Adaptive Fusion NestedUNet for Change Detection Using Optical Remote Sensing Images

Junwei Li, *Member, IEEE*, Shijie Li, and Feng Wang

***Abstract*—Change detection (CD) is a major topic in remote sensing research. Deep learning (DL)-based CD methods have made great progress. However, existing CD methods have difficulty in exploiting the different semantic and detailed information in deep and shallow features, which often leads to blurred target boundaries of identified changes. In addition, most CD methods based on the NestedUNet structure focus on improving accuracy, ignoring the importance of efficiency. Therefore, in this article, an adaptive fusion NestedUNet for CD (AFNUNet) is proposed. AFNUNet compresses the model parameters and computational cost by using the encoder based on the inverted bottleneck structure and the decoder based on the depthwise convolution. The correlation between the final multilevel feature maps extracted by NestedUNet is difficult to model by summation or concatenation. Therefore, an attention mechanism-based adaptive fusion module (AFM) is proposed. The AFM allows the network to adaptively select feature information from the final different layers of features extracted from NestedUNet in both channel and spatial dimensions so that the fused features capture deep rich semantic information while retaining detailed information at shallow boundaries. Finally, a loss function based on the Bray-Curtis distance is introduced for suppressing the sample imbalance problem. Extensive experiments on the WHU-CD, LEVIR-CD, and SYSU-CD datasets demonstrate that AFNUNet surpasses several state-of-the-art (SOTA) CD methods in terms of effectiveness. Moreover, the proposed AFNUNet remarkably reduces Params and FLOPs by 63% and 70% compared to other NestedUNet-based CD models.**

***Index Terms*—Adaptive fusion module (AFM), change detection (CD), deep learning (DL), remote sensing (RS).**

I. Introduction

Change detection (CD) is an essential and challenging topic in remote sensing (RS). It aims to identify the differences in the surface based on bitemporal or multitemporal RS images. This technique is very crucial in various fields, including disaster assessment [1], environmental investigation [2], urban planning [3], [4], forest monitoring [5], and land use dynamics detection [6]. Recently, due to the rapid advancements in satellite RS technology, high-resolution (HR) optical sensors have been increasingly designed for observing the earth. The increasing number of HR optical RS images provides strong support for various RS applications.

In recent decades, plenty of CD methods have been proposed. The traditional CD methods can be categorized into two types: pixel-based change detection (PBCD) methods and object-based change detection (OBCD) methods [6]. The PBCD methods generate difference maps by pixel-by-pixel comparison of paired bitemporal images [7]. Some PBCD methods have been proposed, including algebra-based methods for change vector analysis (CVA) [8], classification-based methods [9], transformation-based methods for principal component analysis (PCA) [10], multivariate alteration detection (MAD) [11], iteratively reweighted multivariate alteration detection (IR-MAD) [12], and machine learning-based methods [13]. Although it is easy to implement the PBCD methods, they ignore the spatial contextual information, which results in a great deal of salt-and-pepper noise during processing. To address this problem, various work has been presented in the literature based on different approaches, such as Markov random fields [14], conditional random fields [15], and level sets [16]. However, the PBCD methods are unsuitable for processing very high-resolution (VHR) images due to the increased variability within the image objects. To perform CD in VHR images, a few scholars have proposed OBCD methods. The OBCD methods first utilize spectral and texture information for segmenting an image into disjoint objects and then compare and analyze the bitemporal objects to obtain the change map [17]. In [18], an OBCD method that was robust to illumination and noise changes was proposed by fusing the texture and luminance differences between various frames. In [19], an OBCD method was proposed to detect abrupt changes and subtle variations by combining the profile and texture of objects from a geometric perspective. In [5], an OBCD method was proposed to identify the changes in the forest land cover by combining image differencing, image segmentation, and statistical testing. Although the OBCD methods use the information of spatial features from HR images, the traditional manual feature extraction methods are more complex and exhibit poor robustness.

Due to the ability of deep learning (DL) techniques in learning the feature information from images effectively, many academics have introduced the DL in RS images CD [20], [21], [22]. The DL-based CD methods can predict the spatial context and pixel classification maps from the original images. As a result,

Manuscript received 20 February 2023; revised 4 May 2023; accepted 4 June 2023. Date of publication 7 June 2023; date of current version 19 June 2023. This work was supported in part by the Programs for Science and Technology Development of Henan Province, China under Grant 222102210004, in part by the Key Research Projects of the University in Henan Province, China under Grant 20B510001, and in part by Weinan Science and Technology under Grant ZDYF-JCYJ-196. *(Corresponding author: Feng Wang.)*

Junwei Li and Shijie Li are with the School of Artificial Intelligence, Henan University, Zhengzhou 450000, China (e-mail: lijunwei@henu.edu.cn; lishijie123@henu.edu.cn).

Feng Wang is with the School of Physics and Electrical Engineering, Weinan Normal University, Weinan 714000, China (e-mail: wangfeng81113@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3283524

they break the boundaries between traditional PBCD and OBCD methods. It is noteworthy that compared to the conventional CD methods, the DL-based methods do not require preprocessing. This not only enables them to avoid the errors caused by preprocessing but also assists in reducing the postprocessing workload.

The existing DL-based RS image CD methods are divided into two main types. The first type is based on the single-branch structure and the second type is based on the double-branch structure. Both single- and double-branch structures require the extraction of features at different scales and a series of processing of these feature maps for obtaining change maps. However, for bitemporal RS images, the fusion strategy is different between the single-branch and double-branch structures. The single-branch structure fuses the prechanged and postchanged images by concatenating and then inputting them into the network for feature extraction. In [20], the concatenated bitemporal images were fed into U-Net to precisely segment the changed regions. In [21], the concatenated prechanged and postchanged images were used as the input of the UNet++ structure to effectively utilize fine-grained and global information for accurate feature maps. In [22], the prechanged and postchanged images and their difference maps were fed into HRNet [23] to obtain better CD performance. Different from the single-branch structure, the double-branch structure first extracts features from both branches of the prechanged image and the postchanged image and then fuses the obtained bitemporal features. In [24], the Siamese network and fully convolutional network (FCN) with skip connections were combined to address the dense prediction problem. In [25], the Siamese convolution was combined with a dual-attention mechanism for enhancing the ability of the model in recognizing change information. In [26], the pretrained SE-ResNet50 [27] was combined with the Siamese structure to effectively extract features. In [28], dual temporal features were effectively constrained in both encoding and decoding. The Siamese encoder was used for the extraction of the correct dual temporal features and the dual decoder was used for their effective fusion. In [30], the Siamese ResNet18 [29] was used for feature extraction, and transformers encoded and decoded the extracted features to model the contextual relationships in the spatial-temporal domain. In [31], a fully transformer-based Siamese structure efficiently demonstrated long-range details in multiscale features.

Most DL-based CD methods already have a good performance in detecting changes between bitemporal RS images. However, the existing CD methods pay little attention to the complementary information between different layers in the network. They tend to extract change information using deep features, ignoring the importance of shallow features containing fine-grained information, which often leads to loss of boundary details and mislocalization of changed regions. There is various work [32], [33], [34] that shows the strong semantic information representation capability of deep networks. However, the overall of small objects and the edge details of large objects are gradually lost with the network's multiple down-sampling and up-sampling. The shallow networks are effective in terms of detailed information representation but are weak in semantic information representation. The changed area in RS images contains both large vegetation changes and small target building changes. This makes the information extraction from different layers of the network have both the same or similar contents and significant differences. Besides, the internal structure of most NestedUNet structure-based CD methods stacks multiple 3×3 convolutional layers to achieve better CD performance, but this also leads to huge parameters and computation.

