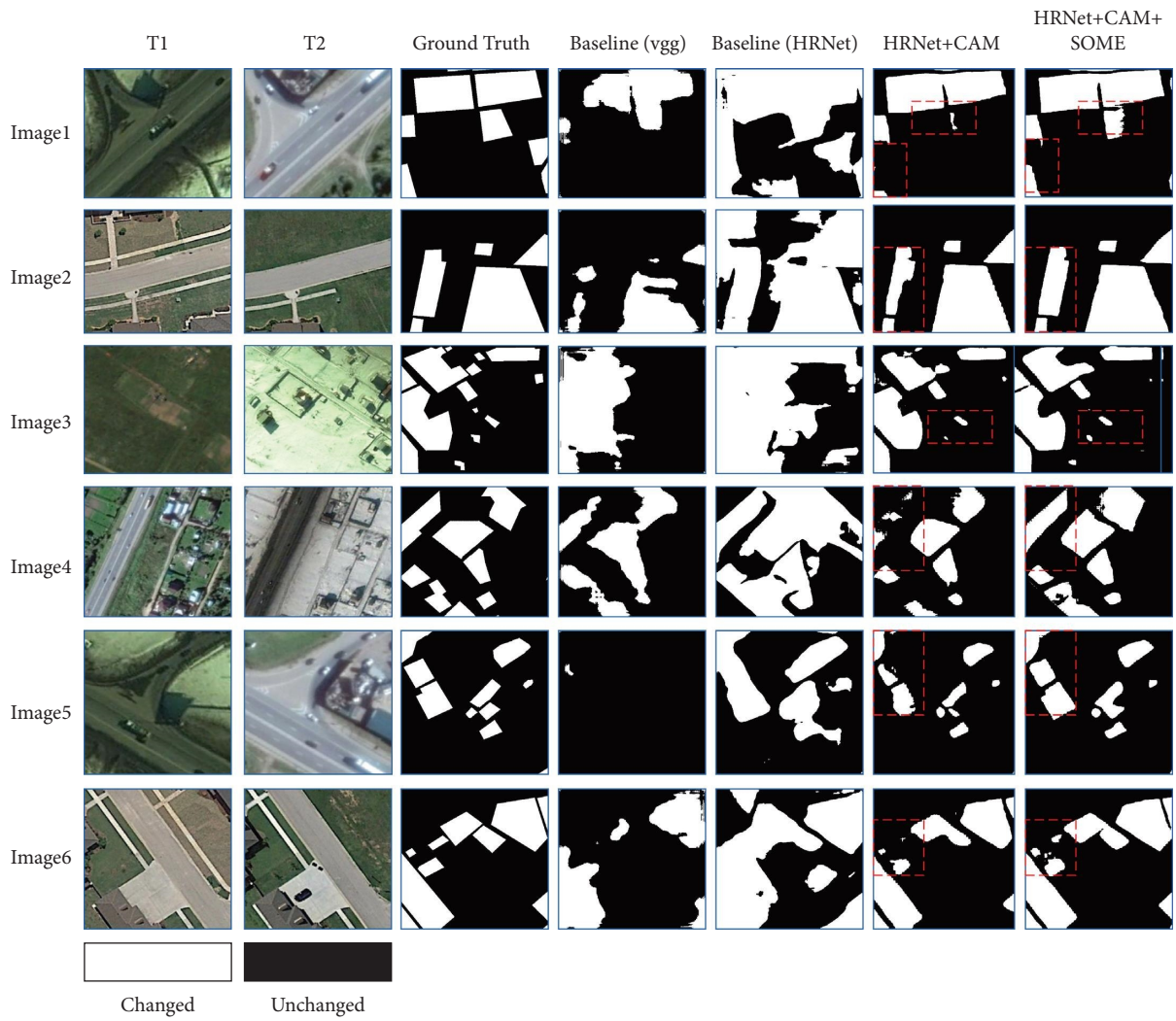

Figure 6: Visualization of the three modular ablation experiments on the DSIFN dataset. Improved areas are marked with red boxes.

