AFPF-Net: Adjacent-Level Feature Progressive Fusion Full Convolutional Network for Remote Sensing Change Detection

Wei Wang, Luocheng Xia, and Xin Wang

***Abstract*—In recent years, convolutional neural networks have achieved good results in the field of change detection (CD) owing to their exceptional feature extraction capabilities. However, accurately detecting objects with completely changing details, given the complex imaging conditions of bitemporal images, remains a formidable challenge. Aiming at the above challenge, we have designed a new method for remote sensing image CD. First, to capture the fine difference features at different scales, the feature difference enhancement module is proposed to enhance the information interactions not only among the bitemporal features but also between the difference features of the previous layer and the rough difference map of the current layer. Second, to accurately capture the entire region of change, the adjacent-level feature progressive fusion module is proposed, which extracts complementary information by progressively fusing high-level and low-level features, therefore enhancing the change features. Finally, based on the above two modules, a full convolution-based adjacent-level feature progressive fusion network (AFPF-Net) is designed. To validate the effectiveness of AFPF-Net, experimental evaluations are performed on two different datasets, the LEVIR-CD and WHU-CD datasets. Compared to the sub-optimal network in the experiments, the F1-score on these two datasets improved by 0.33% and 1.74%, and total model complexity is relatively reduced, achieving better balance between model performance and complexity compared to the experimental state-of-the-art network.**

***Index Terms*—Adjacent-level feature progressive fusion (AFPF), change detection (CD), convolutional neural network (CNN), feature difference enhancement (FDE), remote sensing image.**

I. Introduction

CHANGE detection (CD) is the process of identifying differences in bitemporal images captured over different periods in the same region. With the increasing ability to acquire high-resolution remote sensing images and the progression of imaging technology, CD has received an increasing amount of attention and has gained applications in many aspects of real life, such as land development CD [1], [2], global resource monitoring [3], urban management [4], and damage assessment [5], [6].

In recent years, deep learning (DL) has demonstrated remarkable progress in computer vision applications, including object detection, image segmentation, and image classification, owing to the rapid improvement in convolutional neural networks (CNNs) [9], [10], [11]. DL-based methods not only significantly reduce the need for manual intervention compared to the conventional hand-crafted feature approach, but they also mitigate the occurrence of errors that may arise from data preprocessing [12]. In addition, as the importance of CNN in image processing has grown, an increasing number of entirely convolutional-based CD networks have emerged [13], [14], [15]. According to the strategy of bitemporal image feature fusion, full convolution-based CD methods can be categorized into early fusion and late fusion [16]. In early fusion methods, the bitemporal images are used as inputs to the CD network after concatenating or differencing. For example, Peng et al. [17] designed a difference-enhanced dense-attention convolutional network that concatenated pairs of bitemporal images as inputs to the networks, from which accurate change features were extracted. Before fusing the temporal features, the later fusion approach extracts the bitemporal image features separately using the Siamese network. For example, Daudt et al. [15] applied the Siamese network before stitching and then performed the fusion of the temporal features.

Despite previous CD methods having made tremendous advancements, detecting changing objects with complete change details is still a challenging task, as shown in Fig. 1. Recent approaches have begun to combine concatenation and difference operations to perform bitemporal feature fusion [18], [19]. However, the captured temporal difference information still contains a substantial quantity of difficult to distinguish "nonsemantic changes" [20], such as those caused by car motion, sensor noise, or human subjective factors. In addition to bitemporal feature fusion multiscale feature fusion [21], [22] is also a crucial component in CD. In general, low-level features possess detailed spatial information but lack comprehensive semantic information, whereas high-level features possess detailed semantic information but are devoid of fine-grained or boundary information. Multiscale feature fusion can integrate the complementary information between them. However, since

Manuscript received 27 February 2024; revised 10 April 2024 and 11 May 2024; accepted 2 August 2024. Date of publication 6 August 2024; date of current version 15 August 2024. This work was supported in part by the National Science Innovation Special Zone Project under Grant 2019XXX00701, in part by the National Key Basic Research Program Project under Grant 624XXXX0206, in part by the Key research and development projects of Hunan Province under Grant 2020SK2134, and in part by the Natural Science Foundation of Hunan Province under Grant 2019JJ80105, and Grant 2022JJ30625. *(Corresponding author: Xin Wang.)*

The authors are with the School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China (e-mail: wangxin@csust.edu.cn).

Our code will be available at https://github.com/is8xia/AFPF-Net/.

Digital Object Identifier 10.1109/JSTARS.2024.3439340

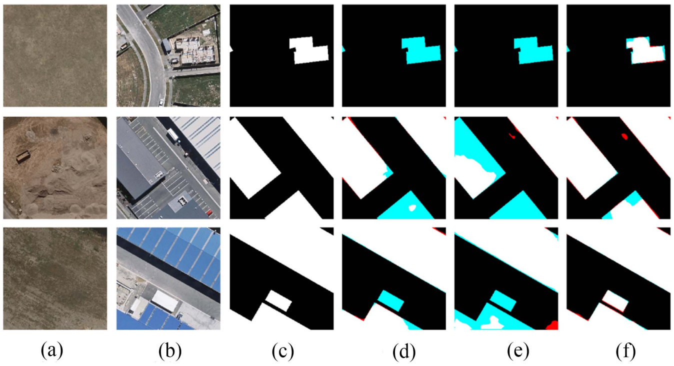


Fig. 1. Visualization results of different methods. (a) T1 image. (b) T2 image. (c) Label. (d) AFCF3D-Net [7]. (e) DMINet [8]. (f) AFPF-Net. TP, TN, FN, and FP are denoted as white, black, blue, and red. These methods cannot accurately detect the complete change object, but AFPF-Net obtains more accurate change detection results.

the semantic difference exists between features at low and high levels, the noise of low-level features and the rough boundaries of high-level features may interfere with the direct fusion of change information, making it difficult to accurately detect changing objects with complete change details.

Therefore, we propose a new adjacent-level feature progressive fusion fully convolutional network (AFPF-Net), which can achieve better detection results by enhancing the information exchange between the bi-temporal features, and the adjacent-layer feature fusion. In AFPF-Net, after extracting multiscale temporal features from the Siamese CNN, the number of channels for each scale feature is reduced from {64, 128, 256, 512} to {64, 64, 64, 64} by using channel reduction (CR). Then, the feature difference enhancement (FDE) module is used to combine the feature concatenation, element-wise subtraction, and the upper layer of variation features to extract the fine difference information. After that, the adjacent-level feature progressive fusion (AFPF) module is designed to acquire and integrate the supplementary information in the features with different scales to enhance and improve different aspects of the changing objects. The main contributions are as follows.

1) To enhance the details of change regions and extract reliable change information fully, the FDE module is designed. After extracting and fusing bitemporal features, the residual connection is used to aggregate the information from the previous layer difference to enhance the discrepancy among the current bitemporal features and capture more accurate change features.
2) To address the incompleteness of change areas and the noise interference resulting from the direct fusion of different scale features, the AFPF module is designed. Through the knowledge review branch and the boundary compensating branch, complementary information between the cross-layers is used to improve the change features. The enhanced features have more complete change regions and smoother boundaries.
3) Based on the above two modules, a fully convolutional network, AFPF-Net, for remote sensing change detection is designed. Among the state-of-the-art (SOTA) methods that are already compared, superior detection results are achieved compared to pure transformer networks or networks with a combination of transformer and convolution.
4) A set of experiments is carried out on the LEVIR-CD and WHU-CD datasets in order to assess the effectiveness of AFPF-Net. The F1-score on these two datasets is improved by 0.33% and 1.74%, respectively, compared to the suboptimal network.

The rest of this article is organized as follows. Section II presents DL approaches for CD methods, including both non-transformer and transformer-based techniques. Section III details the design concept of the method. Section IV summarizes and analyzes the results of the experiment. Finally, Section V concludes this article.