To solve the mentioned problems, an end-to-end network based on the NestedUNet, called adaptive fusion NestedUNet (AFNUNet), is proposed. AFNUNet applies an inverted bottleneck structure in the feature extraction stage and uses depthwise convolution in the feature fusion stage to improve operational efficiency. For effective aggregation using complementary information from different levels of features, an adaptive fusion module (AFM) based on channel attention and spatial attention [35] is designed. The AFM uses the softmax attention guided by feature information at multiple semantic levels so that different levels of features receive different attention. This improves the ability of the network to discriminate the changed regions and retain boundary detail information. In addition, a loss function based on the Bray-Curtis distance (BCD) [36] is introduced for improving the performance of the model in identifying differences between bitemporal images. The main contributions of this work are as follows.

1) We propose a network, AFNUNet, performing CD based on optical RS images. The proposed network has strong detailed and semantic information representation capability. Moreover, it has strong competitiveness in terms of model parameters and computational cost.
2) We propose a module that can effectively fuse multiple semantic-level features, namely AFM. The AFM adaptively fuses multiple semantic-level features in both channel and spatial dimensions to effectively utilize the complementary information and enhance attention to the boundaries of changed regions.
3) A loss based on BCD is proposed in this work. This loss balances the effect of changed and unchanged samples on the network and improves the ability of the network to identify changed regions.

The rest of this article is organized as follows. Section II reviews related work on the U-Net series-based CD methods. Section III is a detailed description of the proposed method. The implementation details, comparison experiments, ablation experiments, and analysis are presented in Section IV. Finally, Section V concludes this article.

II. Related Work

The U-Net series have been widely used for biomedical image segmentation tasks. U-Net [32] used the concatenation operation for the aggregation of shallow and deep features at the corresponding scales for the segmentation task. UNet++ [33] incorporated features extracted from deep and shallow networks by using different operations, such as nested and dense skip connections, instead of using simple feature concatenation for both encoding and decoding layer features of the same size. UNet3+ [34] used a more dense skip connection, allowing each decoder layer to incorporate full-scale feature maps. The CD task can be viewed as segmenting the changed regions in the image. Therefore, many scholars have proposed networks for performing the CD task with various improvements based on the U-Net series.

Daudt et al. [24] proposed three CD networks based on U-Net, the first one was based on an early fusion strategy and the other two networks were Siamese extensions of the first one. Sun et al. [37] proposed a U-Net that aggregated long short-term memory to handle multiscale spatial features. Chen et al. [28] proposed a U-Net based on the squeeze and excitation module which was used for capturing the response among feature channels.

Peng et al. [21] proposed UNet++ with multiside output fusion, where multiside output fusion was used for fusing the multiside output feature maps of the UNet++ backbone to obtain a highly accurate final change map. Peng et al. [38] proposed a simplified UNet++ with the dense attention unit, where the dense attention approach used high-level features with rich semantic information to guide the selection of low-level features to capture the change features. Zhang et al. [39] proposed UNet++ with a multiside fusion strategy, where the multiside fusion strategy was used to effectively predict changed targets at different scales. Fang et al. [40] combined NestedUNet with the Siamese network. The ensemble channel attention module was used to aggregate and refine the feature mappings obtained from the NestedUNet backbone at multiple semantic levels. Raza et al. [41] proposed UNet++ with efficient encoders and decoders. The attention-based bitemporal feature fusion strategy was used to refine multiscale features and avoid loss during downsampling. Liu et al. [42] proposed UNet++ with spatial-temporal-channel attention, where the spatial-temporal-channel attention mechanism enabled selective feature extraction. Li et al. [43] proposed pseudo-Siamese UNet++, where each branch was based on UNet++ and not sharing weights to extract heterogeneous input image difference features. Du et al. [44] combined transformer and UNet++, and the transformer was used to effectively model the global semantic relationship of features extracted from the convolutional neural network (CNN).

Zhao et al. [45] combined UNet3+ with the Siamese network, where the Siamese network was used for feature extraction and UNet3+ was used for full-scale feature fusion to reduce localization error. Mo et al. [46] proposed Siamese UNet3+ with channel- and spatial-based attention modules, where the attention module was used to effectively identify change features.

However, the U-Net-based CD methods [24], [37] only aggregate feature maps at the same scale in the encoder and decoder networks and do not fully utilize the feature information at different scales, resulting in less accurate localized changed regions. The UNet++-based CD methods [21], [38], [39], [40], [41], [42], [43], [44] have shown excellent performance through dense skip connections. However, they often have huge parameters and computational costs hardly to meet the requirements of high efficiency. In addition, some work [21], [39] fuses features of different levels extracted by UNet++ with equal weight, ignoring the semantic gap between them. Some work [41], [43] directly concatenates and fuses features from different levels of UNet++ extraction, increasing the difficulty of modeling their relevance in the network. Some work [38], [42], [44] utilizes deep features of UNet++ for detecting changed targets, losing some of the fine-grained information of shallow features.

In [40], this work accounts for the weighted fusion of different levels of features in the channel dimension, ignoring the importance of location information of the spatial dimension. As a result, the above work lacks sufficient attention to the boundaries of shallow networks, making the detection results incomplete. With full-scale skip connections, the UNet3+-based CD methods [45], [46] achieve good performance but also further increase the computational cost.

The main purpose of this article aims to utilize the complementary information in different levels of feature maps efficiently and effectively to enhance the CD performance. Different from existing UNet++ and UNet3+-based CD methods that utilize standard convolution for encoding and decoding, we employ an inverted bottleneck structure in the encoding stage for reducing the number of parameters and improving CD performance [47], [48], and depthwise convolution in the decoding stage to further improve efficiency [49]. An attention-based feature fusion module is then used to adaptively select the required information from the feature maps extracted from the UNet++ backbone in the channel and spatial dimensions, respectively, thus, enhancing the attention to changed target boundaries.

III. Methodology

In this section, the general framework of the proposed network is first introduced. Then the efficient channel attention-based feature extractor and adaptive fusion module are presented in detail. Finally, we describe the network optimization strategy.

A. *Network Architecture*

The proposed AFNUNet is a standard encoder-decoder architecture, as presented in Fig. 1. The prechanged and postchanged images are concatenated and used as the input of the proposed network. Multiscale features are obtained by feature extractors. The convolutional block for feature fusion is shown in Fig. 1(b). The 1×1 convolutional layer is used to aggregate the features from the encoding and decoding layers and the 5×5 depthwise convolutional layer is used to improve the decoding efficiency while further perceiving the change information. To attain low-level texture characteristics and high-level semantic features, a dense skip connection mechanism is applied between encoders and decoders. Assuming $x^{i,j}$ denotes the output of node $\mathrm{X}^{i,j}$, where i denotes ith down-sampling layer along the encoder direction and j denotes jth convolutional layer along the skip pathway. The accumulation of feature maps is mathematically described as follows:

$$x^{i,j} = \begin{cases} \mathcal{P}(\mathcal{R}([x^{t1}, x^{t2}])), & i = 1, j = 0 \\ \mathcal{P}(\mathcal{R}(x^{i-1,j})), & i \neq 1, j = 0 \\ \mathcal{C}([\mathcal{U}(x^{i+1,j-1}), [x^{i,k}]_{k=0}^{j-1}]), & j > 0 \end{cases} \tag{1}$$