Secondly, the feature maps of different sizes after the output of the feature encoding module are first channel-normalized. Then, an up-sampling operation and a tandem splicing operation are performed to obtain the information of eigenvalues with the same dimensionality. Finally, the eigenvalues of dual-temporal images are different, and the absolute values are taken to obtain dual-temporal feature fusion information at different scales. In the feature decoding module, the feature maps of different sizes obtained by fusing the differences are subjected to different levels of up-sampling operations and fused for output. We quantitatively evaluated the performance of the benchmark network as shown in the second row of Tables 1 and 2. The fourth and fifth columns in Figures 5–7 visualize that the results with HRNet as the backbone network are better than those with vgg_16 as the backbone network. The general outline of HRNet as the backbone network is shown in the visualization.

5.3. Ablation Studies of CAM. We have designed the contextual aggregation module (CAM). This module can amplify the convolutional neural network receptive field without increasing the computational effort, obtaining not only global but also (for different channels) detailed local contextual information. As can be seen from Table 1, the addition of the CAM module to the baseline network results in a significant improvement in all metrics. On the CDD dataset, the precision (P), recall (R), $F1$ score, and overall accuracy (OA) were 95.55%, 86.59%, 90.85%, and 96.73%, respectively, with the addition of this module to the baseline network. Compared to the baseline network, P, R, $F1$, and OA are improved by 0.71%, 0.82%, 0.78%, and 0.70%, respectively. On the DSIFN dataset, the precision (P), recall (R), $F1$ score, and overall accuracy (OA) were 90.02%, 84.43%, 87.14%, and 94.84%, respectively, after adding the module. Compared to the baseline network, P, R $F1$, and OA improved by 3.59%, 0.88%, 2.18%, and 0.48%, respectively. On the SYSU-CD dataset, the precision (P), recall (R), $F1$ score, and overall accuracy (OA) were 88.07%, 81.97%, 87.77%, and 92.25%, respectively, with the addition of this module to the baseline network. Compared to the baseline network, P, R, $F1$, and OA are improved by 1.23%, 2.09%, 1.53%, and 2.51%, respectively. From the sixth column of Figures 5–7, it is clear that the CAM module has improved the boundaries compared to the baseline network, and some