II. Related Work

Since the proposed network relies on the CNN, in this section, the CD method based on DL is mainly introduced. Recently, the DL approach has become the main solution for CD of remote sensing images because of its excellent feature expression capability. In terms of fusion strategies for bitemporal images, DL-based approaches can be simply categorized into early fusion and late fusion [16]. Early fusion methods use the image after concatenation or difference operation of the bitemporal image as the input. Late fusion approaches use a Siamese network architecture to separately retrieve the characteristics of the bitemporal images.

A. CNN-Based Model

To cope with uncorrelated change regions and extract more accurate change regions, a number of recent studies have improved the generalization capability of the network through the perspectives of bitemporal information fusion and multiscale feature aggregation, which are the key components for extracting changes from dual-temporal remotely sensed images.

The dual-temporal information fusion strategy [23], [24], [25], as an important part of remote sensing change detection, can provide reliable change information and enhance the details of the change region. In the process of fusing bitemporal information, researchers have widely used difference or splice operations to extract and fuse features. To fully extract the change information between bitemporal images, Zhu et al. [23] proposed a feature comparison module to capture feature difference maps at different scales. In addition, the number of channels for features at different scales is unified so that the contribution of each feature is the same, which reduces the information loss in the fusion process. However, extracting the change information between pairs of features at different scales separately still makes it difficult to accurately capture the complex change information in different features. Therefore, Lei et al. [24] proposed a difference enhancement module, which subtracted the bitemporal features to obtain the difference feature map, then utilizes the attention mechanism to capture the weights of the real changed regions and map them back to the original features. Thus, it serves to guide for the next layer of feature extraction. Zhong and Wu [25] proposed a novel network (T-UNet) based on a three-branch encoder, and to allow features from different branches to interact and fuse effectively, a multibranch spatial domain cross-attention module was proposed. This module allows full interaction of the difference feature maps at different scales, suppresses pseudochanges induced by the difference computation or noise, and completes the real changes that may be missing in the high-level feature information.

In addition, the multiscale feature fusion strategies [18], [26], [27], [28], [29], also serve as an important components in remote sensing change detection, providing both high-level features rich in semantic information and low-level features rich in fine-grained information. In order to fully utilize the information of different scale features, the deeply supervised image fusion network designed by Zhang et al. [26] fused multiscale features which were taken out of a two-stream architecture using a Siamese network and then employed in a disparity discrimination network for CD. To extract accurate change maps, a convolutional block attention module [27] was used in the disparity discrimination network. To meet the demand for fusing multiscale features, Fang et al. [28] proposed a NestedUNet-based densely connected Siamese network (SNUNet) for fusing multiscale features. Liu et al. [29] proposed a very lightweight Siamese network based on SNUNet, which eliminated duplicate connections at the cost of losing a small amount of accuracy and greatly reduced the number of parameters. With further research, Wang et al. [18] proposed a deeply supervised network (ADS-Net) based on an attention mechanism using intermediate layer fusion. After feature extraction process from each layer, the bitemporal features obtained during the encoding phase were connected to the result of the preceding layer in the decoding section. Additionally, the differential feature maps were connected to the bitemporal feature maps and used as inputs to each decoding layer. To further strengthen the intrinsic connection between the temporal features at each level and to capture more representative change features, Li et al. [19] proposed a guided progressive refinement model. Initially, different scale change features are aggregated, and then the fused features are used to iteratively refine the multiscale features, so that the pseudochange information in the low-level features is filtered out and the rough boundary in the high-level features is further polished. Wang et al. [30] proposed a new spatial–spectral cross-fusion network for a remote sensing image change detection model. By misplacing and reorganizing temporal features at different levels in the channel domain, not only the semantic differences between different features are reduced, but also the semantic information in each feature is enhanced. Li et al. [31] introduced an online uncertainty estimation branch, compelling the network to allocate more attention toward the actual area of change. A knowledge review strategy was also introduced to increase the distinguishability of the different features by continuously learning the parts where conflicts between low-level and high-level features occur.

Furthermore, to capture more precise regions of change, several approaches have been identified through extensive research, such as fusion of multiscale features across layers [32], 3-D convolution [7], and attentional mechanisms [33], [34], [35], [36], [37]. It is worth noting that various forms of attentional mechanisms have emerged as they have become increasingly influential in computer vision. For example, Zheng et al. [33] utilized spatial, channel squeezing, and channel excitation modules to recalibrate the space and channels so that the network focuses on more useful features. Eftekhari et al. [34] proposed a parallel spatial channel attention mechanism to learn the details of changes more stably and achieve more accurate CD results compared to serial. Chen et al. [35] proposed a biattention fully convolutional Siamese network (DASNet). This network addressed the issue of an unbalanced penalty between changed and unchanged feature pairs by weighting the dual-edge contrast loss. Additionally, the network utilized a biattention mechanism to accurately identify the change regions, resulting in enhanced model performance. The attention mechanisms are becoming increasingly popular in the CD domain due to the significant improvement of the attention mechanisms for network performance and the presence of plug-and-play properties.

However, the information interaction among bi-temporal features and the extraction of complementary information among features with changes at different scales are still not fully investigated in the current research. Therefore, we design a module to extract features efficiently and introduce another module to realize feature complementation.

B. *Transformer-Based Model*

Transformer [38] has recently made significant progress in the domain of natural language processing. Due to its powerful representation capabilities, researchers have applied it to computer vision tasks in areas, such as semantic segmentation [39], [40], target detection [41], and image classification [42]. Subsequently, the transformer has also started to be used for CD tasks. For example, Chen et al. [43] proposed a bitemporal image transformer (BITNet), in which the context in the bitemporal images is better modeled by the transformer encoder to eliminate irrelevant changes and distinguish relevant changes. Feng et al. [44] proposed the intrascale cross-interaction and interscale feature fusion network (ICIF-Net). By leveraging the efficiency of CNNs in extracting local features and the capabilities of Transformer in global modeling, parallel processing together provides change targets and fine-grained details in CD from remotely sensed images. Mao et al. [45] proposed a transformer-based multiscale feature fusion network, which utilized a transformer to capture correlations between changing regions and other regions over long distances and then aggregated these features for a stronger semantic and localized representation. Although the introduction of transformers has improved the performance of CD networks, it has also added a large number of parameters. Therefore, AFPF-Net adopts a fully convolutional structure, which reduces the number of parameters while also surpassing other compared SOTA methods.

III. Methodology

This section first introduces the overall structure and workflow of AFPF-Net, then provides a comprehensive description of the

