FDFF-Net: A Full-Scale Difference Feature Fusion Network for Change Detection in High-Resolution Remote Sensing Images

Feng Gu, Pengfeng Xiao, *Senior Member, IEEE*, Xueliang Zhang, *Member, IEEE*, Zhenshi Li, and Dilxat Muhtar

***Abstract*—Deep-learning techniques have made significant advances in remote sensing change detection task. However, it remains a great challenge to detect the details of changed areas from high-resolution remote sensing images. In this study, we propose a full-scale difference feature fusion network (FDFF-Net) for change detection, which can alleviate pseudochanges and reduce the loss of change details during detection. In the encoding stage, a dense difference fusion module is proposed to effectively mine and fuse the multiple differences for each feature level between bitemporal images, leading to a substantial reduction in missed detection of change areas. Additionally, the different levels of difference features are aggregated through a full-scale skip connection, allowing the network to detect multiple changed objects with various sizes. In the decoding stage, a strip spatial attention module is designed to enhance the perception of the change areas, which improves the ability to detect detailed changes. The experiments on three change detection datasets, CDD, LEVIR-CD, and S2Looking, demonstrate that FDFF-Net outperforms the compared state-of-the-art methods and can detect more complete changes of small objects and clear contours of changed areas.**

***Index Terms*—Attention mechanism, change detection, deep learning, difference feature fusion.**

I. Introduction

CHANGE detection is the technique that identifies changes occurred on the Earth's surface by using images acquired on the same geographical area at different times [1], which has extensive applications in varied fields, such as urban landscape monitoring [2], agricultural investigation [3], land cover mapping [4], and natural resource management [5]. With the advancement of Earth observation technology, a large quantity of high-resolution remote sensing images with abundant ground object information is available [6]. However, change detection methods relying on manual features suffer from limited generalization ability and robustness, resulting in inferior performance when applied to complex scenes. Deep learning has revolutionized remote sensing change detection by enabling the automatic learning of representative and discriminatory features directly from images [7].

Various deep neural networks have been used for change detection and have achieved admirable results [8], [9], [10], [11]. Full convolutional neural network (FCN) [12] replaces the last fully connected layer of convolution neural network (CNN) [13] with a convolution layer, which is suitable for pixel-level prediction, and thus used in pixelwise change detection as a basic structure [14], [15], [16], [17], [18]. UNet is a typical FCN [19], which obtains the features of different levels in the encoding stage, restores the image resolution in the decoding stage, and recovers the semantic information lost in the downsampling through the skip connection. By utilizing aggregated multilevel features, UNet has achieved excellent accuracy in change detection task [20], [21]. Based on UNet, UNet++ [22] uses denser skip connections to aggregate features at different levels. Comparing the change detection results on the same dataset, UNet++ has achieved higher accuracy than UNet [23], which shows that the efficient use of features at different levels can achieve advantages in change detection [24], [25]. Moreover, UNet3+ [26] has been used as a backbone network for change detection, which has a denser skip connection than UNet++.

In addition to connecting the multilevel features of the encoder with the decoder, the strategy of aggregating features at different levels in both the encoder and decoder is used to improve the detection capability. For example, CLNet [27] extracted multiscale and multilevel features in the encoder to improve the ability of detecting pixel-level changes. ADS-Net [28] used various weights to fuse the difference features of different levels in the decoder to generate the predicted change map. Mining and utilizing multilevel features of images to improve the identification capacity of networks on various ground objects is a crucial aspect to be considered. However, it is equally important to focus on the time correlation between images during change detection, in order to suppress the interference of pseudochanges. Therefore, after obtaining features of different phases, the method of fusing the features from different phases needs to be further considered.

The change detection feature fusion methods could be divided into prefusion and postfusion [29]. The prefusion method

Manuscript received 30 June 2023; revised 4 October 2023 and 24 October 2023; accepted 15 November 2023. Date of publication 21 November 2023; date of current version 3 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42071297 and in part by the Fundamental Research Funds for the Central Universities under Grant 020914380119. *(Corresponding author: Pengfeng Xiao.)*

The authors are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: 502022270111@smail.nju.edu.cn; xiaopf@nju.edu.cn; zxl@nju.edu.cn; lzhenshi@outlook.com; 502022270062@smail.nju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3335287

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

involves concatenating images from different phases as the input of the network, then fuses them and extracts the features in the encoder, and finally generates the change map in the decoder. However, this approach fails to capture the multilevel features of individual original images, leading to a loss of details that cannot be adequately compensated during upsampling. Postfusion method, on the other hand, allows the bitemporal images to be fed into the dual-stream structure to extract respective features and then fuses them in the decoder to generate a change map, which has been proved better than the prefusion method [30]. Some studies made improvements based on the two fusion paradigms, such as introducing the gate module [31] and using a differential pyramid [32] to highlight the changing features.

The utilization of attention mechanisms [33] in feature fusion can also significantly improve the results of change detection since the attention mechanism is able to improve the effectiveness of feature fusion and obtain discriminant feature representations. The attention model used for change detection mainly includes the convolution-based attention model [34], [35] and the self-attention model [36], [37]. The convolution-based attention module adaptively selects and enhances input features in both channel and spatial dimensions, enhancing change features while restraining irrelevant features. For example, DDCNN [25] integrated both spatial attention and channel attention in the network, which enhanced the representation of detailed change. DSA-Net [38] introduced a cross-layer connection module guided by the designed spatial attention model to focus on the change areas. The self-attention model can obtain global context information by calculating the correlation between each pixel and other pixels in the image. For example, H-SALENet [39] utilized multilayer and multihead self-attention to enhance the representation capability of hierarchical and long-range-dependent features, and CSANet [40] combined self-attention to improve the feature representation of images at different phases and used cross-temporal attention to integrate different embedded features.

Despite the remarkable improvement achieved by deep-learning-based approaches in change detection, certain limitations persist that impact the detection ability of intricate change details. First, the existing methods focus on obtaining the multilevel features of a single-temporal image and neglect to explore the connections between bitemporal image features. If the bitemporal image features are concatenated directly, there will be a problem of heterogeneity feature fusion when utilizing the concatenated features as a supplement of the difference features to recover the lost semantic information. On the other hand, subtracting the bitemporal image features can overcome this problem. However, due to the limited exploration of feature differences, it results in missed detections of the changed areas. Second, the convolution-based attention module has a limited receptive field, making it unable to obtain long-range dependence. While the utilization of a self-attention module enables the network to obtain long-range contextual information, it also increases computational complexity due to calculating the correlation between each pixel and other pixels within the feature map. Therefore, in the task of change detection, when designing attention modules to capture the spatial relationship of long range, the shape characteristics of changed objects, especially the small objects, should be taken into consideration to avoid unnecessary connections between distant locations.

In this study, we design a full-scale difference feature fusion network (FDFF-Net) to overcome the aforementioned problems. Considering the integrity of the detected change areas, the dense difference fusion module (DDFM) is designed to mine the multiple differences between the same level features before feature fusion, which helps to comprehensively explore the difference between features. Moreover, we employ a full-scale skip connection to concatenate the multilevel difference features, enabling the effective detection of changes in ground objects of varying sizes. Furthermore, to enhance the representation of change area edges and strip-shaped small objects, We propose a strip spatial attention module (SSAM). This module combines a spatial attention mechanism with strip pooling, applied prior to each upsampling, thereby enhancing the representation of change area edges and small strip-shaped objects. The main contributions of this study are as follows.

1) An FDFF-Net is proposed for change detection, which makes full use of multilevel differences between bitemporal images, allowing to detect a variety of changed objects.
2) A DDFM is designed to mine and fuse the multiple differences for each feature level between bitemporal images, which improves the completeness of the detected changed objects and the ability to alleviate pseudochanges.
3) A SSAM is designed to capture local spatial details and long-range dependencies of discrete regions, helping to detect the details of changed areas.