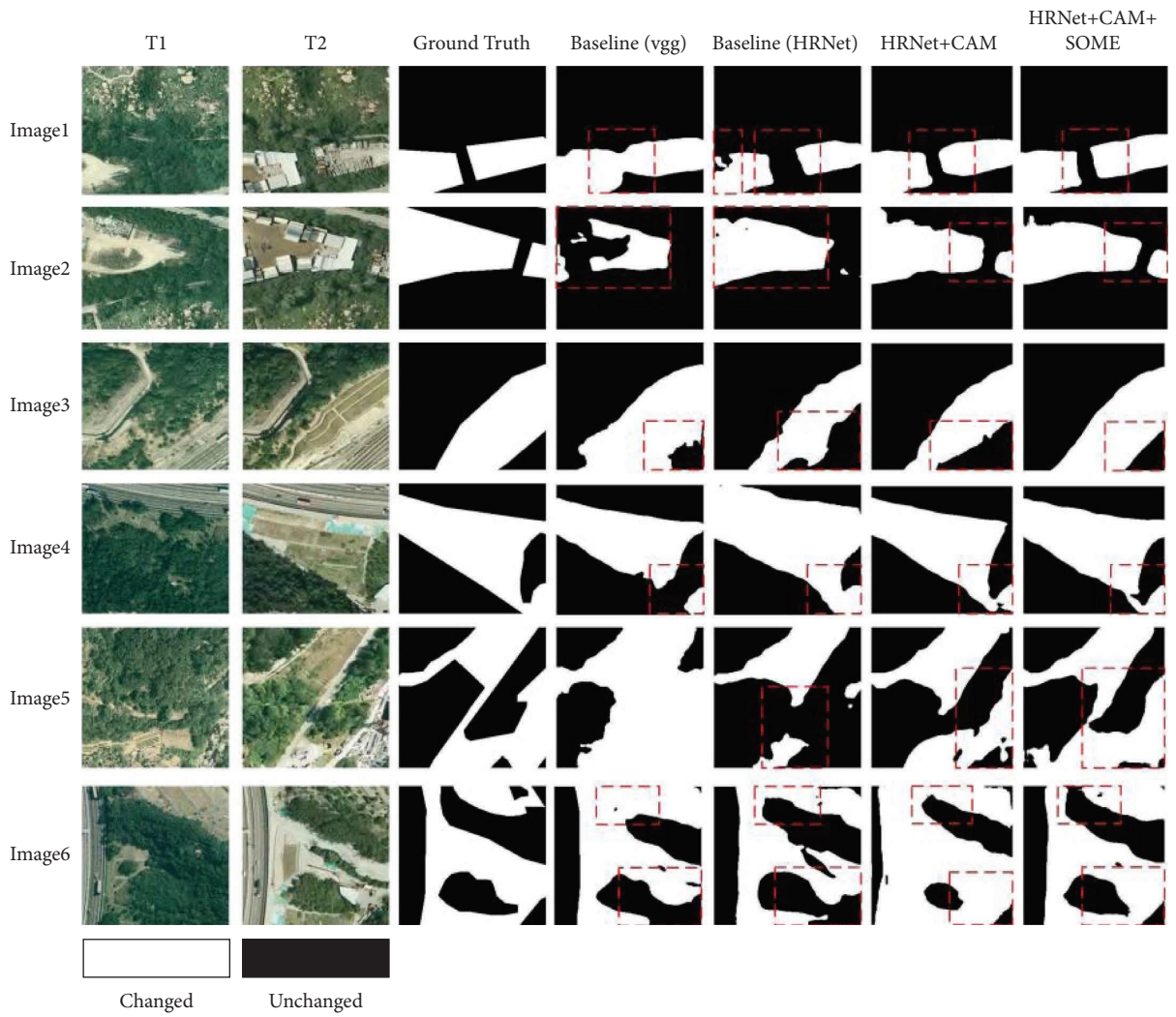

Figure 7: Visualization of the three modular ablation experiments on the SYSU-CD dataset. Improved areas are marked with red boxes.

small targets can be seen in the sixth column of Figures 5 and 6. Its outline has been fully revealed, but there are still some detailed features that have not fully emerged and need to be extracted further. This indicates that parallel atrous convolution of multiple channels ensures maximum information extraction.

5.4. Ablation Studies of SOEM. We also studied the contribution of the SOEM module to the network. Side output embedding module (SOEM) can fuse shallow and deep features and can improve the accuracy of small-volume targets and edge detection. And it enables the shallow layer to be trained more fully to avoid gradient disappearance and too slow convergence. As can be seen in Table 1, the significant gains in precision (P), recall (R), $F1$ score, and overall accuracy (OA) were achieved with the addition of the SOEM module compared to the first two ablation experiments. On the CDD dataset, the addition of this module improves P, R, $F1$, and OA by 1.86%, 2.21%, 2.06%, and 1.83%, respectively, compared to the baseline network. Compared to baseline + CAM, P, R, $F1$, and OA improved by 1.15%, 1.39%, 1.28%, and 1.15%, respectively. On the DSIFN dataset, P, R, $F1$, and OA improved by 4.97%, 5.38%, 4.91%, and 1.01%, respectively, compared to the baseline network after adding this module. Compared to baseline + CAM, P, R, $F1$, and OA were improved by 1.02%, 4.50%, 2.73%, and 0.53%, respectively. On the SYSU-CD dataset, the addition of this module improves P, R, $F1$, and OA by 4.36%, 4.05%, 2.47%, and 5.15%, respectively, compared to the baseline network. Compared to baseline + CAM, P, R, $F1$, and OA improved by 3.13%, 1.96%, 0.94%, and 2.64%, respectively. As can be seen in the seventh column of Figures 5–7, the addition of this module not only improves the detection performance in general but also makes its edges more complete. And the seventh column of Figures 5 and 6 can find that some small targets are displayed more accurately. At the same time, the detailed features extracted by this module make the predicted result maps closer to the real labels. Therefore, using three modules together enables the network to be used to its best advantage.