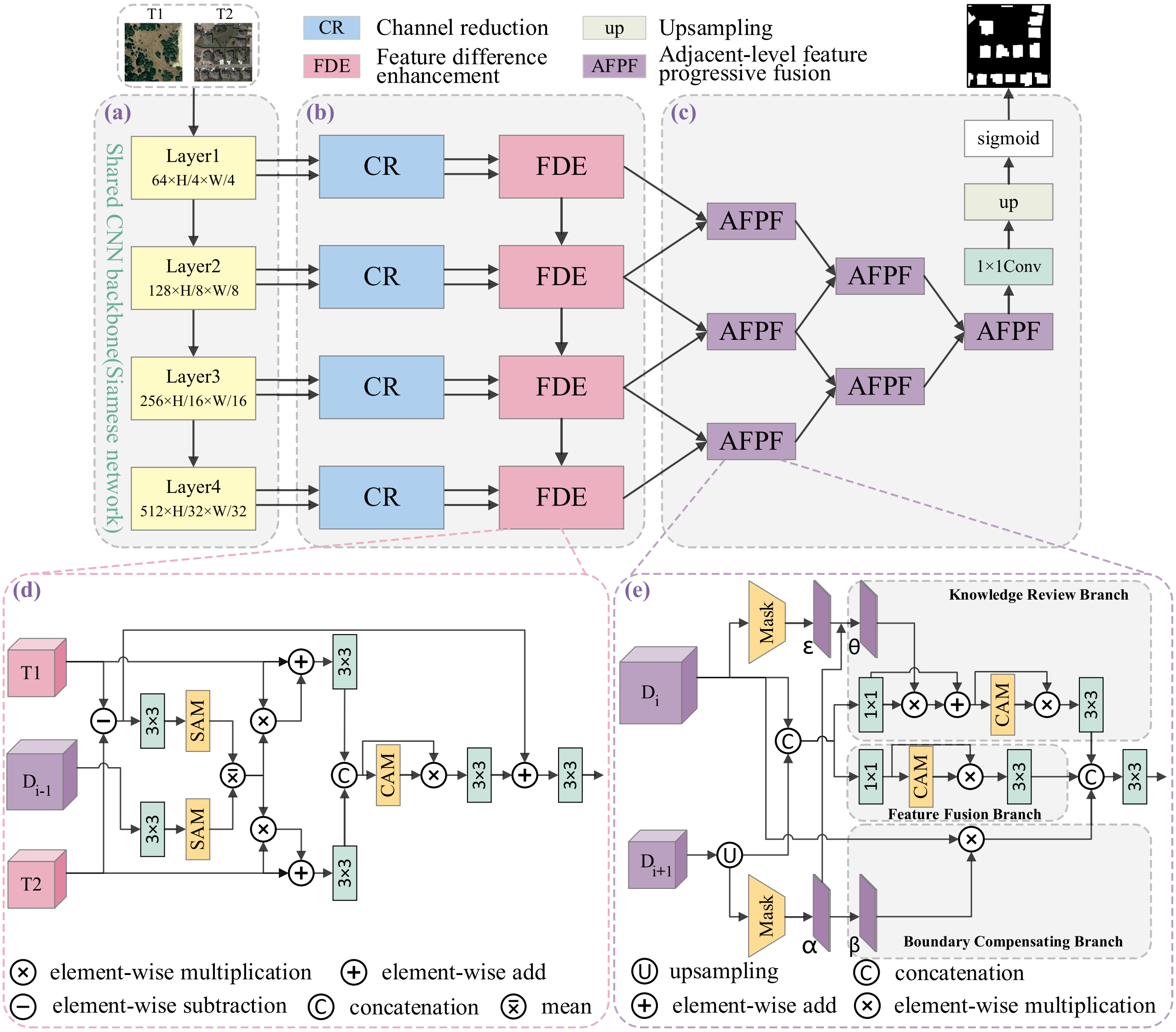


Fig. 2. Overall structure of AFPF-net. Initially, the temporal features are captured through a weight-sharing backbone network, followed by channel reduction through CRs. The FDEs are responsible for capturing temporal features at multiple levels, and finally, the final change map is generated by progressively fusing temporal difference features through the AFPFs.

FDE module and AFPF module, and finally provides the hybrid loss function.

A. Overall Structure

The whole structure of AFPF-Net is shown in Fig. 2, which includes four modules: the feature extraction module (Backbone), the CR module, the feature difference enhancement (FDE) module, and the AFPF module. To capture more accurate feature difference maps between bitemporal feature pairs, the FDE is proposed to capture change features. To overcome the semantic differences between different scale features and to extract complementary information between low-level and high-level features, the AFPF module is proposed to fuse different scale features. The network utilizes the pretrained ResNet18 [46] as the Backbone. A Siamese network with shared weights is used to capture features from the bitemporal image pairs T1 and T2. The extracted multiscale features may be denoted as F_i^1 and F_i^2, i $\in \{1, 2, 3, 4\}$. Then, the number of extracted multiscale feature channels is unified to 64 using the CR module. Next, feature extraction and fusion are performed on the bitemporal feature pairs by FDE, and residual connections are used to aggregate the difference information of the previous layer to obtain a fine difference feature. Finally, the feature difference maps at various scales are input into AFPF, and the complementary information between neighboring layers is fused to produce the final prediction maps.

B. Feature Difference Enhancement (FDE) Module

Since rich change information is included in the low-level feature difference map, FDE first extracts the variation information from it. However, the low-level feature difference map also contains pseudochanges, so the residual connection is used in

FDE to aggregate the previous layer variance information, which not only enhances the difference of the bitemporal features in the current layer but also removes some of the pseudochanges present in the variance information of the previous layer. This is due to the receptive field is relatively large in the deep network, and the extracted high-level feature difference maps have a lower proportion of pseudochanges than low-level feature difference maps. The weight of pixels in pseudochange areas is significantly reduced, and the weight of pixels in the region of the real changes is relatively increased when the weights of the two difference maps are averaged. The function of FDE is closely related to the above two points, and its detailed design is illustrated in Fig. 2(d).

Specifically, the different scales of features extracted from the Backbone are initially processed via the CR module in Fig. 2(b), which unifies the channel number of different scales of features to 64, which reduces the use of memory and computation for subsequent operations. As shown in Fig. 2(d), the CR-processed features are fed into the FDE module, whose inputs are composed of three parts: F_i^1, F_i^2, and the difference features after enhancement of the previous layer. When $i = 1$, only the lowest level temporal feature pairs are used as inputs, and when i $\in \{2, 3, 4\}$, the inputs are all composed of three parts. The rough feature difference map D_i^r is first computed in the FDE using element-wise subtraction, followed by an absolution operation. Then the difference feature extraction is performed by 3 × 3 convolution, and the attention map $\hat{D}_i$ is obtained by the spatial attention module (SAM) [27]. This process can be represented as follows:

$$D_i^r = \text{Conv}_{3\times3}\left(\left|F_i^1 \ominus F_i^2\right|\right) \tag{1}$$

$$\hat{D}_i = \text{SAM}\left(D_i^r\right) \tag{2}$$

where $\ominus$ denotes an element-wise subtraction operation, $|\cdot|$ is an absolution operation, $\text{Conv}_{3\times3}(\cdot)$ represents a 3 × 3 convolutional layer, a batch normalization (BN), and a ReLu function. The subsequent convolutional structure is consistent with that here. Meanwhile, the change feature D_{i-1}^r of the previous layer is downsampled by a 3 × 3 convolution to make its spatial dimension consistent with F_i^1, after which the change attention map $\hat{D}_{i-1}$ is obtained by SAM. The refined change attention map is obtained by averaging the weights of the two change attention maps $\hat{D}_i$ and $\hat{D}_{i-1}$. This process can be represented as follows:

$$\hat{D}_{i-1} = \begin{cases} \text{null } i = 1 \\ \text{SAM}\left(\text{Conv}_{3\times3}\left(D_{i-1}^r\right)\right) \ 1 < i \leq 4 \end{cases} \tag{3}$$

$$\hat{D} = \frac{\hat{D}_i + \hat{D}_{i-1}}{2} \tag{4}$$

where $\hat{D}$ denotes the refined change attention map. We further emphasize the temporal feature change areas by elementwise multiply. After that, to improve the feature representation, the original temporal features are combined with the augmented temporal features by addition. Finally, the change information is extracted by 3 × 3 convolution. This process can be represented as follows:

$$\hat{F}_i^1 = \text{Conv}_{3\times3}\left(\hat{D} \otimes F_i^1 \oplus F_i^1\right) \tag{5}$$

$$\hat{F}_i^2 = \text{Conv}_{3\times3}\left(\hat{D} \otimes F_i^2 \oplus F_i^2\right) \tag{6}$$

where $\oplus$ denotes an elementwise addition operation, $\otimes$ denotes an elementwise multiply operation, $\hat{F}_i^1$ and $\hat{F}_i^2$ denote the temporal features after change enhancement at T1 and T2 moments, respectively. After the results of concatenating $\hat{F}_i^1$ and $\hat{F}_i^2$ are input into the channel attention module (CAM) [27] to capture the channel correlation, the channel-enhanced features are fed into a 3 × 3 convolution to eliminate some insignificant channels. Then, the refined difference features are added with the previous rough feature difference map D_i^r to compensate for the lost difference information, and finally, the refined difference features are extracted by 3 × 3 convolution. This process can be represented as follows:

$$\hat{F}_i = \text{Cat}\left(\hat{F}_i^1, \hat{F}_i^2\right) \tag{7}$$

$$D_i = \text{Conv}_{3\times3}\left(\text{Conv}_{3\times3}\left(\text{CAM}\left(\hat{F}_i\right) \otimes \hat{F}_i\right) \oplus D_i^r\right) \tag{8}$$

where $\text{Cat}(\cdot)$ denotes concatenation operation, D_i represents the acquired fine temporal change features. The FDE module is employed on four different scales of bitemporal features to concurrently extract and merge the bitemporal image feature information, generating fine multiscale difference features.