II. Methodology

A. Network Architecture

FDFF-Net adopts the typical encoder–decoder architecture (see Fig. 1). The bitemporal images are first fed into a fully convolutional dual-stream feature extraction structure with shared weights to obtain the features of different levels. In the encoding stage, DDFM is proposed to connect the features of bitemporal images to obtain difference feature. By mining multiple differences for each feature level, a complete representation of the changes is acquired. In the decoding stage, the multilevel of difference features is concatenated through a full-scale skip connection [26]. This dense connection can integrate fine-grained details and coarse-grained semantics, allowing the network to detect multiple changed objects with various sizes. SSAM is designed to enhance the perception of changing areas, especially small objects, during the fusion of full-scale difference features. Capturing local spatial features and long-range dependencies can help networks focus on change areas, especially the detailed changes.

In the encoding stage, FDFF-Net uses VGG16 with parameters pretrained on ImageNet to extract the multilevel features of bitemporal images (T1 and T2), respectively. For each input image, the encoder can extract five features at different levels, with respective sizes of 256×256, 128×128, 64×64, 32×32, and 16×16. The channels of those feature maps are 64, 128, 256, 512, and 512, respectively. These feature

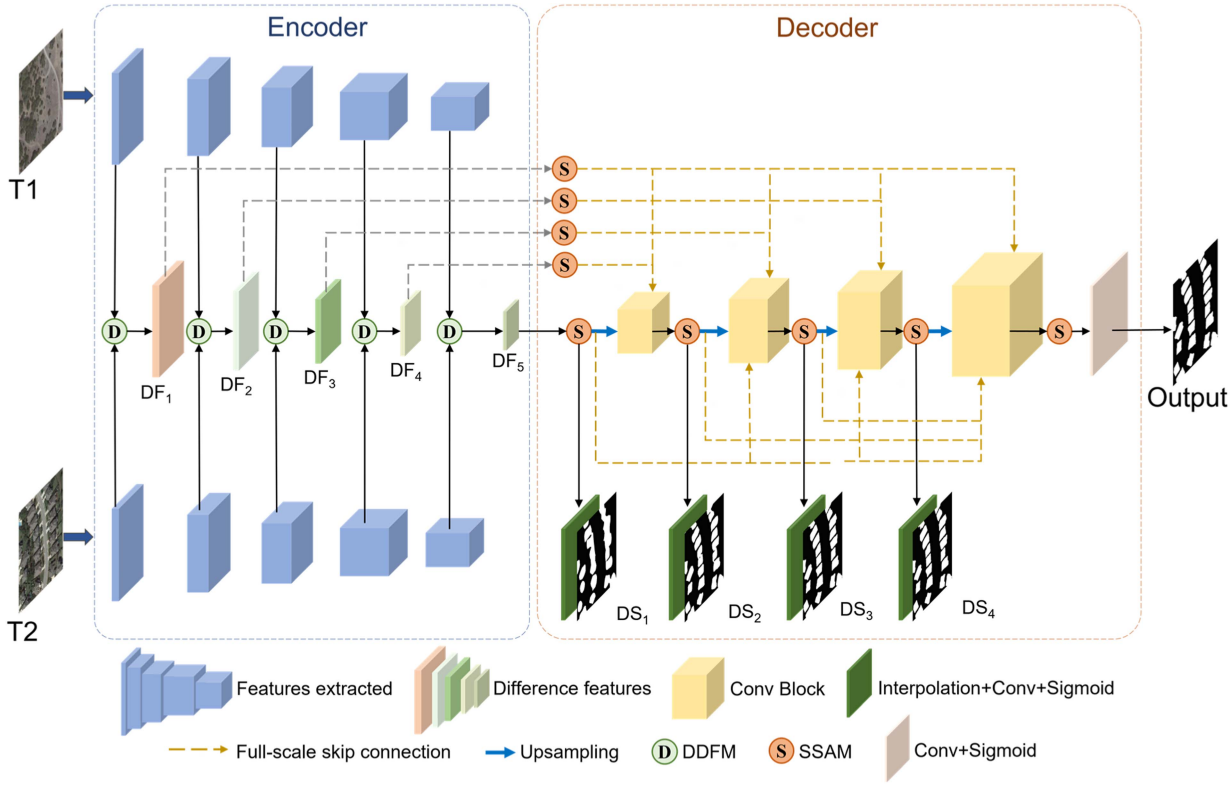


Fig. 1. Structure diagram of the FDFF-Net. DDFM represents the dense difference fusion module, and SSAM represents the strip spatial attention module.

maps are fed into DDFM to obtain the difference features (DF_1, DF_2, DF_3, DF_4, and DF_5) with reduced channels of 64. In the decoding stage, after connecting the difference features at different levels by full-scale skip connection, each decoder layer contains five difference features with a channel number of 320. Then, a conv block is used to fuse difference features. SSAM is used to enhance the attention to detailed changes before each upsampling. After four upsampling operations, the spatial resolution is restored to 256 × 256, generating the predicted change map by a 1 × 1 convolution and activating by a Sigmoid function. Moreover, in each decoder layer, four change maps in the intermediate layer (DS_1, DS_2, DS_3, and DS_4) are obtained by a 1 × 1 convolution and bilinear interpolation, resulting in maps of equal size to the input image. Moreover, during the training stage, individual loss values are calculated for each intermediate layer of the decoder, which helps to alleviate the vanishing gradient problem. The total loss value is determined through the utilization of the four intermediate layer change maps and final predicted change map, which is computed by summing up the five individual losses, where each loss is assigned an equal weight of 1.

B. Dense Difference Fusion Module

The accuracy of detected changes in Siamese architecture-based change detection networks is influenced by the connection methods employed for bitemporal features. If only the original image features are subtracted to generate the difference features, the difference between the features is not completely mined, which may result in the occurrence of missed detection. Furthermore, seasonal variation, shadows, and illumination can interfere with the detection of actual changes, resulting in the detection of pseudochanges. Therefore, we propose a DDFM to extract the multiple difference features between bitemporal images, which helps to obtain substantive change and improve the integrity of detected change objects.

The DDFM has three branches (see Fig. 2). In the first branch, DDFM calculates the value of elementwise addition between bitemporal features to enhance the representation of contour information. In this branch, the channel dimension of the feature map is reduced to 64 by a convolution layer of 1 × 1. The details are given as follows:

$$DF_a = f^{1\times1}\ (F1 + F2) \tag{1}$$

where $F1$ and $F2$ denote the bitemporal image features, respectively, and $f^{1\times1}$ denotes the 1 × 1 convolution operation.

In the second branch, DDFM uses multiple convolutions to obtain features at different scales. If only the fixed receptive field of convolution is used to extract the difference features, the context information is unable to be effectively utilized, resulting in limited ability to distinguish changes and background. While applying convolution kernels of varying sizes during processing, the extraction of multiscale features is possible. However, additional parameters are introduced to increase the computational effort. Therefore, DDFM uses atrous convolutions [41] to replace the larger convolution kernel in the second branch for extracting difference features at different scales. This branch consists of three convolution operations with the 3 × 3 kernel

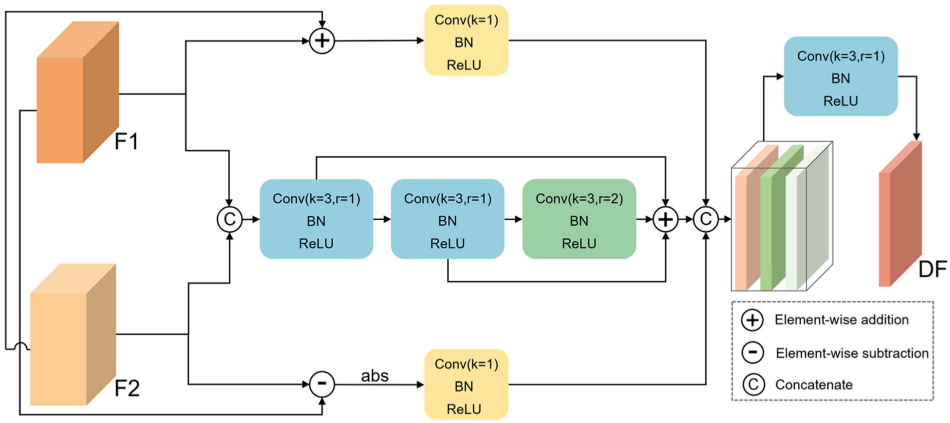


Fig. 2. DDFM. $F1$ and $F2$ are the input features from different phase images and DF is the difference feature obtained using this module.