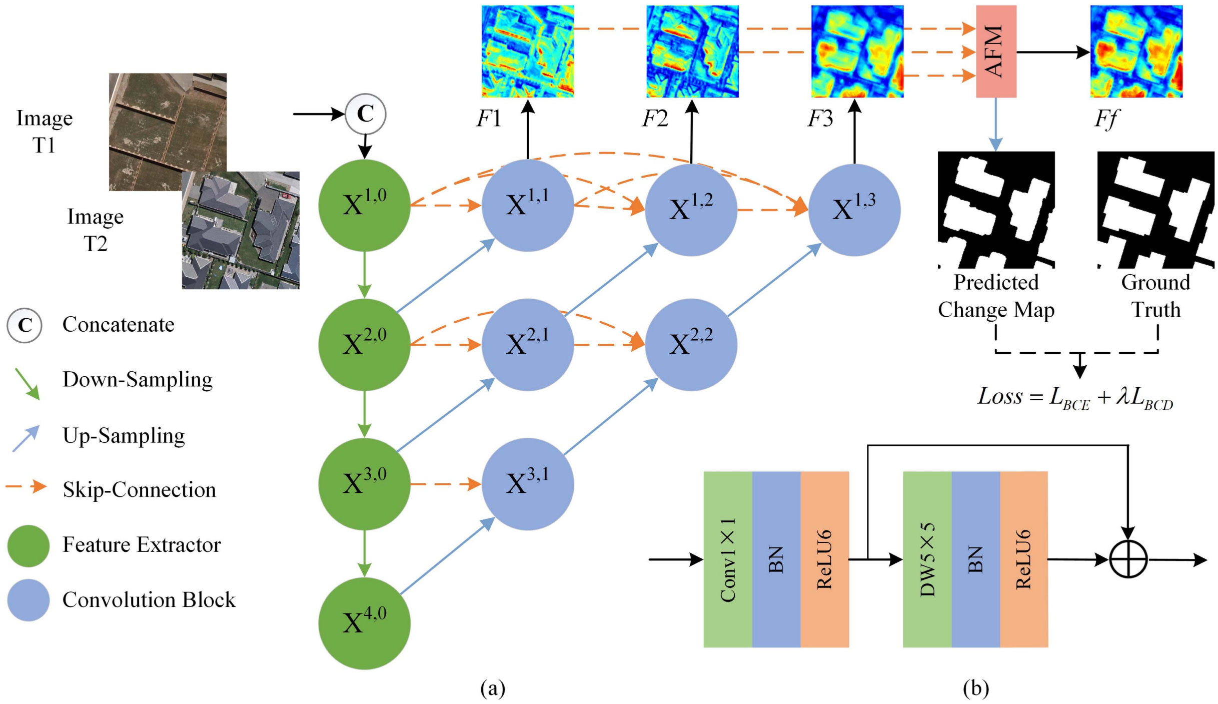

Fig. 1. Architecture of the proposed AFNUNet. (a) Overview of the proposed AFNUNet. (b) Convolution block.

where x^{t1} and x^{t2} denote the input bitemporal images, and [ ] denotes the concatenation operation. The function $\mathcal{R}(\cdot)$ denotes the encoding operation. The function $\mathcal{P}(\cdot)$ denotes the down-sampling for the feature maps using a 2×2 max pooling operation. The function $\mathcal{U}(\cdot)$ denotes the use of the $Upsample$ method for up-sampling. The function $\mathcal{C}(\cdot)$ denotes the operation of fusion using the convolution block.

In the proposed AFNUNet, the outputs of the same hierarchical nodes have the same size. In the down-sampling stage, the output features of each decoder are doubled in the number of channels and halved in size compared to the input feature mapping. In the up-sampling phase, each node has two or more inputs. Considering node $\mathrm{X}^{1,2}$ as an example, nodes $\mathrm{X}^{1,0}$, $\mathrm{X}^{1,1}$ from the same level and up-sampled node $\mathrm{X}^{2,1}$ are concatenated for performing convolution block operation to obtain the node $\mathrm{X}^{1,2}$. Finally, multiple features extracted from the backbone of the proposed AFNUNet which have different semantic information of the same scale are fed into the AFM for further enhancing the extraction of detailed information regarding the changed regions.

B. Feature Extractor

We analyze some CD networks based on the UNet++ structure [21], [33], [38], [39], [40], [42], and they all use a double-layer 3×3 standard convolution in the encoding stage, which is one of the reasons for their inefficiency. In the proposed method, the encoder is redesigned for higher efficiency. The inverted bottleneck structure has been demonstrated for its ability to improve performance while reducing the number of parameters [47], [48]. The efficient channel attention (ECA) [50] not only focuses on the channel of interest by using cross-channel interactions, but it also has a lower computational complexity. In this work, the inverted bottleneck structure and ECA are combined to form a feature extractor for suppressing the incomplete changed target profile caused by multiple down-sampling operations and obtaining more feature information about the changed region during the feature extraction stage.

Fig. 2 shows the structure of the feature extractor, composed of three convolutional layers and one ECA layer. The first 3×3 convolutional layer conducts the dimension-raising operation on the input feature map, while the second and third 1×1 convolutional layers perform the channel numbers doubling and halving operations, respectively, i.e., a thick middle and thin end structure. The loss of feature information due to the replacement 3×3 convolution is reduced by performing expansion between two 1×1 convolutions. Wang et al. [50] used the global average pool (GAP) to obtain aggregated features. To extract more feature information, we make some changes in the ECA. We use global max pooling (GMP) to create a branch parallel to the GAP. This process is expressed as follows:

$$w = y \otimes (\sigma(C1D_k(\mathrm{MaxPool}(y)) + C1D_k(\mathrm{MaxPool}(y)))) \tag{2}$$

where y denotes the output of the third convolutional layer, MaxPool and AvgPool are utilized to generate two aggregated vectors, $C1D_k$ denotes a fast 1-D convolution of kernel size k (in this work, $k = 3$), σ denotes the sigmoid function, and w represents the output obtained after ECA. The encoding stage

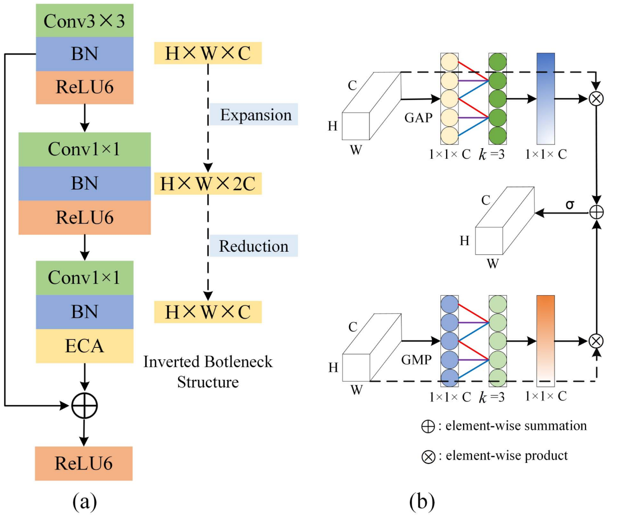

Fig. 2. Architecture of the proposed feature extractor. (a) Overview of the feature extractor. The inverted bottleneck structure refers to expanding the number of channels in the middle layer using 1×1 convolution and then reducing the number of channels to match the initial convolution layer by 1×1 convolution. (b) Improved ECA.