C. Adjacent-Level Feature Progressive Fusion (AFPF) Module

In CNN, low-level features offer rich spatial detail information, and high-level features offer abundant semantic information. However, low-level and high-level features have semantic gap, and fusing them directly will cause information loss or semantic confusion. Therefore, it is important to capture the complementary information among the low-level and high-level features effectively and recognize the features with conflicting parts. So, the AFPF module aims to utilize the complementary information between cross-layer features to refine semantic information and spatial details of change features. Fig. 2(e) shows the detailed design of the AFPF. Temporal difference features D_i are used as input of AFPF, i $\in \{1, 2, 3\}$. Owing to the input of AFPF are temporal difference features of adjacent layers, low-level features are denoted by D_i, and high-level features are represented by D_{i+1}, thus i maximizes to 3. Then, the multibranching structure with boundary compensating and knowledge review is used to extract the complementary information among D_i and D_{i+1}.

Specifically, in AFPF, D_{i+1}^u is obtained by using an upsampling operation on D_{i+1}. D_{i+1}^u has the same spatial dimensions as D_i. α and ε are the predicted change maps obtained by Mask for the temporal difference features D_{i+1}^u and D_i, respectively. Conflicting attention can represent the part where conflict occurs between α and ε. This process can be represented as follows:

$$D_{i+1}^u = \text{up}\left(D_{i+1}\right) \tag{9}$$

$$\alpha = \text{Mask}\left(D_{i+1}^{u}\right) \tag{10}$$

$$\varepsilon = \text{Mask}\left(D_{i}\right) \tag{11}$$

$$\theta = \varepsilon \cdot (1-\alpha) + \alpha \cdot (1-\varepsilon) \tag{12}$$

where θ denotes the conflict attention map between D_i and D_{i+1}^{u}. We also bring in boundary compensating attention to obtain boundary information. Boundary information is weaker in high-level features and can be used to locate changing objects. Therefore, boundary compensating attention is guided by utilizing low-level features to make up for the high-level feature deficiency of detailed information. Specifically, boundary compensating attention masks change areas generated from high-level features; thus, AFPF-Net is forced to allocate greater attention to unchanged boundary areas. This process can be represented as follows:

$$\beta = 1 - \alpha \tag{13}$$

where β denotes the boundary compensating attention map. After D_i and D_{i+1}^{u} are concatenated by using AFPF, the feature transition is performed by two 1×1 convolutions to ensure the network captures the change region from different angles. This process can be represented as follows:

$$D_{k}^{f} = \text{Conv}_{1\times 1}\ \left(\text{Cat}\left(D_i, D_{i+1}^{u}\right)\right) \tag{14}$$

$$D_{f}^{f} = \text{Conv}_{1\times 1}\ \left(\text{Cat}\left(D_i, D_{i+1}^{u}\right)\right) \tag{15}$$

where Cat$(\cdot)$ denotes feature concatenation operation, $\text{Conv}_{1\times 1}$ $(\cdot)$ denotes a 1×1 convolutional layer, a BN, and a ReLu function, D_{k}^{f} and D_{f}^{f} represents the input features of the two branches of the AFPF, and D_i represents the input feature of the boundary compensating branch. Conflict attention maps and boundary compensating attention maps are inserted into both branches to enable the network to extract boundary information as well as change regions after refinement. In addition, CAM and 3×3 convolution are also injected into the two branches of AFPF to enhance the feature representation capability and remove some irrelevant channels. This process can be represented as follows:

$$\hat{D}_{k}^{r} = D_{k}^{f}\ \otimes \theta \oplus D_{k}^{f} \tag{16}$$

$$D_{k}^{r} = \text{Conv}_{3\times 3}\ \left(\text{CAM}\left(\hat{D}_{k}^{r}\right) \otimes \hat{D}_{k}^{r}\right) \tag{17}$$

$$D_{f}^{r} = \text{Conv}_{3\times 3}\ \left(\text{CAM}\left(D_{f}^{f}\right) \otimes D_{f}^{f}\right) \tag{18}$$

$$D_{b}^{r} = D_i\ \otimes \beta \tag{19}$$

where $\oplus$ denotes an element-wise addition operation, $\otimes$ denotes an element-wise multiply operation, D_{k}^{r}, D_{f}^{r}, and D_{b}^{r} are the knowledge review branch, the feature fusion branch, and the enhanced features from the boundary compensating branch, respectively. Finally, D_{k}^{r}, D_{f}^{r}, and D_{b}^{r} are concatenated to generate the final refined temporal difference features by 3×3 convolution. This process can be represented as follows:

$$D_{i}^{c} = \text{Conv}_{3\times 3}\ \left(\text{Cat}\left(D_{k}^{r}, D_{f}^{r}, D_{b}^{r}\right)\right) \tag{20}$$

where D_{i}^{c} denotes the final refined difference features. Through the gradual fusion of neighboring features, not only the complementary information between multiscale temporal difference features is fully captured, which enhances the network's capacity to detect real changes, but also the low-level features are utilized to further refine the boundaries of the change area, eventually generating a complete change object.

D. Loss Function

In the CD tasks, the unchanged regions are generally much more numerous than the changed regions, leading to the problem of category weight imbalance in the network training process. To alleviate the effect of sample imbalance, a hybrid loss is adopted, involving dice loss and binary cross-entropy loss [47]. The binary cross-entropy loss can be expressed as follows:

$$L_{\text{bce}} = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)\right] \tag{21}$$

where *N* is the number of samples, y_n denotes the ground truth value of pixel *n*. If $y_n = 1$, it means that a change has occurred, otherwise, $y_n = 0$. $\hat{y}_n$ denotes the probability that a change has occurred, and $1 - \hat{y}_n$ denotes the probability that no change has occurred. The dice loss can be expressed as follows:

$$L_{\text{dice}} = 1 - \frac{2\sum_{n=1}^{N} y_n \hat{y}_n}{\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} \hat{y}_n}. \tag{22}$$

Eventually, the hybrid loss of AFPF-Net can be expressed as follows:

$$L = L_{\text{bce}} + L_{\text{dice}}. \tag{23}$$

IV. Experiments

A. Experiment Preparation

1) Datasets: The performance of AFPF-Net and current SOTA networks is validated on two popular high-resolution remote sensing building CD datasets, namely LEVIR-CD [48] and WHU-CD [49]. LEVIR-CD consists of 637 pairs of images with a spatial resolution of 0.5 m and a spatial size of 1024 × 1024. The dataset LEVIR-CD was sliced into 256 × 256 pixels, and 7120, 1024, and 2048 pairs of images were obtained, which were used in the training set, the evaluation set, and the test set, respectively. WHU-CD consists of a pair image with a spatial resolution of 0.075 m and a spatial size of 32507 × 15354. This dataset is used for the training set, evaluation set, and test set. After cropping it to a size of 256 × 256 pixels, the dataset was randomly divided into three subsets of 5947, 744, and 743 images, which were used in the training set, evaluation set, and test set, respectively.