size, including two normal convolutions with a dilation rate of 1 and one atrous convolution with a dilation rate of 2. The number of channels in each output feature map is 64. In addition, the dense connection is used to correct the difference feature. The operations are as follows:

$$DF_b = DF_{b1} + DF_{b2} + DF_{b3} \tag{2}$$

$$DF_{b1} = f_{r=1}^{3\times3}(F) \tag{3}$$

$$DF_{b2} = f_{r=1}^{3\times3}(DF_{b1}) \tag{4}$$

$$DF_{b3} = f_{r=2}^{3\times3}(DF_{b2}) \tag{5}$$

where F denotes the feature after concatenating bitemporal image features ($F1$ and $F2$) together; DF_{b1}, DF_{b2}, and DF_{b3} denote the difference features obtained using the three convolution operations, respectively; $f_{r=1}^{3\times3}$ denotes the 3×3 convolution operation; $f_{r=2}^{3\times3}$ denotes the 3×3 convolution operation with a dilation rate of 2.

In the third branch, DDFM calculates the absolute value of elementwise subtraction between bitemporal features. Although the direct subtraction of bitemporal features cannot obtain the accurate difference features, the approximate location of the change areas can be retained. In this branch, the channel dimension of the feature map is also reduced to 64 by a convolution layer of 1×1. The details are as follows:

$$DF_c = f^{1\times1}(|F1 - F2|). \tag{6}$$

Finally, DDFM concatenates multiple feature maps obtained by the three branches. Then, these features are fused to obtain difference feature by a 3×3 convolution operation, as follows:

$$DF = f_{r=1}^{3\times3}([DF_a; DF_b; DF_c]) \tag{7}$$

where $[;]$ denotes the concatenate operation, and DF is the output difference feature.

C. Strip Spatial Attention Module

Attention mechanisms can enhance the perception of change areas and obtain discriminant feature representations. Most of the changed objects have regular edges and the changes of small objects are always strip-shaped. Therefore, we design an SSAM to capture local spatial features as well as to expand the perception of long range. The detailed structure is shown in Fig. 3, which can be divided into three stages.

In the first stage, the channel dimension information of the feature map is aggregated as F_{Max} and F_{Avg} using max pooling and average pooling, and then concatenated before being sent to the second stage

$$F_{\text{Max}} = \text{MaxPool}(F) \tag{8}$$

$$F_{\text{Avg}} = \text{AvgPool}(F) \tag{9}$$

where MaxPool and AvgPool denote the max pooling and average pooling operations along the channel dimension.

In the second stage, the first step is to utilize a 7×7 convolution to obtain the relation between each pixel and surrounding pixels and obtain local change details to generate F_S. The second step involves the utilization of strip pooling [42] to acquire extensive spatial correlations in both the horizontal and vertical directions. Due to its long and narrow kernel shape, strip pooling establishes long-range dependencies of discrete regions, focuses on capturing details and prevents interference from unrelated areas. Then, the current location and its neighbor feature by one-dimensional (1-D) convolution with a kernel size of 3. Finally, the horizontal feature and vertical feature are restored to the same size as the original feature map to obtain F_H and F_V by interpolation. The details are as follows:

$$F_S = f^{7\times7}([F_{\text{Max}}; F_{\text{Avg}}]) \tag{10}$$

where $f^{7\times7}$ denotes a 7×7 convolution and $[;]$ denotes a concatenate operation

$$F_H = I\left(f^{3\ 1D}(SP_H([F_{\text{Max}}; F_{\text{Avg}}]))\right) \tag{11}$$

$$F_V = I\left(f^{3\ 1D}(SP_V([F_{\text{Max}}; F_{\text{Avg}}]))\right) \tag{12}$$

where I denotes the interpolation operation; $f^{3\ 1D}$ denotes the convolution operation with a kernel size of 3 in 1-D space; SP_H denotes the pooling calculation in the horizontal direction; and SP_V denotes the pooling calculation in the vertical direction.

In the third stage, F_S, F_H, and F_V are concatenated and fused by a 2-D convolution of 1×1 size, and then the fusion feature

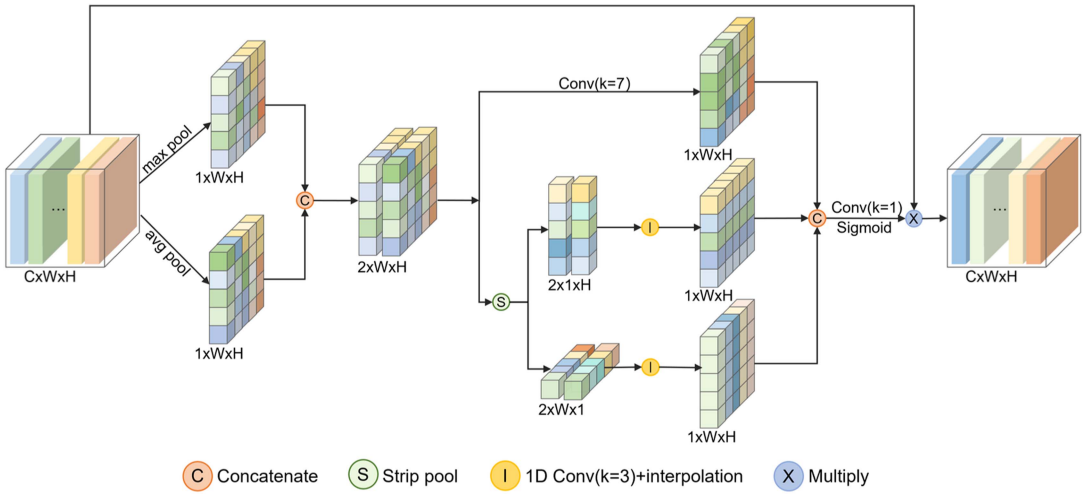


Fig. 3. SSAM. The module obtains two parts of features, one extracts long-range strip feature, and the other extracts local spatial features.

is activated using the Sigmoid function to obtain the weight matrix M. Finally, the M is elementwise multiplied by the input features to obtain the F_{SSAM} after strip spatial attention refinement, as follows:

$$M = \sigma\left(f^{1\times 1}\left([F_S; F_H; F_V]\right)\right) \tag{13}$$

$$F_{\text{ssam}} = M \otimes F \tag{14}$$

where $f^{1\times 1}$ denotes the 1×1 convolution in 2-D space; σ represents the Sigmoid function; and $\otimes$ denotes the pixel multiplication.

D. Loss Function

In the change detection task for remote sensing images, there is a huge quantitative difference between the unchanged pixels and the changed pixels, leading to a proportion imbalance of changed and unchanged areas. To weaken the effect of imbalance distribution of samples, a hybrid loss function is employed, which combines binary cross-entropy loss and dice coefficient loss.

Binary cross entropy [43] is usually chosen as the loss function of binary change detection, and its representation is shown as follows:

$$L_{\text{bce}} = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log p_n + (1-y_n)\log(1-p_n)\right] \tag{15}$$

where N denotes the total number of pixels; y_n denotes the ground truth value of the nth pixel; and p_n denotes the probability that the nth pixel is predicted to be changed.

Dice coefficient loss [44] is adopted to alleviate the effect of class imbalance, and its representation is shown as follows:

$$L_{\text{dice}} = 1 - \frac{2PY}{P+Y} \tag{16}$$

where P and Y denote the probability that all pixels are predicted to be in the change class and ground truth value, respectively.