Table 3: Results of the quantitative evaluation of the different methods on the CDD and DSIFN datasets.

Models	CDD				DSIFN			
	Precision (%)	Recall (%)	*F*1 (%)	OA (%)	Precision (%)	Recall (%)	*F*1 (%)	OA (%)
FCN-PP	0.8264	0.8060	0.8047	0.9536	0.5640	0.6703	0.6126	0.8559
FC-siam-conc	0.8441	0.8250	0.8250	0.9572	0.4183	0.5963	0.4917	0.7905
FC-siam-diff	0.8578	0.8364	0.8373	0.9575	0.5151	0.6554	0.5769	0.8366
Unet++_MSOF	0.8954	0.8711	0.8756	0.9673	0.5983	0.6591	0.6273	0.8668
IFN	0.9496	0.8608	0.9030	0.9771	0.6711	0.6754	0.6733	0.8886
TCANet (ours)	**0.9670**	**0.8798**	**0.9213**	**0.9788**	**0.9104**	**0.8893**	**0.8987**	**0.9537**

The best ones are marked in bold.

Table 4: Results of the quantitative evaluation of the different methods on the SYSU-CD datasets.

Models	SYSU-CD			
	Precision (%)	Recall (%)	*F*1 (%)	OA (%)
FCN-PP	0.7432	0.7584	0.7507	0.8509
FC-siam-conc	0.8254	0.7103	0.7635	0.8660
FC-siam-diff	0.8913	0.6121	0.7257	0.8611
Unet++_MSOF	0.8752	0.7260	0.7694	0.8884
IFN	0.8784	0.8333	0.8565	0.9111
TCANet (ours)	**0.9120**	**0.8393**	**0.8702**	**0.9489**

The best ones are marked in bold.

5.5. Comparative Experiments. In order to show the superiority of our proposed method, a comparison with five other change detection networks is presented in this paper. Five change detection networks include full convolutional network with pyramid pool (FCN-PP) [45], fully convolutional siamese-concatenation (FC-siam-conc) [46], fully convolutional siamese-difference (FC-siam-diff) [46], Unet++_MSOF [26], and IFN [22]. Table 3 shows the experimental comparison of our method with the other five methods on the CDD and DSIFN datasets. Table 4 shows the experimental comparison of our method with the other five methods on the SYSU-CD dataset. As shown in Figure 8, the performance of different methods on CDD, DSIFN, and SYSU-CD datasets is quantitatively analyzed using a line graph format. Figures 9–11 show the visualization of the different methods on CDD, DSIFN, and SYSU-CD datasets.

As shown in Table 3, we quantitatively evaluated the results of TCANet and different methods on the CDD and DSIFN datasets. As shown in Table 4, we quantitatively evaluated the results of TCANet and different methods on the SYSU-CD dataset. As shown in Figure 8, our network has the highest performance metrics on three datasets. On the CDD dataset, precision, recall, *F*1 score, and OA of TCANet were 96.70%, 87.98%, 92.13%, and 97.88%, respectively, compared to IFN by 1.00%, 0.18%, 0.55%, and 0.17%, respectively. On the DSIFN dataset, these three metrics for TCANet were 91.04%, 88.93%, 89.87%, and 95.37%, respectively, compared to IFN by 2.19%, 3.73%, 3.18%, and 6.51%, respectively. On the SYSU-CD dataset, precision, recall, *F*1 score, and OA of TCANet were 91.20%, 83.93%, 87.02%, and 94.89%, respectively, compared to IFN by 3.36%, 0.60%, 1.37%, and 3.78%, respectively. Figures 9–11 show the visualization of the different methods on the three datasets. The red boxes represent the improvement areas. In comparing the visualization results of the experiments, it can be seen that the prediction maps of our proposed method are closer to the real labels, thus demonstrating the effectiveness of our proposed method.