2) Evaluation Metrics: To accurately evaluate the performance of CD networks, four more common evaluation metrics [50], [51], intersection over union (IoU), F1-score (F1), recall (Rec), and precision (Pre) are used to evaluate the network performance. The detailed definition of the above evaluation

TABLE I
DETAILED DESCRIPTION OF ALL COMPARISON EXPERIMENTS

FC-EF [15]	Concatenating the bi-temporal images makes them into a single image, which is then fed into the network.
FC-Siam-Conc [15]	A shared weight concatenation network is employed to capture multi-scale features and then fuse features of the same scale with the corresponding layers of the decoder through jump connections.
FC-Siam-Diff [15]	Unlike FC-Siam-Conc, element-wise subtraction and absolution operations are performed on feature pairs of the same scale to generate a coarse change map after extracting multi-scale features, which is subsequently fused by a jump connection with the decoder counterpart layer.
ChangeFormer [52]	This is a novel architecture for non-fully convolutional networks. A Siamese network form is used, but the encoder consists of a pure transformer and the decoder uses a multi-layer perceptron (MLP)
BITNet [43]	This is a network based on CNNs and transformers, where temporal features are represented in the form of semantic tokens, the context is modeled by the transformer to enhance the relationship among the tokens, and then a decoder consisting of the transformer is used to refine the original features.
ICIF-Net [44]	This is an intra-scale cross-interaction and inter-scale feature fusion network that exploits the abilities of CNNs in extracting local features and transformers in global modeling to parallel process fine-grained details and changing targets of bi-temporal images.
SNUNet [28]	This is a densely connected CD network that enhances the information interaction between features by connecting all of them at different scales and uses deep supervision to obtain accurate change features.
AFCF3D-Net [7]	Utilizing 3D convolution, to effectively extract and merge feature information from bi-temporal images. Adjacent-level feature cross-fusion is employed to extract complementary information from cross-level features.
DMINet [8]	This is a two-branch, multi-level cross-temporal network that unifies self-attention and cross-attention, allowing attention to focus precisely on the regions where real change is happening and suppressing regions of non-change and task-irrelevant change.

metrics can be expressed as follows:

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{24}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{25}$$

$$\text{F1} = \frac{2\text{Pre} \cdot \text{Rec}}{\text{Pre} + \text{Rec}} \tag{26}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{27}$$

where TP, FP, TN, FN denote the numbers of true positive, false positive, true negative, and false negative, respectively.

B. Experimental Environment

The AFPF-Net is implemented in the Pytorch architecture, and all experiments are tested on the Intel Xeon Gold 5315Y (CPU), and NVIDIA A800 (GPU). Data augmentation is performed on the input training images using flipping and cropping. The Adam optimizer is used to optimize the AFPF-Net with momentum, weight decay, β_1 and β_2 initially set to 0.9, 0.0001, 0.9, and 0.99, respectively. The learning rate is adjusted using the poly learning scheme as $(1 - (\text{cur\_iteration}/\text{max\_iteration}))^{\text{power}} \times \text{lr}$, where power and max_ iteration are set to 0.9 and 20 000, respectively, batch size and learning rate are initialized to 32 and 0.0001, respectively.

C. Comparison With SOTA Methods

To confirm the effectiveness of AFPF-Net, several representative change detection networks from recent years have been selected for comparative experiments. FC-EF [15], FC-Siam-Conc [15], and FC-Siam-Diff [15] are fully convolutional-based networks. AFCF3D-Net [7] is based on 3-D convolution, attention mechanisms, and multiscale. DMINet [8] is network based on convolution, attentional mechanism, and transformer. SNUNet [28] is a network based on convolution, attentional mechanism, and multiscale. BITNet [43] and ICIF-Net [44] are networks based on convolution and transformers. ChangeFormer [52] is based on multiscale, MLP, and transformer. Table I provides a detailed the description of the comparative experiments. To ensure the fairness of the comparisons, they were reproduced using their released source code and under their default hyperparameters.

1) Experimental Results Analysis: To validate the effectiveness of AFPF-Net, experimental evaluations were conducted on two different datasets. Table II shows the experimental outcomes for all comparison networks and AFPF-Net. To enhance readability, the optimal and second-optimal outcomes are denoted as red and blue, respectively. ChangeFormer, BITNet, ICIF-Net, and DMINet are based on a concatenation of transformer and CNN, or pure transformer. From Table II, it can be seen that the pure CNN-based AFPF-Net outperforms the next best network on the LEVIR-CD dataset and the WHU-CD dataset in terms of IoU and F1 by 0.56%/0.33% and 3.07%/1.74%, respectively.

TABLE II
Comparison Results of Two CD Datasets

	LEVIR-CD	WHU-CD
Model	Pre. / Rec. / F1. / IoU.	Pre. / Rec. / F1. / IoU.
FC-EF(2018)	88.57 / 81.90 / 85.10 / 74.07	78.60 / 73.20 / 75.81 / 61.04
FC-Siam-Diff(2018)	89.29 / 84.92 / 87.05 / 77.07	83.99 / 79.72 / 81.80 / 69.20
FC-Siam-Conc(2018)	88.40 / 85.78 / 87.07 / 77.10	87.99 / 83.23 / 85.54 / 74.74
ChangeFormer(2022)	91.27 / 86.17 / 88.65 / 79.61	89.86 / 82.31 / 85.92 / 75.31
BITNet(2021)	90.20 / 88.38 / 89.28 / 80.64	86.57 / 88.89 / 87.71 / 78.12
ICIF-Net(2022)	91.76 / 88.54 / 90.12 / 82.02	91.13 / 88.87 / 89.99 / 81.80
SNUNet(2022)	91.75 / 89.37 / 90.55 / 82.72	94.83 / 90.62 / 92.67 / 86.35
AFCF3D-Net(2023)	92.33 / 89.03 / 90.65 / 82.90	93.79 / 91.60 / 92.68 / 86.36
DMINet(2023)	91.87 / 89.73 / 90.79 / 83.13	92.49 / 88.04 / 90.21 / 82.17
AFPF-Net(ours)	91.71 / 90.54 / 91.12 / 83.69	95.72 / 93.15 / 94.42 / 89.43

The optimal and second-optimal results are denoted as red and blue, respectively. All indicators are expressed as percentages (%).

The reason is that pure transformer-based methods focus more on global information and are not very good at detecting the overall detailed information of the change region and small object changes. However, in change detection, although it is important to utilize global information to locate the change region, it is also important to control the detailed information. Although BITNet, ICIF-Net, and DMINet focus on both global and detailed information by combining transformer and CNN, they do not emphasize the generation of feature difference maps. The AFPF-Net extracts multiscale bitemporal features through the Backbone shown in Fig. 2(a). Features captured at higher levels contain more semantic information and pay more attention to global information, while features captured at lower levels contain more detailed information. Then the FDE module generates a fine multiscale change map, and finally, it is input to the AFPF module through a progressive fusion method, in which the low-level features are guided by high-level features to learn the real changes from a global perspective, and the detailed information lacking in the high-level features is compensated by the low-level features. So, the generated change regions have smoother boundaries and more complete content. It should be noted that although AFPF-Net obtains the optimal scores in IoU and F1 on both datasets, its accuracy on LEVIR-CD is not the highest, and some pseudochanges are still not detected. Based on the above analysis, it can be concluded that AFPF-Net relies solely on attention mechanisms and CNNs, and its performance is significantly better than other networks on the two datasets.

2) *Visualization of Results:* Figs. 3 and 4 show the results of the visualization of all methods on both datasets. To highlight the performance gap between AFPF-Net and other networks, some samples are selected for local feature magnification comparisons. As can be seen in Fig. 5, the outline shapes, edge locations, and completeness of the detected change regions in our results are clearer than those generated by the other methods, especially the areas marked by the red boxes. To effectively demonstrate the disparities between network predictions and labels, four distinct colors are utilized to highlight the detection outcomes. White, black, red, and blue are used for TP, TN, FP, and FN, respectively.

a) *Visualization on LEVIR-CD:* The visualization results for the LEVIR-CD dataset are shown in Fig. 3, which shows that the AFPF-Net has fewer blue parts compared to the other methods. As can be seen from the first four rows of Fig. 3, most methods have FP zones (red zones) affected by "nonsemantic changes," while AFPF-Net recognizes a very low percentage of false variations, especially in the fourth row, where AFPF-Net not only eliminates most of the interferences of "nonsemantic changes" but also accurately identifies other networks that fail to check the building changes. In the second and last three rows of Fig. 3, when only small buildings change or the changed objects are relatively numerous, most methods have the problem of losing small changing objects, but AFPF-Net is more accurate in detecting small object changes. Especially in the second and sixth rows, most methods tend to focus on the changes of in large or medium-sized buildings and easily ignore the changes of small buildings. In contrast, the AFPF-Net can take care of all the changing objects, even the changes of small buildings occurring in the upper-left corner of the figure in the sixth row can be detected. This confirms that AFPF-Net has better performance than other comparison networks.

b) *Visualization on WHU-CD:* Fig. 4 illustrates the visualization of the WHU-CD dataset, where most CD approaches fail to recognize all the details of the changing objects, but the AFPF-Net can obtain more complete changing objects. In the top five rows of Fig. 4, most approaches filter out some "nonsemantic changes" regions to some extent, such as shadows and road coverage, and are successful at locating real change zones, but the ability to recognize whole details of change objects remains inadequate, whereas the AFPF-Net is able to obtain more complete change buildings. From the last two rows of Fig. 4, it can be seen that most methods are limited capability in handling the boundaries of change areas, causing the boundaries of change objects irregular or adjacent change regions connected, while AFPF-Net can better handle the boundary details and obtain sharper boundary. Especially in the sixth row of Fig. 4, AFPF-Net successfully identifies the building changes and ensures the integrity of the change region edges.