The hybrid loss function is formulated as the summation of the binary cross-entropy loss and dice coefficient loss, as expressed in the following equation:

$$L = L_{\text{bce}} + L_{\text{dice}} \tag{17}$$

III. Experiment and Analysis

A. Datasets

We design a series of experiments on three change detection datasets CDD [45], LEVIR-CD [46], and S2Looking [47] to evaluate the effectiveness of FDFF-Net.

The CDD dataset is composed of seven pairs of seasonally varying remote sensing images, which include objects at different scales and seasonal variations in natural features. The size of these images is 4725×2700 pixels with a spatial resolution of 3–100 cm. To facilitate the deep-learning task, each image in CDD is cropped into 256×256 pixels and randomly rotated. The training, validation, and test sets consist of 10 000, 2998, and 3000 pairs of cropped images, respectively.

The LEVIR-CD dataset contains a total of 637 pairs of remote sensing images, which focus on building-related changes. The size of these images is 1024×1024 pixels with a spatial resolution of 0.5 m. To alleviate the computational burden on GPUs during training, each image is cropped into 256×256 pixels. Finally, the training, validation, and test sets consist of 7120, 1024, and 2048 pairs of cropped images, respectively.

The S2Looking dataset contains a total of 5000 pairs of side-looking satellite images captured at various off-nadir angles. The size of these images is 1024×1024 pixels with a spatial resolution of 0.5–0.8 m. Each image in this dataset is also cropped into 256×256 pixels without overlapping. The training, validation, and test sets consist of 56 000, 8000, and 16 000 pairs of cropped images, respectively.

Several data augmentation strategies are used for both datasets during the training stage to increase the diversity of the training data.

1) *Random flip:* Image pairs are flipped randomly, including horizontal and vertical flips, with a flip probability of 0.5.

2) *Random rotation:* Rotating the image pairs randomly with a rotation angle between −45° and 45° with a rotation probability of 0.4.
3) *Random fixed-angle rotation:* Image pairs are randomly rotated at a randomly selected angle among 90°, 180°, and 270° with a rotation probability of 0.7.
4) *Adding Gaussian noise:* Adding Gaussian noise to the image pairs at random with an additional probability of 0.3.

B. Implementation Details

FDFF-Net is implemented in the PyTorch framework and uses a single NVIDIA GeForce RTX 2080Ti GPU for training, validating, and testing. During the training stage, the Adam is used as an optimization algorithm, with the batch size of 10 and weight decay of 0.0005. We train the FDFF-Net for 200 epochs on the CDD and LEVIR-CD datasets, and 50 epochs on the S2Looking dataset. The initial learning rate is set as 0.0001, following the learning rate decay strategy that involves a reduction factor of 0.3 after every 30 epochs of training.

C. Evaluation Metrics

To quantitatively evaluate the effectiveness of the FDFF-Net, the precision (P), recall (R), $F1$-score ($F1$), intersection over union (IoU), and overall accuracy (OA) metrics are utilized. Precision measures the extent of false detection of changed pixels, with higher values indicating a lower occurrence of false detections. Recall measures the extent of missed detection of changed pixels, with higher values indicating fewer instances of missed detections. $F1$-score, as a composite metric combining precision and recall through harmonic averaging, presents a comprehensive evaluation of change detection. Intersection over union quantifies the spatial overlap between predicted and ground truth, and offers a focused assessment of the localization accuracy in change detection tasks. Additionally, the overall accuracy quantifies the proportion of correctly predicted pixels (both changed and unchanged) in relation to all the pixels.

D. Ablation Study

We design ablation experiments on CDD and LEVIR-CD datasets to verify the validity of the DDFM and SSAM modules. The baseline model is set up the same as the network proposed in this study except that the absolute value of the bitemporal features was directly used when extracting the difference features.

The accuracies of change detection are improved by using DDFM and SSAM, and the integration of these two modules maximizes the improvement in accuracy. The $F1$-score improves by 0.63% and 0.44%, and IoU improves by 1.19% and 0.75%, compared with the baseline model on the CDD and LEVIR-CD datasets, respectively (see Tables I and II). After replacing the module for extracting difference features in the baseline model with DDFM, the recall is improved by 0.30% and 0.64% on the two datasets, indicating that mining and fusing multiple difference is helpful to reduce missed detections. Furthermore, we conducted experiments on two datasets to analyze the effects of different dilation rates of the three convolution layers on the balance between the expansion of the receptive field and the preservation of local spatial information. The experimental results show that the optimal combination of the dilation rates of the second branch in DDFM is 1, 1, 2 (see Table III). We verify the effectiveness of DDFM in comprehensively mining differences between bitemporal features by visualizing the difference feature map (see Fig. 4). In Fig. 4, it can be seen that DDFM can detect more precise changes than the baseline model.

By adding SSAM to baseline model, the precision increases by 0.82% and 0.49% on the two datasets, respectively. It indicates that SSAM makes the network focus on changed area and reduces false detection. We also added only SAM on baseline model to verify the effectiveness of strip pool branch in SSAM (see Tables I and II). The results show that, compared with SAM, adding SSAM makes $F1$-score and IoU further improved.

TABLE I
CHANGE DETECTION ACCURACIES OF ABLATION STUDY ON CDD DATASET

Method	P	R	F1	IoU	OA
Baseline	0.9496	0.9867	0.9678	0.9376	0.9923
Baseline+DDFM	0.9502	**0.9897**	0.9696	0.9409	0.9926
Baseline+SSAM	0.9578	0.9835	0.9705	0.9427	0.9929
Baseline+SAM	0.9556	0.9850	0.9700	0.9419	0.9928
Baseline+SAM+DDFM	0.9611	0.9854	0.9731	0.9476	0.9936
Baseline+SSAM+DDFM	**0.9688**	0.9795	**0.9741**	**0.9495**	**0.9939**

The best performance highlighted in bold.

TABLE II
CHANGE DETECTION ACCURACIES OF ABLATION STUDY ON LEVIR-CD DATASET

Method	P	R	F1	IoU	OA
Baseline	0.9213	0.9051	0.9131	0.8401	0.9912
Baseline+DDFM	0.9224	**0.9115**	0.9169	0.8466	0.9915
Baseline+SSAM	0.9262	0.9025	0.9142	0.8420	0.9913
Baseline+SAM	0.9272	0.9006	0.9137	0.8411	0.9913
Baseline+SAM+DDFM	0.9269	0.9054	0.9160	0.8451	0.9915
Baseline+SSAM+DDFM	**0.9278**	0.9075	**0.9175**	**0.8476**	**0.9917**

The best performance highlighted in bold.

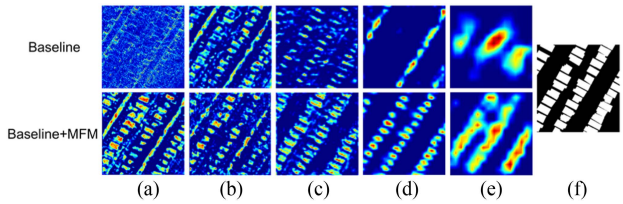


Fig. 4. Heat map comparison of the results of difference feature extraction. (a)–(e) Difference features obtained using different levels of features. (f) Ground truth. The top row shows the heat map for obtaining the difference features by subtracting the image features, while the bottom row shows the heat map for obtaining the difference features using DDFM.

TABLE III
ANALYSIS OF DILATION RATES OF THE SECOND BRANCH CONVOLUTIONAL LAYER OF DDFM

Dilation rate of convolution layers	CDD					LEVIR-CD				
	P	R	F1	IoU	OA	P	R	F1	IoU	OA
(1,1,1)	0.9597	**0.9857**	0.9725	0.9465	0.9934	0.9282	0.9044	0.9161	0.8453	0.9916
(1,1,2)	**0.9688**	0.9795	**0.9741**	**0.9495**	**0.9939**	0.9278	**0.9075**	**0.9175**	**0.8476**	**0.9917**
(1,2,2)	0.9619	0.9840	0.9728	0.9471	0.9935	0.9278	0.9065	0.9170	0.8468	0.9916
(1,2,3)	0.9598	0.9831	0.9713	0.9442	0.9931	**0.9321**	0.9007	0.9161	0.8452	0.9916

The best performance highlighted in bold.