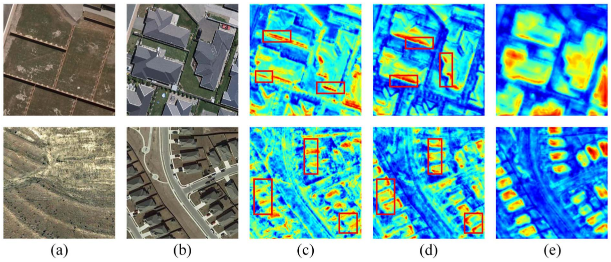

Fig. 3. Visualization mapping of the final feature maps of NestedUNet. (a) Image at time1. (b) Image at time2. (c) $\mathrm{X}^{1,1}$. (d) $\mathrm{X}^{1,2}$. (e) $\mathrm{X}^{1,3}$. Blue indicates lower attention values and red indicates higher attention values.

contains four layers of feature extractors. The depths obtained after the four encoder layers are 64, 128, 256, and 512.

C. Adaptive Fusion Module

The three same-sized groups of feature maps extracted by the NestedUNet backbone contain different semantic information. As shown in Fig. 3, the shallow features retain more texture details and boundary information but are accompanied by more noise. The deeper features are semantically rich and accurately locate changed regions, but some detailed information is lost. Exploring the correlation between them by simple summation or splicing fusion is vulnerable to the interference of semantic gaps between features at different levels. Intuitively, an automatic feature selection strategy is needed to fuse these features for focusing on the shallow boundaries and deep localization of the changed target for obtaining more accurate detection results.

As shown in Fig. 4, the adaptive fusion module (AFM) is designed to improve feature representation by adaptively selecting change information between different levels. Structurally, the AFM is an extension of CBAM [35] and SKNet [51] in integrating features. Three feature maps F_1, F_2, and F_3 are first extracted by using the proposed AFNUNet and then fused by performing element-by-element summation as follows:

$$F = F_1 + F_2 + F_3 \tag{3}$$

where F denotes the fusion result obtained by integrating the information from multiple branches. The AFM suppresses the channels and locations of uninterest and emphasizes those of interest in the channel and spatial dimensions, assigning higher weights to interested channels and locations and lower weights to uninterested channels and locations, allowing the network to select the required information from the appropriate level of features. The mathematical expression for the channel attention submodule is expressed as follows:

$$M_c = \mathrm{MLP}(\mathrm{MaxPool}(F)) + \mathrm{MLP}(\mathrm{AvgPool}(F)) \tag{4}$$

$$F_c = a \cdot F_1 + b \cdot F_2 + c \cdot F_3 \tag{5}$$

where MaxPool and AvgPool are applied on the fused features F for generating two $\mathrm{C} \times 1 \times 1$ aggregation vectors of size. The above vectors are processed through the multilayer perception (MLP) module with shared weights to obtain two vectors of size $3\ \mathrm{C} \times 1 \times 1$. M_c denotes the channel attention map obtained by performing the elemental sum of the two aforementioned vectors. The soft attention (softmax layer) is applied to feature M_c to adaptively select different semantic levels from the channel dimension. a, b, and c denote the soft attention vectors obtained after the application of the softmax layer. The size of a, b, and c are $\mathrm{C} \times 1 \times 1$, where a_i denotes the ith element of a, and so on, and $i \in [0, \mathrm{C})$. The softmax layer is used to obtain $a_i + b_i + c_i = 1$ by summing the specified dimensions to 1. Then, the original feature maps F_1, F_2, and F_3 are subject to elementwise multiplication with the attention weights in different channels for obtaining the feature maps F_c.

Similarly, the spatial attention submodule uses MaxPool and AvgPool in the first step for generating two matrices of size $1 \times \mathrm{H} \times \mathrm{W}$. For efficiency, a kernel-sized 7×7 convolutional layer with shared weights is applied to these two matrices. After the convolutional layer, two matrices of sizes $3 \times \mathrm{H} \times \mathrm{W}$ are obtained. M_s denotes the spatial attention map obtained by the application of elementwise summation of the above two matrices. The soft attention is applied to feature M_s to adaptively select different semantic levels on the spatial dimension. Let a, b, and c denote the soft attention matrixes obtained after the application of the softmax layer. The sizes of a, b, and c are $1 \times \mathrm{H} \times \mathrm{W}$ (where $a_{i,j}$ denotes the jth element of the ith row of a, and so on, where $a_{i,j} + b_{i,j} + c_{i,j} = 1$, $i \in [0, \mathrm{H})$, $j \in [0, \mathrm{W})$). Now, the original feature maps F_1, F_2, and F_3 are subject to elementwise multiplication with the attention weights in different spatial dimensions for obtaining the feature maps F_s

$$M_s = f^{(7\times7)}(\mathrm{MaxPool}(F)) + f^{(7\times7)}(\mathrm{AvgPool}(F)) \tag{6}$$

$$F_s = a \cdot F_1 + b \cdot F_2 + c \cdot F_3. \tag{7}$$

Finally, the feature maps of the two submodules are summed to obtain the final fused features F_f as follows:

$$F_f = F_c + F_s. \tag{8}$$

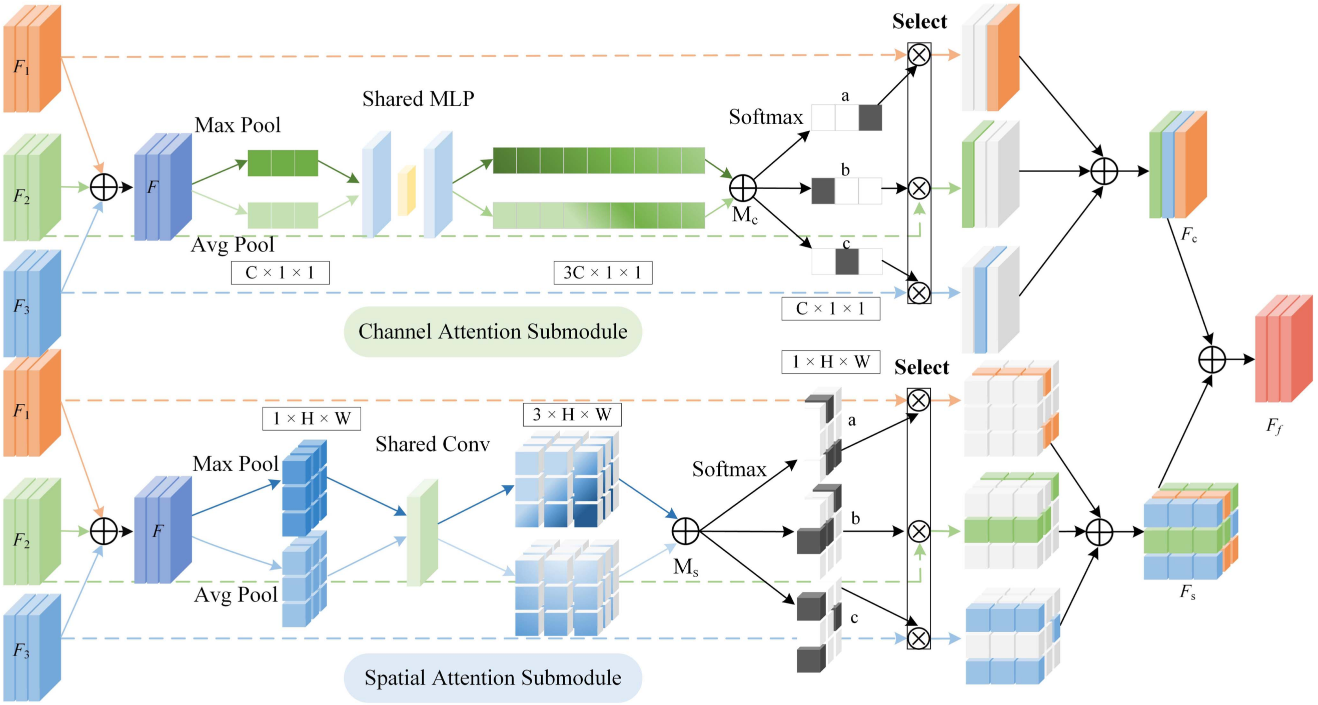

Fig. 4. Aarchitecture of the proposed AFM. Three final feature maps F_1, F_2, and F_3 extracted from the UNet++ backbone are initially fused by summation. The fused feature map F is processed through the MLP and convolution with shared weights for obtaining the channel and spatial attention maps, respectively. After the softmax layer assigns weights to the feature maps F_1, F_2, and F_3, the most required information is adaptively selected from the feature maps F_1, F_2, and F_3, i.e., the detailed boundary information of the shallow network and the rich semantic information of the deep network.

D. Loss Function

The binary cross-entropy (BCE) [52] is a loss function commonly used in binary classification problems and is defined as follows:

$$L_{BCE}(X_n, Y_n) = \sum_{i=1}^{M} [-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)] \tag{9}$$

where $X_n = \{(x_n^{t1}, x_n^{t2}), n = 1, 2, \ldots, N\}$ and $Y_n = \{y_n, n = 1, 2, \ldots, N\}$ denote training images and the ground truth (GT), $\hat{y}_i \in [0, 1]$ denotes the predicted probability of a pixel point in the change map being a change pixel and $y_i = \{0, 1\}$ denotes the probability of a pixel point in the label map being a change pixel, and M denotes the product of the height and width of the image. 0 and 1 denote no change and change, respectively.

The $L_{BCE}(X_n, Y_n)$ assigns equal weights to each pixel during training. There is a problem of fewer changed pixels and more unchanged pixels in CD, which may cause the training process to be dominated by the unchanged pixel class, making the model biased toward the unchanged class and ignoring the changed class, thus increasing the difficulty of the model to identify changes.

The BCD [36] is mainly used in ecological environmental science for calculating the distance between the coordinates and the differences between the samples. To solve the class imbalance in the sample, we introduce it into CD. The value of BCD loss ranges from 0 to 1. The larger the value, the greater the difference between the prediction map and the GT. The loss function based on the BCD $L_{BCD}(X_n, Y_n)$ is defined as follows:

$$L_{BCD}(X_n, Y_n) = \frac{\sum_{i=1}^{M} |\hat{y}_i - y_i|}{\sum_{i=1}^{M} \hat{y}_i + \sum_{i=1}^{M} y_i}. \tag{10}$$

The BCD loss is a region-dependent loss, where the loss of the current pixel is related to both the predicted value of the current pixel and the values of the other points. Since the value of GT is either 0 or 1, the formulation for the BCD loss can be differentiated to yield the gradient

$$\frac{\partial L_{BCD}(X_n, Y_n)}{\partial \hat{y}_j} = \begin{cases} \frac{\sum_{i=1}^{M} \hat{y}_i + \sum_{i=1}^{M} y_i - \sum_{i=1}^{M} |\hat{y}_i - y_i|}{\left(\sum_{i=1}^{M} \hat{y}_i + \sum_{i=1}^{M} y_i\right)^2}, & y_j = 0 \\ -\frac{\sum_{i=1}^{M} \hat{y}_i + \sum_{i=1}^{M} y_i + \sum_{i=1}^{M} |\hat{y}_i - y_i|}{\left(\sum_{i=1}^{M} \hat{y}_i + \sum_{i=1}^{M} y_i\right)^2}, & y_j = 1 \end{cases} \tag{11}$$

where $\hat{y}_j$ and y_j denote any predicted pixel point and its corresponding pixel point in GT.

Note that the positive and negative of the gradient only indicate the direction, so only the values of the gradients need to be compared. The results of the gradients demonstrate that when a pixel value in GT is 1, the resulting gradient is greater than that when a pixel value in GT is 0, indicating that the BCD loss is directional and more biased toward the changed class.

Finally, the objective function $L(X_n, Y_n)$ of the proposed network is defined as follows:

$$L(X_n, Y_n) = L_{BCE}(X_n, Y_n) + \lambda L_{BCD}(X_n, Y_n) \tag{12}$$

When $\lambda = 0$, only the benchmark loss $L_{BCE}(X_n, Y_n)$ is used. We present the impact of λ for different datasets later in this work.

IV. Experiments

In this section, the proposed AFNUNet is evaluated by using three CD datasets, including the WHU building CD (WHU-CD) dataset [4], the LEVIR-CD dataset [53], and the SYSU-CD dataset [54]. We also perform a series of ablation experiments by using each of the three datasets. Finally, an efficiency comparison of the proposed method with different methods is performed.

A. Datasets and Implementation Details

1) Datasets: WHU-CD Dataset: The WHU-CD dataset consists of pairs of images of size 15 354 × 32 507 pixels acquired using the satellite. The images in this dataset cover the area where the 2011 Christchurch, New Zealand earthquake occurred. This area was rebuilt in the subsequent years. We divide each image pair into patches of size 256 × 256 pixels without any overlap and randomly divide the dataset into training, validation, and test sets at a ratio of 8:1:1. Finally, we obtain 5908 training samples, 763 validation samples, and 763 test samples.

LEVIR-CD Dataset: The LEVIR-CD dataset contains 637 pairs of optical RS images. Each image has a resolution of 1024 × 1024 pixels and is collected from Google Earth. These images mainly cover various types of building growth with significant land-use changes. The dataset is divided into training, validation, and test sets. We use the partition method [38] and obtain 3167 training samples, 436 validation samples, and 935 test samples.

SYSU-CD Dataset: The SYSU-CD dataset consists of 20 000 pairs of aerial images of size 256 × 256 pixels. The images are acquired between 2007 and 2014 in Hong Kong. The major changes in the SYSU-CD dataset include vegetation changes, suburban dilation, groundwork before construction, newly built urban buildings, and road expansion.

2) Experimental Settings: The proposed AFNUNet uses AdamW as the optimizer with an initial learning rate = 0.001 and weight decay = 0.0001. The learning rate is reduced by a factor of 0.5 after every 10 epochs. The batch size of AFNUNet is set to 16. In addition, AFNUNet is implemented using the PyTorch DL framework. All the experiments are conducted on a single NVIDIA GeForce RTX 3090 with 24 GB memory.