From the above analysis of the visualization, although transformer-based methods, such as ChangeFormer, BITNet, and ICIF-Net can accurately locate the changed regions in more cases, the full convolution-based AFPF-Net can also locate these regions and is more effective for the detection of completeness than other networks.

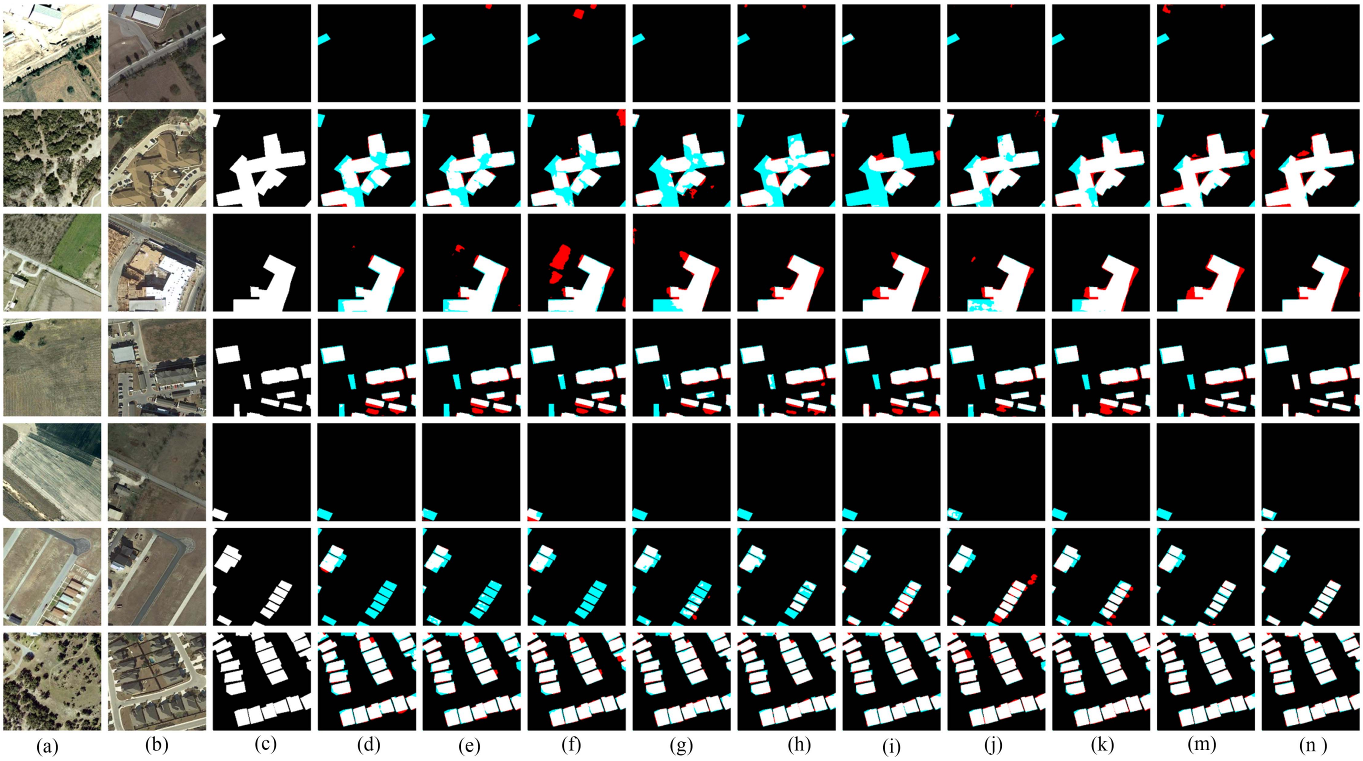


Fig. 3. Visualization of the results on the LEVIR-CD dataset. (a) T1 image. (b) T2 image. (c) Label. (d) FC-EF. (e) FC-Siam-Diff. (f) FC-Siam-Conc. (g) ChangeFormer. (h) BITNet. (i) ICIF-Net. (j) SNUNet. (k) AFCF3D-Net. (m) DMINet. (n) AFPF-Net. TP, TN, FN, and FP are denoted as white, black, blue, and red.

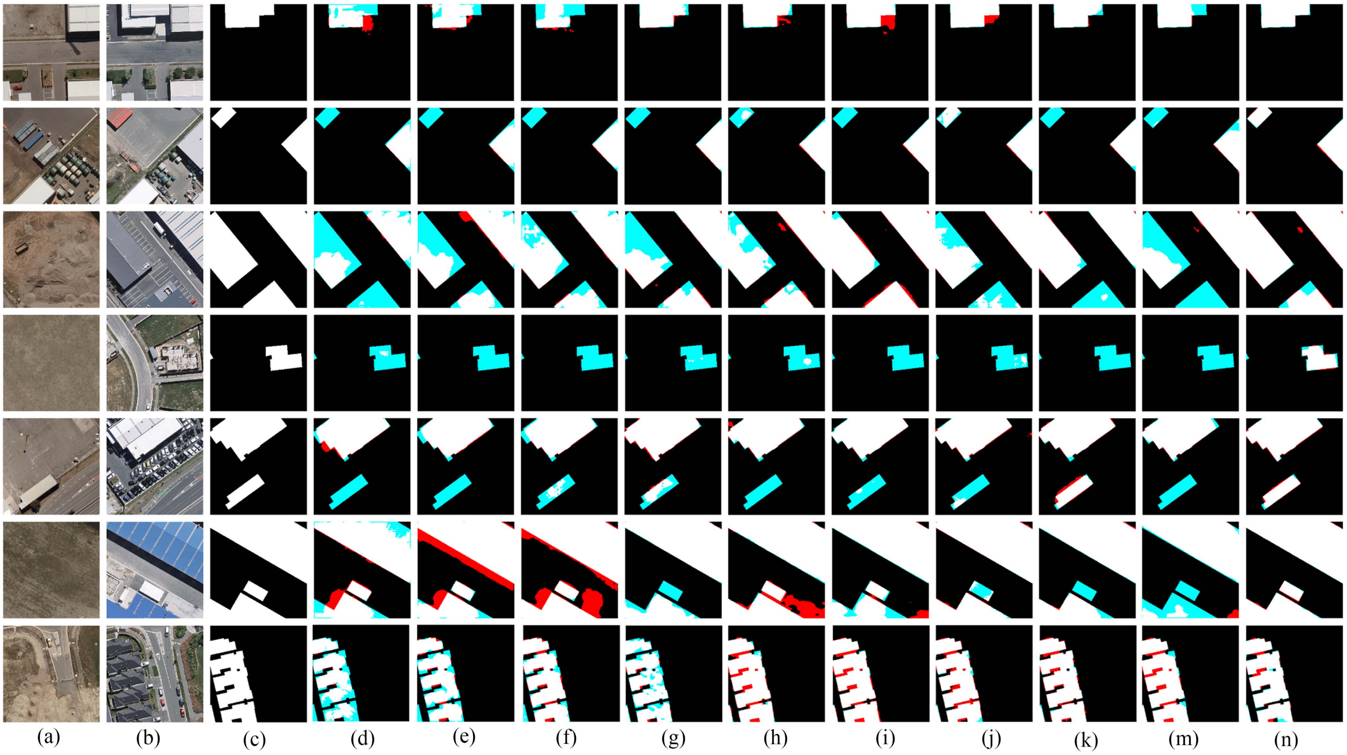


Fig. 4. Visualization of the results on the WHU-CD dataset. (a) T1 image. (b) T2 image. (c) Label. (d) FC-EF. (e) FC-Siam-Diff. (f) FC-Siam-Conc. (g) ChangeFormer. (h) BITNet. (i) ICIF-Net. (j) SNUNet. (k) AFCF3D-Net. (m) DMINet. (n) AFPF-Net. TP, TN, FN, and FP are denoted as white, black, blue, and red.