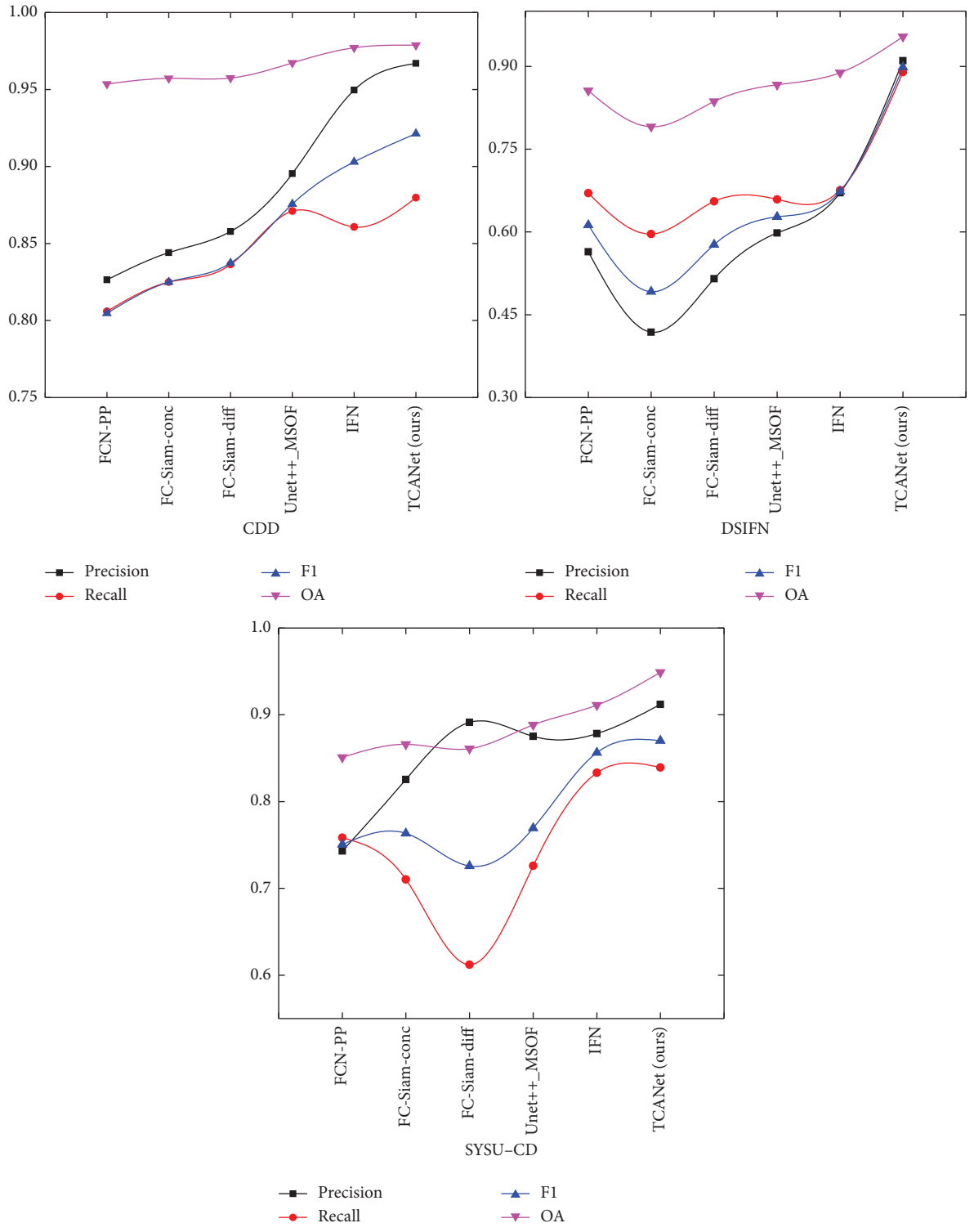

Figure 8: A shows the comparison of different networks on the CDD test set, and B shows the comparison of different networks on the DSIFN test set.