TABLE IV
CHANGE DETECTION ACCURACIES OF ABLATION STUDY OF DIFFERENT SPATIAL ATTENTION MECHANISM

Method	CDD					LEVIR-CD				
	P	R	F1	IoU	OA	P	R	F1	IoU	OA
Baseline+SRU	0.9531	0.9858	0.9692	0.9402	0.9926	**0.9367**	0.8914	0.9135	0.8408	0.9912
Baseline+GLTB	0.9545	**0.9865**	0.9703	0.9422	0.9928	0.9293	0.8992	0.9140	0.8416	0.9913
Baseline+SSAM	**0.9578**	0.9835	**0.9705**	**0.9427**	**0.9929**	0.9262	**0.9025**	**0.9142**	**0.8420**	**0.9913**

The best performance highlighted in bold.

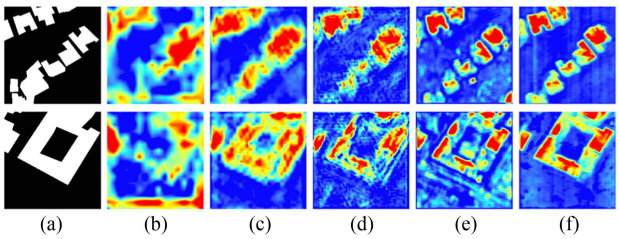


Fig. 5. SSAM output weight maps of difference feature extraction. (a) Ground truth. (b)–(f) Represent the weight maps output by SSAM at different levels of the decoder.

TABLE V
CHANGE DETECTION ACCURACIES OF ABLATION STUDY OF DEEP SUPERVISION

Dataset	Deep supervision	P	R	F1	IoU	OA
CDD	×	0.9668	0.9791	0.9729	0.9473	0.9936
	✓	**0.9688**	**0.9795**	**0.9741**	**0.9495**	**0.9939**
LRVIR-CD	×	**0.9289**	0.9028	0.9156	0.8444	0.9915
	✓	0.9278	**0.9075**	**0.9175**	**0.8476**	**0.9917**

The best performance highlighted in bold.

Moreover, we compared SSAM with the state-of-the-art (SOTA) spatial attention modules (global-local Transformer block (GLTB) [48] and spatial reconstruction unit (SRU) [49]) to demonstrate its superiority, and the results show that SSAM achieves higher $F1$-score and IoU on the two datasets (see Table IV). We confirm that SSAM concentrates on changed areas through the visualization of the output weight map (see Fig. 5). It can be seen that the brightness of the changed regions is higher than unchanged regions.

The FDFF-Net employs a deep supervision strategy [50] to alleviate the issue of gradient vanishing during network training. To evaluate the effectiveness of the deep supervision strategy, an ablation study is also conducted on two datasets separately to compare the network with this network without the deep supervision strategy (see Table V). Compared with the full model, $F1$-score of removing deep supervision from the network decreases by 0.12% and 0.19% on the two datasets, respectively. This indicates that training the intermediate layer of the network can further improve the accuracy of change detection.

E. Comparison

We compare our method with the SOTA methods, FC-Siam-diff [30], STANet [46], IFN [50], SNUNet-CD [51], AMAC [52], and ConvTransNet [53], to verify the advantages and effectiveness of FDFF-Net.

1) FC-Siam-diff [30] is based on the UNet network. This network inputs the absolute value of subtracted bitemporal features into the decoder to generate change maps.
2) STANet [46] incorporates a self-attention mechanism into the network. Spatial–temporal attention module is designed to capture spatial–temporal relations at long range to improve the ability to capture details.
3) IFN [50] uses both attention mechanisms and deep supervision strategies to improve the capability to detect changes. The utilization of an attention module facilitates the efficient fusion of multilevel features. Moreover, the implementation of a deep supervision strategy enhances the discriminative ability of the network to distinguish differences.

TABLE VI
CHANGE DETECTION ACCURACIES OF COMPARED METHODS ON CDD DATASET

Method	P	R	F1	IoU	OA
FC-Siam-diff	0.8290	0.6859	0.7506	0.6009	0.9462
STANet	0.8687	0.9412	0.9034	0.8240	0.9763
IFN	0.9496	0.8608	0.9030	0.8232	0.9791
SNUNet-CD/48	0.9629	0.9615	0.9622	0.9272	0.9924
AMCA*	0.9548	0.9308	0.9539	0.8915	0.9855
ConvTransNet*	**0.9759**	0.9464	0.9609	0.9248	0.9908
FDFF-Net	0.9688	**0.9795**	**0.9741**	**0.9495**	**0.9939**

* presents the results reported in the original article.
The best performance highlighted in bold.

TABLE VII
CHANGE DETECTION ACCURACIES OF COMPARED METHODS ON LEVIR-CD DATASET

Method	P	R	F1	IoU	OA
FC-Siam-diff	0.8754	0.7464	0.8058	0.6747	0.9817
STANet	0.8381	**0.9104**	0.8734	0.7742	0.9833
IFN	0.9233	0.8776	0.8998	0.8180	0.9901
SNUNet-CD/48	0.9067	0.8890	0.8978	0.8145	0.9896
AMCA*	0.9272	0.9067	0.9148	0.8464	0.9873
ConvTransNet*	0.9157	0.9027	0.9091	0.8334	0.9908
FDFF-Net	**0.9278**	0.9075	**0.9175**	**0.8476**	**0.9917**

* presents the results reported in the original article.
The best performance highlighted in bold.

4) SNUNet-CD [51] utilizes the dense skip connections of the UNnet++ architecture as its foundational framework. This network reduces the loss of deep information through dense connection and focuses on representative features by integrating the channel attention module. Here, SNUNet-CD with the best accuracy of 48 channels is selected for comparison.
5) AMCA [52] uses multiple attention modules to enhance and fuse multiscale contextual information. By effectively utilizing feature at different scales, the accuracy of change detection is improved.
6) ConvTransNet [53] combines the transformer and CNN in the encoder, which uses CNN branch and transformer branch to extract local and global features, respectively. In this way, the ability to detect multiple changed objects with various sizes is improved.

FDFF-Net achieves the best accuracy on CDD, LEVIR-CD, and S2Looking datasets (see Tables VI–VIII), as indicated by the higher $F1$-score, IoU, and OA than other methods. Compared with IFN using a pretrained VGG16 for feature extraction during the encoding stage, FDFF-Net achieves an improvement of 7.11% in $F1$-score and 12.63% in IoU on the CDD dataset, and improves by 1.77% in $F1$-score and 2.96% in IoU on the LEVIR-CD dataset. Furthermore, when evaluated on the S2Looking dataset, our method improves by 6.22% in $F1$-score and 6.80% in IoU. In contrast to SNUNet-CD/48, which uses a dense skip connection, FDFF-Net achieves the enhancement of 1.19% and 1.97% in $F1$-score on the CDD and LEVIR-CD datasets, and improves by 10.38% on the S2Looking dataset. Compared with AMAC, which introduces multiple attention modules, there is an enhancement of 2.02% and 0.27% in $F1$-score on the two datasets, respectively. In contrast to ConvTransNet, FDFF-Net achieves the enhancement of 1.32% and 0.84% in the $F1$-score on the CDD and LEVIR-CD datasets, respectively.

TABLE VIII
CHANGE DETECTION ACCURACIES OF COMPARED METHODS ON S2LOOKING DATASET

Method	P	R	F1	IoU	OA
FC-Siam-diff	0.7068	0.2937	0.4150	0.2618	0.9899
STANet	0.2531	**0.7913**	0.3835	0.2373	0.9652
IFN	0.6449	0.5771	0.6091	0.4379	0.9910
SNUNet-CD/48	0.5822	0.5546	0.5681	0.3967	0.9811
FDFF-Net	**0.7266**	0.6248	**0.6719**	**0.5059**	**0.9926**

The best performance highlighted in bold.