TABLE III
Ablation Study on FDE and AFPF Modules

		LEVIR-CD		WHU-CD	
FDE	AFPF	F1	IoU	F1	IoU
×	×	90.13	82.03	93.65	88.05
√	×	90.47	82.60	93.86	88.44
×	√	91.00	83.48	93.99	88.66
√	√	91.12	83.69	94.42	89.43

All experimental indicators are expressed as percentages (%).

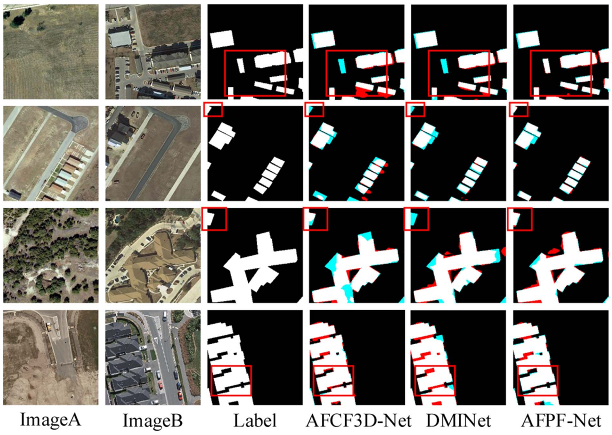


Fig. 5. Detailed comparison of visualization results.

3) Performance Analysis: As shown in Fig. 6, the tradeoffs between the IoU of the other methods and AFPF-Net with the number of floating-point operations (FLOPs) and the number of parameters (Params) are analyzed on the LEVIR-CD and WHU-CD datasets. As can be seen in Fig. 6, FC-EF, FC-Siam-Conc, and FC-Siam-Diff obtain poor results owing to a lack of complex feature extraction structures, but Params and FLOPs are relatively low. The other methods, due to use of transformers or 3-D convolution, achieve good results but with relatively high complexity. At relatively low Params and FLOPs, AFPF-Net achieves better results than other networks.

D. Ablation Experiments

The effectiveness of the modules within the AFPF-Net is evaluated by combining or removing the AFPF-Net modules, ablation experiments were conducted on the WHU-CD and LEVIR-CD datasets. To enhance the readability of the results, it is preferred that the results be labeled in red.

1) Effectiveness of FDE and AFPF: To evaluate the effectiveness of FDE and AFPF, ablation experiments are conducted. From Table III, it can be observed that the detection results on both datasets are worse when any one module is removed than when both modules exist simultaneously, which reflects the synergistic effect of the two modules. FDE is responsible for extracting the fine feature difference maps, and AFPF is responsible for extracting the complementary information in the fine difference maps of different scales, which makes the detected changed objects more complete.

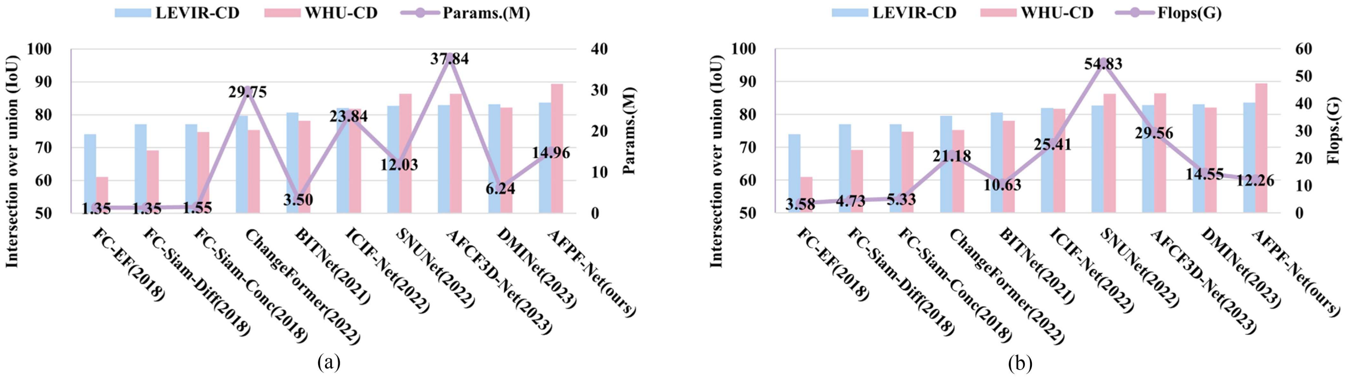
2) Effectiveness of Different Backbone Networks: Ablation experiments are conducted using ResNet18, ResNet34, ResNet50, and ResNet101 as the backbone networks, respectively, as shown in Table IV. Specifically, the number of residual blocks in the other backbones carries on increasing compared to ResNet18, so the quantity of floating-point operations and parameters increases greatly. Except for ResNet34, the other backbones trade computational load for improved network performance. However, although the parameters of ResNet34 are increased, the overall network performance does not exceed ResNet18. Therefore, ResNet18 is selected as the backbone of AFPF-Net.

3) Effectiveness of Different Channel Reductions: The ablation experiments focus on the impact of channel numbers, with channel numbers set to 64, 96, and 128, respectively. From Table V, it can be seen that the channel number reduced to 64 performs the best. Interestingly, when the number of multi-scale feature channels is 96 or 128, the detection performance is worse. We believe that the reason is that too many channels can easily lead to overfitting of the network. Therefore, a rise in the quantity of channels does not inherently result in enhanced performance.

4) Effectiveness of Connectivity Between FDEs: To verify the effectiveness of skip connections between FDEs, we conducted comparative experiments using dense connections. The results are shown in Table VI, and there is little difference in the number of parameters between the two connection methods. A dense connection fuses the difference features after each previous layer FDE refinement with the current layer feature difference map, which enhances the change features but also causes feature redundancy, leading to an overfitting phenomenon.

5) Effectiveness of FDE: To verify the effectiveness of FDE, four variants of the FDE module, which are module without previous layer of difference features (Pldf) and skip connections, module without Pldf, module without skip connections, and the complete FDE module, are set to go for the ablation experiments. From Table VII, it can be noticed that the effect of module without both Pldf and skip connections is worse than modules only using Pldf or skip connections, proving that these two branches are effective. Using only Pldf or skip connections has worse performance than the complete FDE, which is especially evident on WHU-CD, where the IoU metric is worse by 1.8%. Therefore, an FDE using both Pldf and skip connections is the best choice.

6) Effectiveness of AFPF: To verify the effectiveness of AFPF, four variants of the AFPF module, which are module without conflict attention (conflict_att) and boundary compensating attention (boundary_att), module without conflict_att



(a) (b)

Fig. 6. All method Params. and FLOPs results on both CD datasets. (a) IoU versus Params. (b) IoU versus FLOPs.

TABLE IV
ABLATION STUDY ON AFPF-NET WITH DIFFERENT BACKBONES

			LEVIR-CD		WHU-CD	
Backbone	Params.(M)	FLOPs(G)	F1	IoU	F1	IoU
resnet18	14.96	12.26	91.12	83.69	94.42	89.43
resnet34	25.07	17.10	90.84	83.22	94.19	89.03
resnet50	27.66	18.47	91.20	83.82	94.53	89.63
resnet101	46.65	28.23	91.37	84.10	93.72	88.19

All experimental indicators are expressed as percentages (%).