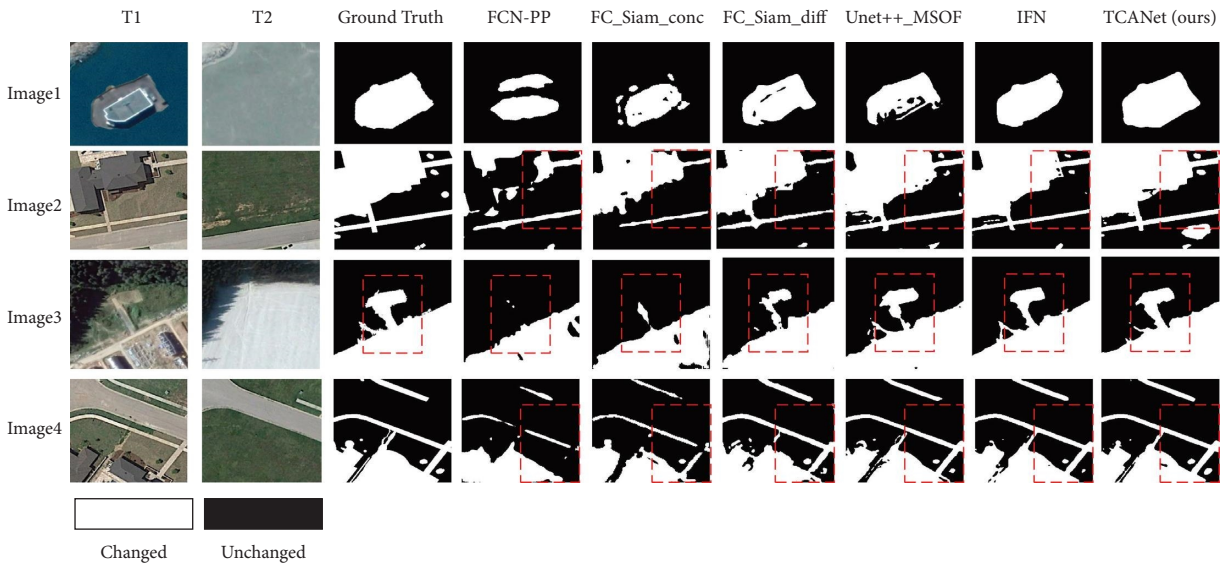

Figure 9: Visualization of different methods on the CDD test set.

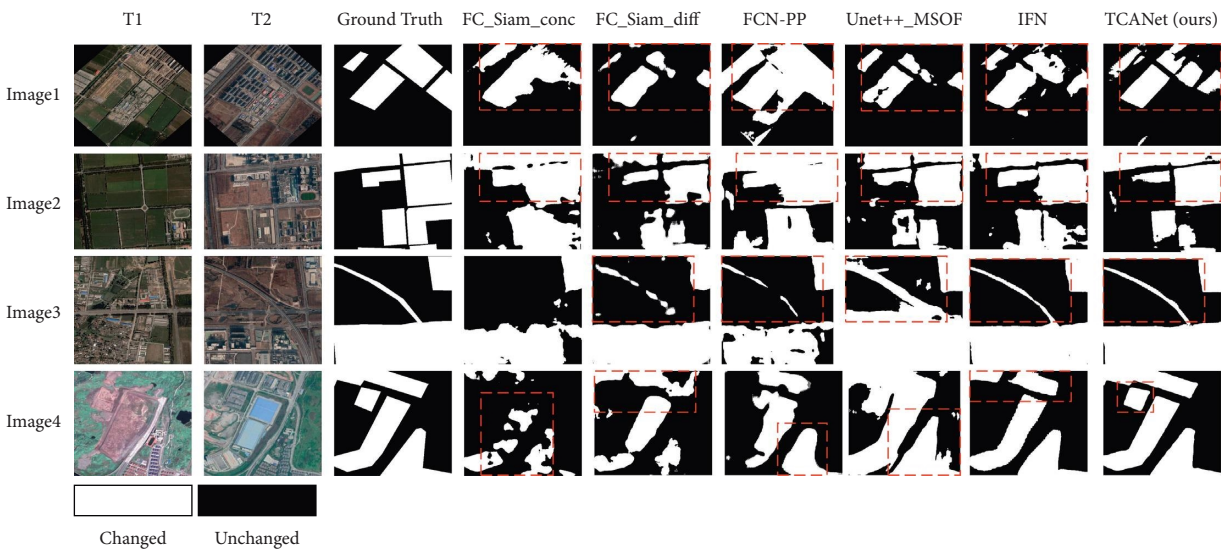

Figure 10: Visualization of different methods on the DSIFN test set.