We conduct a visual comparative analysis between FDFF-Net and other methods to validate the advantages of our method in detecting small object change detection and changed object contours detection. The visualization comparison results show that FDFF-Net detects the changes of multiple ground objects while still retaining change details. For example, in Fig. 6, there are red boxes in the second, third, and fifth rows, with multiple small car changes. Compared with other methods, FDFF-Net can not only detect more complete small object changes but also identify the gap between the cars well. In the fourth row of Fig. 6, the detection of changes in the small trees within the red box is exclusively accomplished by our method, which due to FDFF-Net mines the multiple differences between bitemporal images, enabling the precise identification of changes in various objects.

FDFF-Net detects the contours of the changes more clearly than other methods. In Fig. 7, the two fences within the red boxes labeled in the third and fourth rows are detected completely without overlap, and the results in the fifth row are close to SNUNet-CD/48, but the fine fence contours in the red boxes are captured and detected. This indicates that the attention mechanism used by FDFF-Net enhances the perception of change details and that the use of SSAM makes the network accurate in the detection of strip-shaped objects. In Fig. 8, all five methods have good detection effects, but FDFF-Net has obvious advantages for the detection of building edges. The building edge details, such as the red boxes in the first, fourth, and fifth rows, are marked with closely connected buildings and small gaps. The other four methods are unable to detect such change details, while FDFF-Net can detect them.

IV. DISCUSSION

A. Advantage of FDFF-Net for Alleviating Pseudochanges

The precision of change detection is always influenced by the pseudochanges caused by seasonal variation, shadows, and

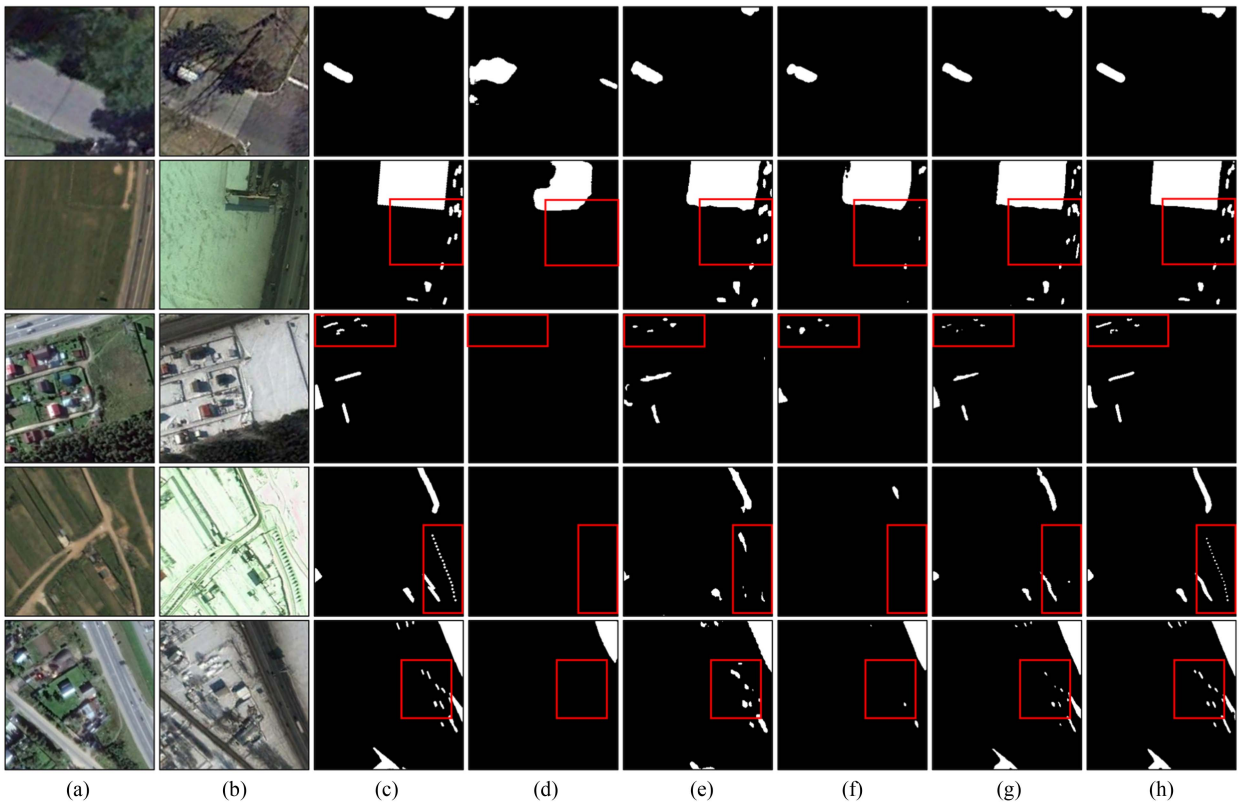


Fig. 6. Visualization comparison of different methods for change detection on CDD dataset, which shows our superiority on detecting changes of small objects. (a) Pretemporal images. (b) Posttemporal images. (c) Ground truth. (d) FC-Siam-diff. (e) STANet. (f) IFN. (g) SNUNet-CD/48. (h) FDFF-Net.

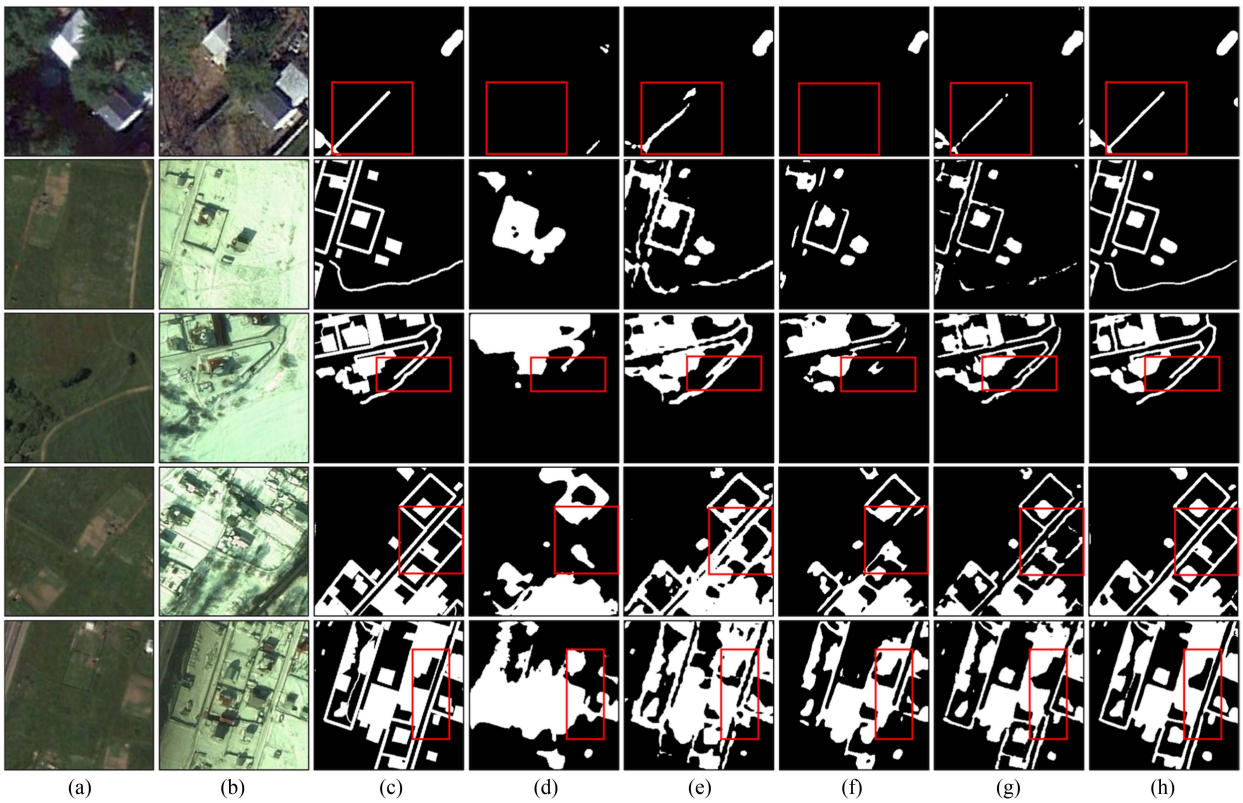


Fig. 7. Visualization comparison of different methods for change detection on CDD dataset, which shows our superiority on detecting contours of changed objects. (a) Pretemporal images. (b) Posttemporal images. (c) Ground truth. (d) FC-Siam-diff. (e) STANet. (f) IFN. (g) SNUNet-CD/48. (h) FDFF-Net.