3) Comparative Method and Evaluation Metrics: We compare the proposed AFNUNet with DL-based CD models, which mainly include the UNet++ structure-based and attention-based methods. Fully convolutional-early fusion (FC-EF) [24] takes the concatenated bitemporal images as input. Fully convolutional-Siamese-difference (FC-Siam-Diff) [24] and fully convolutional-Siamese-concatenation (FC-Siam-Conc) [24] are Siamese extensions of FC-EF. NestedUNet (UNet++) [33] reduces the semantic gap between feature maps by nesting dense skip connections. UNet++ with multiple side output fusion (UNet++_MSOF) [21] applies the UNet++ backbone for performing the CD task. Difference-enhancement dense-attention convolutional neural network (DDCNN) [38] simplifies the UNet++ structure and models the correlation between different levels of features to obtain change features by the dense attention method in the decoder. Difference-enhancement unit is used to selectively aggregate high-level difference features. Siamese NestedUNet (SNUNet) [40] incorporates a Siamese network and a NestedUNet and uses the ensemble channel attention module to suppress the localization error and semantic vacancy. Bitemporal image transformer (BIT) [30] combines CNN and transformer to model context in the space-time domain. CNN-transformer network with multiscale context aggregation (MSCANet) [55] captures features at different scales by CNN and encodes and aggregates multiscale features using the transformer. Full-scale connected Siamese network (SiUNet3+) [45] combines the Siamese network with a modified UNet3+ to produce a discriminative and precisely located change map.

In this work, precision (P), recall (R), F1 score ($F1$), and intersection over union (IoU) [56], [57] are used for quantitative evaluation of the performance of different methods. The $F1$ and IoU evaluation metrics quantify the overall performance of a model used for the CD task. The higher the values of these metrics, the better the prediction result. The aforementioned evaluation metrics are computed as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{13}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{14}$$

$$F1 = \frac{2PR}{P{+}R} \tag{15}$$

$$IoU = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{16}$$

where true positive (TP), false positive (FP), and false negative (FN) denote the number of true positives, the number of false positives, and the number of false negatives, respectively.

B. Comparison Experiments

1) WHU-CD Dataset: To evaluate the effectiveness of the proposed AFNUNet, we first conduct a comparison experiment using the WHU-CD dataset containing only semantic changes in buildings. The results are shown in Table I, indicating that AFNUNet achieves the highest $F1$ and IoU with 92.32% and 85.73%, compared to the improvement of 1.53% and 2.60% over SiUNet3+, respectively.

To intuitively understand the prediction results of different methods using the WHU-CD test set, we present the visualization results in Fig. 5. As shown in the first three rows of Fig. 5, for larger changes and with less interference, all methods can identify significantly changed buildings. However, AFNUNet is more sensitive to building boundaries by the role of AFM and identifies more complete buildings. As shown in the fourth row of Fig. 5, BIT and AFNUNet overcome the pseudochanges in the road surface and identify the changed large buildings. And compared to BIT, AFNUNet extracts buildings with more boundary detail. Furthermore, as can be seen in the last row of

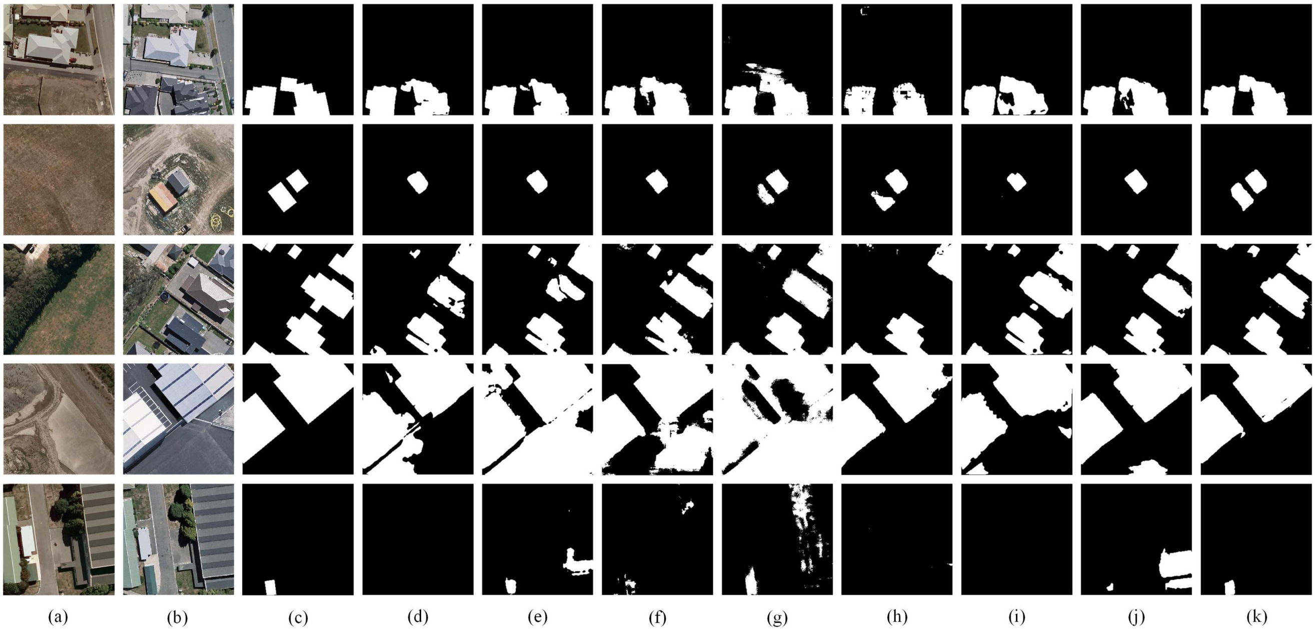

Fig. 5. Visualization results on the WHU-CD test set. (a) Image at time1. (b) Image at time2. (c) Ground truth. (d) UNet++. (e) UNet++_MSOF. (f) DDCNN. (g) SNUNet. (h) BIT. (i) MSCANet. (j) SiUNet3+. (k) AFNUNet.

TABLE I
COMPARISON OF EXPERIMENTAL RESULTS USING THE WHU-CD DATASET

Method	P(%)	R(%)	F1(%)	IoU(%)
FC-EF [24]	81.88	83.05	82.46	70.16
FC-Siam-Diff [24]	28.00	**95.21**	43.28	27.61
FC-Siam-Conc [24]	31.47	93.79	47.13	30.83
UNet++ [33]	87.71	84.52	86.09	75.57
UNet++_MSOF [21]	80.91	91.35	85.81	75.15
DDCNN [38]	88.35	92.70	90.47	82.60
SNUNet [40]	87.70	92.72	90.14	82.05
BIT [30]	90.98	87.51	89.21	80.52
MSCANet [55]	89.68	90.01	89.84	81.56
SiUNet3+ [45]	89.81	91.79	90.79	83.13
AFNUNet	**92.40**	92.23	**92.32**	**85.73**

The bold entities highlight the best results for different metrics for different methods and distinguish the quality of the methods.

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS USING THE LEVIR-CD DATASET

Method	P(%)	R(%)	F1(%)	IoU(%)
FC-EF [24]	84.92	78.21	81.43	68.67
FC-Siam-Diff [24]	88.98	82.29	85.51	74.69
FC-Siam-Conc [24]	91.48	77.65	83.99	72.40
UNet++ [33]	92.04	86.22	89.03	80.23
UNet++_MSOF [21]	91.59	86.62	89.04	80.24
DDCNN [38]	**92.95**	84.96	88.77	79.81
SNUNet [40]	90.08	88.03	89.04	80.25
BIT [30]	90.15	89.94	90.04	81.89
MSCANet [55]	87.26	**92.58**	89.84	81.55
SiUNet3+ [45]	91.38	88.02	89.67	81.27
AFNUNet	91.95	89.97	**90.95**	**83.40**

The bold entities highlight the best results for different metrics for different methods and distinguish the quality of the methods.

Fig. 5, the proposed AFNUNet accurately locates small changes in the building with less noise. The experimental results show that the proposed AFNUNet performs well on the WHU-CD dataset and exhibits strong resistance to interference.