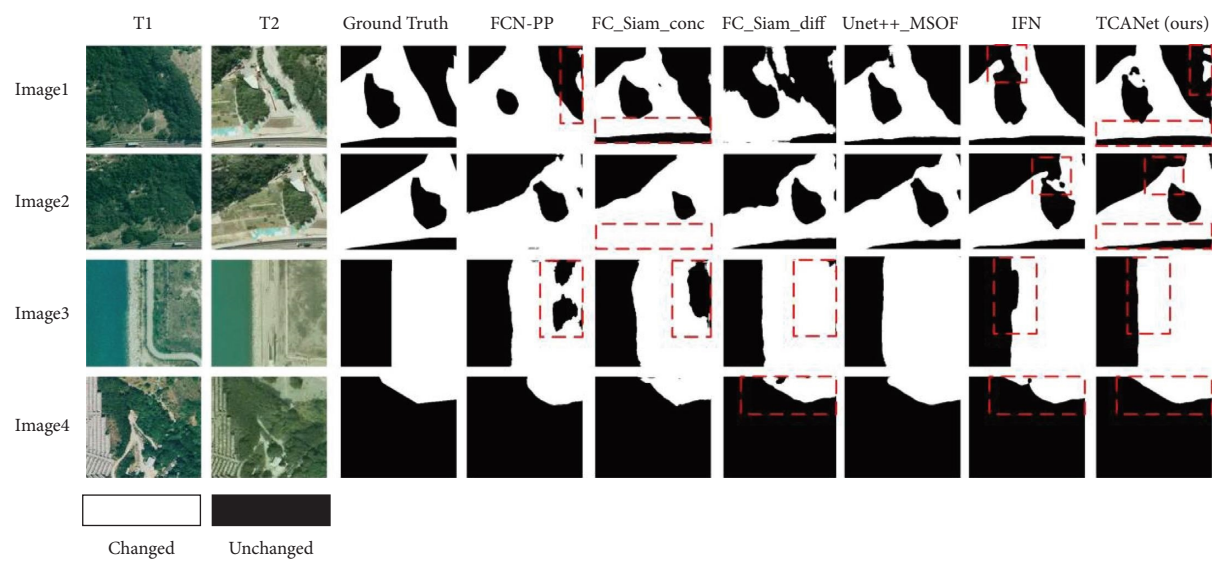

Figure 11: Visualization of different methods on the SYSU-CD test set.

6. Conclusion

In this paper, a twins context aggregation network (TCA-Net) is investigated. This study extracts features separately by feeding dual-temporal images into two networks with shared parameters. In the feature extraction stage, the limitations of the traditional "encoder-decoder" structure are considered. We introduce a parallel multiscale branching HRNet to reduce the loss of spatial information. In addition, we designed separate contextual aggregation modules (CAM) for each branch, expanding their effective receptive field and integrating more contextual information. Then, the two feature values obtained from the feature encoding stage are differenced, and the absolute values are taken to obtain dual-temporal feature fusion information at different scales. Finally, we input the fused feature maps to the side output embedding module to facilitate the detection of edges and small targets. Our proposed architecture shows a greatly improved approach to existing architecture and achieves the better results on three remote sensing image datasets (CDD, DSIFN, and SYSU-CD datasets). One of the limitations of the method is that in order to avoid intensive computation, HRNet reduces the space size of the input data in the early layers.

In the future, we will plan to improve the training speed and accuracy of change detection by reducing the computation and trying to merge more parallel branches of HRNet. Since transformer has been widely used in the remote sensing field, many scholars have migrated transformer to the semantic segmentation field, so I will plan to introduce transformer to change detection in the next step.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key Research and Development Program under Contract 2017YFB0504203, Planned project of Gansu Science and Technology Department under Contract 21JR7RA310, and Youth Science Fund Project of Lanzhou Jiaotong University under Contract 2021029.

References

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: a brief review," *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 7068349, 13 pages, 2018.

[2] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8599–8603, IEEE, Vancouver, BC, Canada, May 2013.