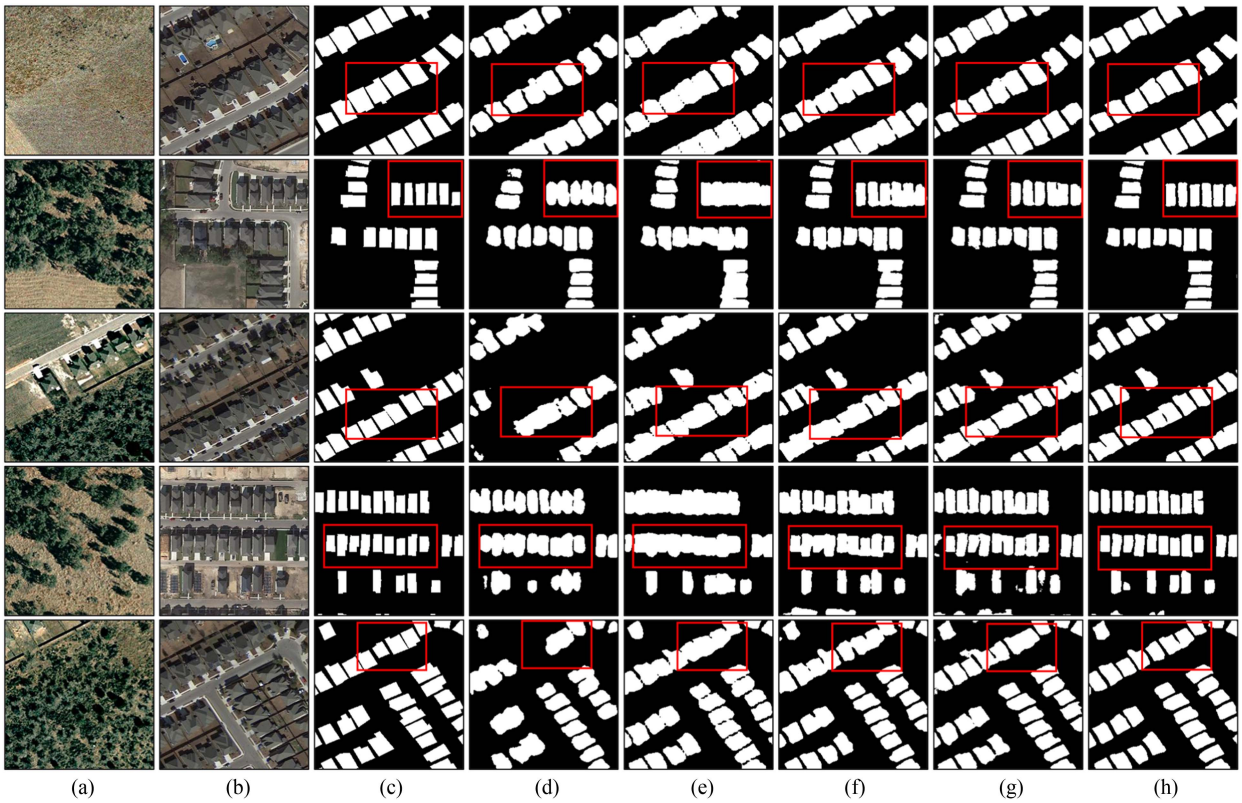


Fig. 8. Visualization comparison of different methods for change detection on the LEVIR-CD dataset, which shows our superiority on detecting contours of changed objects. (a) Pretemporal images. (b) Posttemporal images. (c) Ground truth. (d) FC-Siam-diff. (e) STANet. (f) IFN. (g) SNUNet-CD/48. (h) FDFF-Net.

viewing angles' differences [54]. Two effective strategies for alleviating pseudochanges involve expanding the receptive field of the convolution layer and leveraging the deep features of changed ground objects. For instance, the positional bias of the same object in two images is caused by different satellite imaging angles. By expanding the receptive field beyond the object, positional bias can be captured, preventing the network from erroneously flagging nonoverlapping areas as changed. Similarly, within the deep feature map, the pixel count of changed ground objects is reduced, minimizing positional bias between the same objects in the two images, which makes it easier for the receptive field of the convolutional layer to cover all pixels. We employed both two approaches to alleviate pseudochanges. First, we designed DDFM to expand the receptive field during the process of extracting difference features, making the FDFF-Net obtain the local feature of different ranges. Second, we used a full-scale skip connection methodology to merge deep features with shallow features, providing a continuous refinement to change detection. Experimental results show that FDFF-Net can alleviate these pseudochanges. For example, in Fig. 6, the snow cover on the grassland caused by seasonal variation is not classified as change, and the shadows of the trees are also not identified as changes. Furthermore, in contrast to other methods, FDFF-Net achieves the best $F1$-score and IoU on the S2Looking dataset, whose images are captured at various off-nadir angles (see Table VIII).

Although the experiments prove that our method has advantages in alleviating pseudochanges, there exists potential for refinement. This is attributed to the complexity of remote sensing images, where multiple ground objects coexist, with varying dimensions for even identical objects. Consequently, a static enhancement of the receptive field might fall short of comprehensively addressing the complex scenes of remote sensing images.

B. Limitations and Prospects

Although FDFF-Net has achieved impressive performance, it still has limitations. In the decoding stage, FDFF-Net utilizes the multilevel features extracted by the encoder to supplement the lost change information due to downsampling. However, compared with the deep feature, the shallow feature contains more noise due to fewer convolution operations, which will interfere accuracies of the detection results [55]. Therefore, there is a need to optimize the backbone of feature extraction to reduce noise in shallow feature.

Due to the limited receptive field of convolution, semantic information will be lost during the feature extraction process, and this loss will increase as network layers get deeper. The advantage of transformers in extracting global feature may solve this problem [56], [57]. It is assumed that combining the CNN, which extracts local feature, and transformer, which extracts global feature, as the backbone network of feature extraction can improve the completeness of changed objects.

In addition to the use of transformers to extract image features in the encoding stage, single-temporal image features can also be

decoded to reconstruct features of objects and eliminate noise. Multitask learning framework for change detection [21], [58] has an extra dual decoder to obtain segmentation results of images, which offers semantic constraints for change detection. The network with multitask architecture can not only alleviate the noise and improve the robustness of the network but also further explore the application in "from–to" change detection.

V. Conclusion

In this study, we propose an FDFF-Net for change detection. Our method incorporates two key components, namely the DDFM and the SSAM. The DDFM is designed to mine multiple differences between bitemporal images and the SSAM is designed to enhance change features. Aggregating DDFM and SSAM shows superiority in the detection of change details. Extensive evaluations conducted on three datasets, CDD, LEVIR-CD, and S2Looking, substantiate that FDFF-Net outperforms the compared SOTA methods and is effective in detecting change details. Furthermore, in comparison with other methods, FDFF-Net demonstrates enhanced capability in alleviating the pseudochanges caused by shadows, viewing angles' differences, and seasonal variation.

References

[1] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 609–630, Mar. 2013.

[2] D. Wen, X. Huang, L. Zhang, and J. A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 609–625, Jan. 2016.

[3] H.-J. Gim, C.-H. Ho, S. Jeong, J. Kim, S. Feng, and M. J. Hayes, "Improved mapping and change detection of the start of the crop growing season in the US Corn Belt from long-term AVHRR NDVI," *Agricultural Forest Meteorol.*, vol. 294, Nov. 2020, Art. no. 108143.