2) LEVIR-CD Dataset: We also conduct experiments on another building CD dataset. The metric results of the comparison methods on the LEVIR-CD test set are shown in Table II. The proposed AFNUNet boosts performance on $F1$ and IoU to 90.95% and 83.40%, compared to the improvement of 0.91% and 1.51% over BIT, respectively.

The visualization results of the different methods are shown in Fig. 6. As presented in the first two rows in Fig. 6, it is difficult for other networks to identify the changed buildings when the changed area is small. Through the effect of the BCD loss balance classes, AFNUNet can identify subtle changes and thus locate the changed target accurately. When there are many small changed buildings (see rows 3 and 4 in Fig. 6), all comparison methods perform well. AFNUNet accurately identifies more changed buildings. In the detection of the large changed building (see row 5 in Fig. 6), our AFNUNet can also extract it more completely. The results demonstrate that AFNUNet achieves good performance in this dataset and effectively extracts the overall features of the changed buildings.

3) SYSU-CD Dataset: Finally, we perform experiments by using the SYSU-CD dataset. Different from the WHU-CD and LEVIR-CD datasets, the SYSU-CD dataset has more complex change scenarios. The quantitative results of the SYSU-CD test set are shown in Table III. AFNUNet achieved the highest $F1$ and IoU with 80.09% and 66.79%, respectively.

Fig. 7 visualizes the results of the comparison methods. For vessel changes (first two rows of Fig. 7), most methods lose part of the change information. In this case, AFNUNet obtains a more complete detection result. In the case of building area changes, see the last three rows of Fig. 7, the changes identified by most methods are quite limited due to the more complex scenes (e.g., shadow interference, tree growth) and irregular

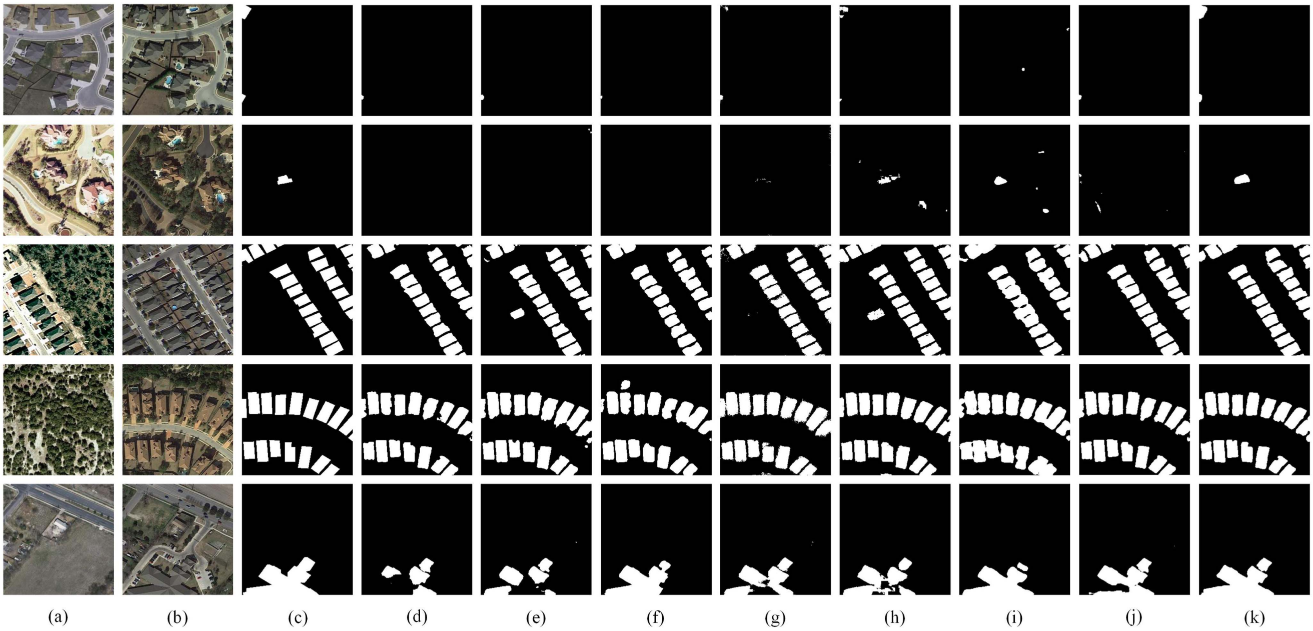

Fig. 6. Visualization results on the LEVIR-CD test set. (a) Image at time1. (b) Image at time2. (c) Ground truth. (d) UNet++. (e) UNet++_MSOF. (f) DDCNN. (g) SNUNet. (h) BIT. (i) MSCANet. (j) SiUNet3+. (k) AFNUNet.

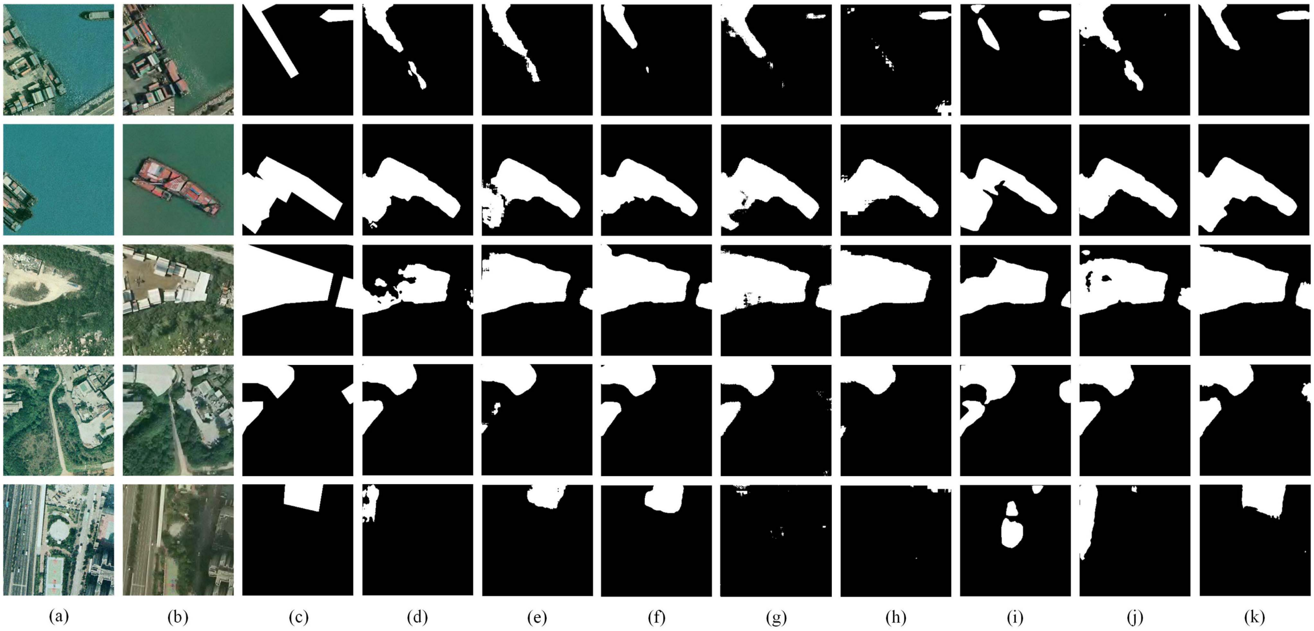

Fig. 7. Visualization results on the SYSU-CD test set. (a) Image at time1. (b) Image at time2. (c) Ground truth. (d) UNet++. (e) UNet++_MSOF. (f) DDCNN. (g) SNUNet. (h) BIT. (i) MSCANet. (j) SiUNet3+. (k) AFNUNet.

change areas. However, the proposed AFNUNet still obtains relatively complete change results.