[3] H. Palangi, L. Deng, Y. Shen et al., "Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 694–707, 2016.

[4] E. Choi and J. Kim, "Robust change detection using channel-wise co-attention-based siamese network with contrastive loss function," *IEEE Access*, vol. 10, pp. 45365–45374, 2022.

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

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, IEEE, Piscataway, NJ, USA, June 2015.

[7] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 784–788, 2018.

[8] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 6000–6010, 2017.

[9] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks*, pp. 729–734, IEEE, Montreal, QC, Canada, July 2005.

[10] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun, "Spectral networks and deep locally connected networks on graphs," 2014, https://arxiv.org/abs/1312.6203.

[11] G. Xian, C. Homer, and J. Fry, "Updating the 2001 National Land Cover Database land cover classification to 2006 by using Landsat imagery change detection methods," *Remote Sensing of Environment*, vol. 113, no. 6, pp. 1133–1147, 2009.

[12] A. P. Tewkesbury, A. J. Comber, N. J. Tate, A. Lamb, and P. F. Fisher, "A critical synthesis of remotely sensed optical image change detection techniques," *Remote Sensing of Environment*, vol. 160, pp. 1–14, 2015.

[13] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.

[14] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 772–776, 2009.

[15] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3677–3693, 2019.

[16] H. Chen, C. Wu, B. Du, and L. Zhang, "Deep Siamese multi-scale convolutional network for change detection in multi-temporal VHR images," in *Proceedings of the 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pp. 1–4, IEEE, Shanghai, China, August 2019.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds., pp. 234–241, Springer, Berlin, Germany, 2015.

[18] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 4063–4067, IEEE, Athens, Greece, October 2018.

[19] F. Rahman, B. Vasu, J. Van Cor, J. Kerekes, and A. Savakis, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *Proceedings of the 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 958–962, IEEE, Anaheim, CA, USA, November 2018.

[20] 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 Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

[21] E. Guo, X. Fu, J. Zhu et al., "Learning to measure change: fully convolutional siamese metric networks for scene change detection," 2018, https://arxiv.org/abs/1810.09111.

[22] C. Zhang, P. Yue, D. Tapete et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.

[23] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[24] J. Chen, Z. Yuan, J. Peng et al., "DASNet: dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2021.

[25] L. Lan, D. Wu, and M. Chi, "Multi-temporal change detection based on deep semantic segmentation networks," in *Proceedings of the 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pp. 1–4, IEEE, Shanghai, China, August 2019.

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

[27] B. Hou, Y. Wang, and Q. Liu, "Change detection based on deep features and low rank," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2418–2422, 2017.

[28] X. Yu, J. Fan, J. Chen, P. Zhang, Y. Zhou, and L. Han, "NestNet: a multiscale convolutional neural network for remote sensing image change detection," *International Journal of Remote Sensing*, vol. 42, no. 13, pp. 4898–4921, 2021.

[29] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: a densely connected Siamese network for change detection of VHR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[30] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: looking wider to see better," 2015, https://arxiv.org/abs/1506.04579.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, IEEE, Piscataway, NJ, USA, July 2017.

[32] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, https://arxiv.org/abs/1511.07122.

[33] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," 2014, https://arxiv.org/abs/1412.7062.

[34] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, https://arxiv.org/abs/1706.05587.

[35] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, Amsterdam, Netherlands, September 2018.

[36] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, IEEE, Piscataway, NJ, USA, June 2018.

[37] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[39] H. Zhang, K. Dana, J. Shi et al., "Context encoding for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, New Orleans, LA, USA, June 2018.

[40] H. Zhao, Y. Zhang, and S. Liu, "Psanet: point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 267–283, Amsterdam, Netherlands, September 2018.

[41] K. Sun, Y. Zhao, and B. Jiang, "High-resolution representations for labeling pixels and regions," 2019, https://arxiv.org/abs/1904.04514.

[42] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, Piscataway, NJ, USA, June 2019.

[43] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[44] 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," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, no. 2, pp. 565–571, 2018.

[45] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 982–986, 2019.

[46] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing*, Athens, Greece, October 2018.