TABLE V
ABLATION STUDY ON DIFFERENT CHANNEL REDUCTIONS

			LEVIR-CD		WHU-CD	
CR	Params.(M)	FLOPs(G)	F1	IoU	F1	IoU
64	14.96	12.26	91.12	83.69	94.42	89.43
96	19.58	21.53	91.02	83.52	93.83	88.37
128	26.03	34.50	91.06	83.58	93.73	88.20

All experimental indicators are expressed as percentages (%).

TABLE VI
ABLATION STUDY ON CONNECTIVITY BETWEEN FDES

			LEVIR-CD		WHU-CD	
Method	Params.(M)	FLOPs(G)	F1	IoU	F1	IoU
skip connection	14.96	12.26	91.12	83.69	94.42	89.43
dense connection	15.22	12.37	91.04	83.55	94.15	88.95

All experimental indicators are expressed as percentages (%).

TABLE VII
ABLATION STUDY ON FDE

		LEVIR-CD		WHU-CD	
pldf	skip connection	F1	IoU	F1	IoU
×	×	90.90	83.32	93.35	87.53
√	×	90.89	83.31	93.54	87.87
×	√	91.00	83.48	93.41	87.63
√	√	91.12	83.69	94.42	89.43

All experimental indicators are expressed as percentages (%).

module without boundary_att and the complete module, are set to go for the ablation experiments. From Table VIII, it can be viewed that the effect of modules without both conflict_att and boundary_att is worse than those only using conflict_att or boundary_att, proving that these two branches are effective. Modules using only conflict_att or boundary_att have worse performance than the full AFPF.

Based on the results obtained from the previous six ablation experiments, it can be concluded that the network architecture depicted in Fig. 2 is optimal.

TABLE VIII
Ablation Study on AFPF

		LEVIR-CD		WHU-CD	
conflict_att	boundary_att	F1	IoU	F1	IoU
×	×	90.84	83.21	93.98	88.64
√	×	90.93	83.37	93.97	88.63
×	√	90.96	83.43	94.00	88.67
√	√	91.12	83.69	94.42	89.43

All experimental indicators are expressed as percentages (%).

V. Conclusion

In this article, a novel network called the AFPF-Net for the task of remote sensing image CD is proposed. To fully extract reliable variation information and augmented detail in the change region, the FDE module is proposed to extract and fuse the temporal features, using residual connections to aggregate the difference information of the previous layer to optimize the difference of the current layer's bitemporal features. To address the incompleteness of change objects and the noise interference caused by the direct fusion of various scale features, the AFPF module is proposed to improve the change features, utilizing complementary information among adjacent layers. Using two publicly accessible datasets (LEVIR-CD and WHU-CD), the efficacy of the proposed model is confirmed. Moreover, the ability of the proposed method to detect small targets needs to be improved, so our future research will be devoted to increasing the model's efficiency while accurately detecting small targets.

References

[1] Y. Hu, Y. Dong, and Batunacun, "An automatic approach for land-change detection and land updates based on integrated NDVI timing analysis and the CVAPS method with GEE support," *ISPRS J. Photogrammetry Remote Sens.*, vol. 146, pp. 347–359, 2018.

[2] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, 2008.

[3] R. E. Kennedy et al., "Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects," *Remote Sens. Environ.*, vol. 113, no. 7, pp. 1382–1396, 2009.

[4] H. Liu, M. Yang, J. Chen, J. Hou, and M. Deng, "Line-constrained shape feature for building change detection in VHR remote sensing imagery," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 10, 2018, Art. no. 410.

[5] F. Bovolo and L. Bruzzone, "A split-based approach to unsupervised change detection in large-size multitemporal images; application to tsunami-damage assessment," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1658–1670, Jun. 2007.

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

[7] Y. Ye, M. Wang, L. Zhou, G. Lei, J. Fan, and Y. Qin, "Adjacent-level feature cross-fusion with 3-D CNN for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5618214.

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

[9] W. Wang, C. Tang, X. Wang, and B. Zheng, "A ViT-based multiscale feature fusion approach for remote sensing image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4510305.

[10] W. Wang, T. Hu, X. Wang, and J. Li, "BFRNet: Bidimensional feature representation network for remote sensing images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5621213.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4095–4104.

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

[14] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 101–119, 2022.

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

[16] W. Wiratama and D. Sim, "Fusion network for change detection of high-resolution panchromatic imagery," *Appl. Sci.*, vol. 9, no. 7, 2019, Art. no. 1441.

[17] 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.

[18] 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 Observation Geoinf.*, vol. 101, 2021, Art. no. 102348.

[19] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628711.

[20] W. Wang, C. Liu, G. Liu, and X. Wang, "CF-GCN: Graph convolutional network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5607013.

[21] J. Wu et al., "A multiscale graph convolutional network for change detection in homogeneous and heterogeneous remote sensing images," *Int. J. Appl. Earth Observation Geoinf.*, vol. 105, 2021, Art. no. 102615.

[22] 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 Observation Geoinf.*, vol. 105, 2021, Art. no. 102591.

[23] S. Zhu, Y. Song, Y. Zhang, and Y. Zhang, "ECFNet: A Siamese network with fewer FPS and fewer FnS for change detection of remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6001005.

[24] T. Lei et al., "Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4507013.

[25] H. Zhong and C. Wu, "T-UNet: Triplet unet for change detection in high-resolution remote sensing images," *Geo-spat. Inf. Sci.*, pp. 1–18, 2024.

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

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

[28] 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.

[29] B. Liu, H. Chen, Z. Wang, W. Xie, and L. Shuai, "LSNET: Extremely lightweight Siamese network for change detection of remote sensing image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 2358–2361.

[30] J. Wang et al., "SSCFNet: A spatial-spectral cross fusion network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4000–4012, 2023.

[31] Z. Li, C. Tang, X. Li, W. Xie, K. Sun, and X. Zhu, "Towards accurate and reliable change detection of remote sensing images via knowledge review and online uncertainty estimation," 2023, *arXiv:2305.19513*.

[32] J. Ma, G. Shi, Y. Li, and Z. Zhao, "MAFF-Net: Multi-attention guided feature fusion network for change detection in remote sensing images," *Sensors*, vol. 22, no. 3, 2022, Art. no. 888.

[33] J. Zheng et al., "MDESNet: Multitask difference-enhanced Siamese network for building change detection in high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 15, 2022, Art. no. 3775.

[34] A. Eftekhari, F. Samadzadegan, and F. Dadrass Javan, "Building change detection using the parallel spatial-channel attention block and edge-guided deep network," *Int. J. Appl. Earth Observation Geoinf.*, vol. 117, 2023, Art. no. 103180.

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

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

[37] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, "Deep multiscale Siamese network with parallel convolutional structure and self-attention for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406512.

[38] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5998–6008.

[39] J. Gu et al., "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 12084–12093.

[40] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 6877–6886.

[41] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, May. 2021.

[42] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 16473–16483.

[43] 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.

[44] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410213.

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

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

[47] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis.*, Oct. 2016, pp. 565–571.

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

[49] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multi-source building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

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

[51] X. Song, Z. Hua, and J. Li, "GMTS: GNN-based multi-scale transformer Siamese network for remote sensing building change detection," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 1685–1706, 2023.

[52] 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.

Wei Wang received the B.S., M.S., and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 1997, 2003, and 2010, respectively.

He is currently a Professor with the School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha. His research interests include remote sensing image processing, computer vision, and deep learning.

Luocheng Xia received the B.E. degree in information engineering college from the Jiangxi University of Technology, Jiangxi, China, in 2022. He is currently working toward the M.E. degree in computer technology with the School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, China.

His research interests include remote sensing image processing and computer vision.

Xin Wang received the B.S. and M.S. degrees in information and communication engineering from the Wuhan University of Technology, Wuhan, China, in 1998 and 2006, respectively.

She is currently a Lecturer with the School of Computer and Communication, Changsha University of Science and Technology, Changsha, China. Her research interests include signal processing, computer vision, and pattern recognition.