[4] S. Berberoglu and A. Akin, "Assessing different remote sensing techniques to detect land use/cover changes in the eastern Mediterranean," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 11, no. 1, pp. 46–53, Feb. 2009.

[5] S. Ye, J. Rogan, Z. Zhu, and J. R. Eastman, "A near-real-time approach for monitoring forest disturbance using Landsat time series: Stochastic continuous change detection," *Remote Sens. Environ.*, vol. 252, Jan. 2021, Art. no. 112167.

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

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

[8] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[9] J. Zabalza et al., "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," *Neurocomputing*, vol. 185, pp. 1–10, Apr. 2016.

[10] M. Gong, T. Zhan, P. Zhang, and Q. Miao, "Superpixel-based difference representation learning for change detection in multispectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2658–2673, May 2017.

[11] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

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

[13] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.

[14] K. Nemoto, R. Hamaguchi, M. Sato, A. Fujita, T. Imaizumi, and S. Hikosaka, "Building change detection via a combination of CNNs using only RGB aerial imageries," *Proc. SPIE*, vol. 10431, Oct. 2017, Art. no. 104310J.

[15] A. M. El Amin, Q. Liu, and Y. Wang, "Zoom out CNNs features for optical remote sensing change detection," in *Proc. 2nd Int. Conf. Image, Vis. Comput.*, 2017, pp. 812–817.

[16] S. T. Seydi, M. Hasanlou, and M. Amani, "A new end-to-end multi-dimensional CNN framework for land cover/land use change detection in multi-source remote sensing datasets," *Remote Sens.*, vol. 12, no. 12, Jan. 2020, Art. no. 2010.

[17] H. Zhang, X. Tang, X. Han, J. Ma, X. Zhang, and L. Jiao, "High-resolution remote sensing images change detection with Siamese holistically-guided FCN," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4340–4343.

[18] 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.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. New York, NY, USA: Springer, 2015, pp. 234–241.

[20] L. Li, C. Wang, H. Zhang, B. Zhang, and F. Wu, "Urban building change detection in SAR images using combined differential image and residual U-net network," *Remote Sens.*, vol. 11, no. 9, Jan. 2019, Art. no. 1091.

[21] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, Sep. 2021.

[22] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[23] E. Bousias Alexakis and C. Armenakis, "Evaluation of UNet and UNet++ architectures in high resolution image change detection applications," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 43B3, pp. 1507–1514, Aug. 2020.

[24] A. Raza, H. Huo, and T. Fang, "EUNet-CD: Efficient UNet++ for change detection of very high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 3510805.

[25] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.

[26] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 1055–1059.

[27] Z. Zheng, Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang, "CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 247–267, May 2021.

[28] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "ADS-Net: An attention-based deeply supervised network for remote sensing image change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, Sep. 2021, Art. no. 102348.

[29] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "ChangeNet: A deep learning architecture for visual change detection," in *Proc. Eur. Conf. Comput. Vis.*, 2019, vol. 11130, pp. 129–145.

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

[31] X. Yang, L. Hu, Y. Zhang, and Y. Li, "MRA-SNet: Siamese networks of multiscale residual and attention for change detection in high-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 22, Jan. 2021, Art. no. 4528.

[32] X. Zhang et al., "DifUnet++: A satellite images change detection network based on UNet++ and differential pyramid," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8006605.

[33] S. Ghaffarian, J. Valente, M. van der Voort, and B. Tekinerdogan, "Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review," *Remote Sens.*, vol. 13, no. 15, Jan. 2021, Art. no. 2965.

[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[35] J. Huang, Q. Shen, M. Wang, and M. Yang, "Multiple attention Siamese network for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406216.

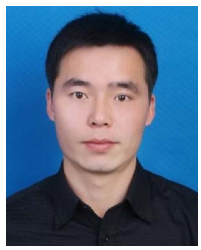
[36] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," in *Proc. Int. Conf. Learn. Representations*, Mar. 2020.
[37] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 207–210.
[38] Q. Ding, Z. Shao, X. Huang, and O. Altan, "DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102591.
[39] H. Cheng, H. Wu, J. Zheng, K. Qi, and W. Liu, "A hierarchical self-attention augmented Laplacian pyramid expanding network for change detection in high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 182, pp. 52–66, Dec. 2021.
[40] R. Song, W. Ni, W. Cheng, and X. Wang, "CSANet: Cross-temporal interaction symmetric attention network for hyperspectral image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 2022, Art. no. 6010105.
[41] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 636–644.
[42] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4002–4011.
[43] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol.*, 2020, pp. 1–7.
[44] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 240–248.
[45] 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. XLII-2, pp. 565–571, May 2018.
[46] 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, Jan. 2020, Art. no. 1662.
[47] L. Shen et al., "S2Looking: A satellite side-looking dataset for building change detection," *Remote Sens.*, vol. 13, no. 24, Dec. 2021, Art. no. 5094.
[48] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.
[49] J. Li, Y. Wen, and L. He, "SCConv: Spatial and channel reconstruction convolution for feature redundancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6153–6162.
[50] 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.
[51] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8007805.
[52] X. Xu, Z. Yang, and J. Li, "AMCA: Attention-guided multiscale context aggregation network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5908619.
[53] W. Li, L. Xue, X. Wang, and G. Li, "ConvTransNet: A CNN–transformer network for change detection with multiscale global–local representations," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5610315.
[54] X. Hou, Y. Bai, Y. Li, C. Shang, and Q. Shen, "High-resolution triplet network with dynamic multiscale feature for change detection on satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 103–115, Jul. 2021.
[55] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 101–119, May 2022.
[56] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.
[57] S. Chu, P. Li, M. Xia, H. Lin, M. Qian, and Y. Zhang, "DBFGAN: Dual branch feature guided aggregation network for remote sensing image," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, Feb. 2023, Art. no. 103141.
[58] Q. Shen, J. Huang, M. Wang, S. Tao, R. Yang, and X. Zhang, "Semantic feature-constrained multitask Siamese network for building change detection in high-spatial-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 189, pp. 78–94, Jul. 2022.

Feng Gu received the B.S. degree in remote sensing science and technology from the China University of Mining and Technology, Beijing, China, in 2022. He is currently working toward the M.S. degree in remote sensing of resources and environment with Nanjing University, Nanjing, China.

His research interests include change detection and deep learning for remote sensing.

Pengfeng Xiao (Senior Member, IEEE) was born in Hunan, China, in 1979. He received the B.M. degree in land resource management from Hunan Normal University, Changsha, China, in 2002, and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2007.

From 2007 to 2009, he was a Lecturer with the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor from 2010 to 2018. Since 2019, he has been a Professor with Nanjing University. He was a Visiting Scholar with the Department of Geography, University of Giessen, Germany, from 2011 to 2012, and the Department of Environmental Science, Policy, and Management, University of California at Berkeley, USA, from 2014 to 2015. He has authored four books and more than 100 articles. His current research interests include high-resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover change.



Xueliang Zhang (Member, IEEE) received the B.S. degree in geographical information system and the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2010 and 2015, respectively.

From 2014 to 2015, he was a Visiting Student with Informatics Institute, University of Missouri, Columbia, USA. From 2016 to 2018, he was an Associate Researcher with the Department of Geographic Information Science, Nanjing University. He is currently an Associate Professor with the Department of Geographic Information Science, Nanjing University. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.

Zhenshi Li received the B.S. degree in geographic information science from Hohai University, Nanjing, China, in 2019, and the M.S degree in cartography and geographic information system in 2022 from Nanjing University, Nanjing, China, where he is currently working toward the Ph.D. degree in cartography and geographic information system.

His research interests include semantic segmentation, weakly supervised deep learning, and intelligent interpretation for remote sensing.

Dilxat Muhtar received the B.S. degree in geographic information science in 2022 from Nanjing University, Nanjing, China, where he is currently working toward the M.S. degree in cartography and geographic information system.

His research interests primarily revolve around self-supervised and transfer learning for remote sensing.