C. Ablation Study

To evaluate the proposed AFNUNet, AFM, and BCD loss, a series of ablation experiments are conducted. Tables IV–VI present the detection accuracy obtained using the WHU-CD, LEVIR-CD, and SYSU-CD datasets, respectively.

The experimental results demonstrate that the AFM improves the detection accuracy by 1.27% in terms of $F1$ and 2.13% in terms of IoU for the WHU-CD dataset; 0.58% in terms of $F1$ and 0.97% in terms of IoU for the LEVIR-CD dataset; and 1.02% in terms of $F1$ and 1.40% in terms of IoU for the SYSU-CD dataset. The contribution of AFM is shown in Fig. 8(e), i.e., the network extracts richer feature information and identifies more complete boundaries of the changed targets when using the AFM. The BCD loss effectively suppresses the

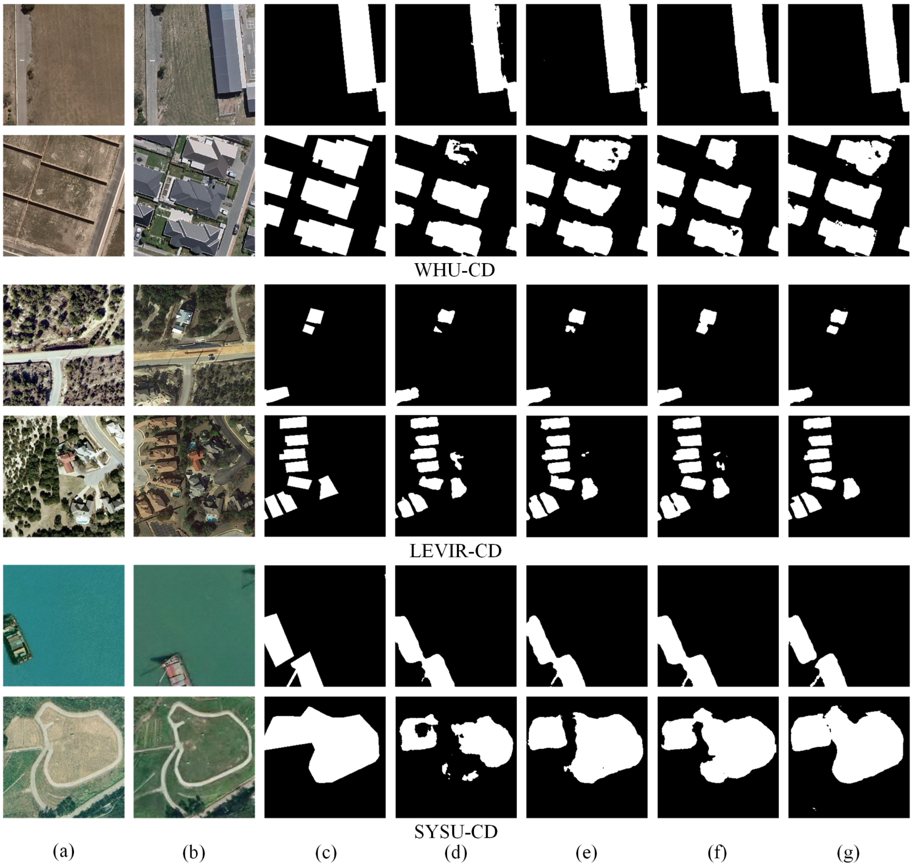

Fig. 8. Examples of ablation experiments performed using the proposed method. (a) Image at time1. (b) Image at time2. (c) Ground truth. (d) Proposed AFNUNet without BCD loss and AFM. (e) Proposed AFNUNet without BCD loss. (f) Proposed AFNUNet without AFM. (g) Proposed AFNUNet.

TABLE III
COMPARISON OF EXPERIMENTAL RESULTS USING THE SYSU-CD DATASET

Method	P(%)	R(%)	F1(%)	IoU(%)
FC-EF [24]	79.12	66.93	72.52	56.88
FC-Siam-Diff [24]	**92.39**	39.90	55.73	38.63
FC-Siam-Conc [24]	80.36	71.19	75.50	60.64
UNet++ [33]	82.05	73.72	77.67	63.49
UNet++_MSOF [21]	84.70	71.95	77.81	63.67
DDCNN [38]	76.52	**78.83**	77.66	63.48
SNUNet [40]	82.36	74.42	78.19	64.19
BIT [30]	83.18	72.92	77.72	63.56
MSCANet [55]	74.58	75.48	75.02	60.03
SiUNet3+ [45]	80.31	77.30	78.77	64.98
AFNUNet	82.66	77.68	**80.09**	**66.79**

The bold entities highlight the best results for different metrics for different methods and distinguish the quality of the methods.

TABLE IV
ABLATION EXPERIMENTS PERFORMED USING THE WHU-CD DATASET

Method	P(%)	R(%)	F1(%)	IoU(%)
w/o BCD Loss, AFM	89.29	91.04	90.15	82.07
w/o BCD Loss	91.20	91.65	91.42	84.20
w/o AFM	90.84	91.40	91.12	83.69
AFNUNet	**92.40**	**92.23**	**92.32**	**85.73**

The bold entities highlight the best results for different metrics for different methods and distinguish the quality of the methods.

TABLE V
ABLATION EXPERIMENTS PERFORMED USING THE LEVIR-CD DATASET

Method	P(%)	R(%)	F1(%)	IoU(%)
w/o BCD Loss, AFM	**92.36**	88.07	90.16	82.08
w/o BCD Loss	92.26	89.27	90.74	83.05
w/o AFM	91.71	89.52	90.59	82.81
AFNUNet	91.95	**89.97**	**90.95**	**83.40**

The bold entities highlight the best results for different metrics for different methods and distinguish the quality of the methods.

imbalance between the changed and unchanged samples, thus accurately identifying the differences between the bitemporal images. When this loss function is added to the training process, the detection accuracy improves by 0.97% in terms of $F1$ and 1.62% in terms of IoU for the WHU-CD dataset; 0.43% in terms of $F1$ and 0.73% in terms of IoU for the LEVIR-CD dataset; and 1.08% in terms of $F1$ and 1.48% in terms of IoU for the SYSU-CD dataset.

The contribution of BCD loss is presented in Fig. 8(f). The results illustrate that the model identifies more change features when using the BCD loss. Fig. 8(g) is the visualization of AFNUNet's prediction maps, it combines the advantages of AFM and BCD loss resulting in well-defined change target boundaries, rich change information, and less noise.

We also visualize the three final feature maps F_1, F_2, and F_3 in the network and the feature map F_f after using the AFM to

TABLE VI
ABLATION EXPERIMENTS PERFORMED USING THE SYSU-CD DATASET

Method	P(%)	R(%)	F1(%)	IoU(%)
w/o BCD Loss, AFM	**84.43**	72.89	78.24	64.25
w/o BCD Loss	81.26	77.36	79.26	65.65
w/o AFM	81.19	77.54	79.32	65.73
AFNUNet	82.66	**77.68**	**80.09**	**66.79**

The bold entities highlight the best results for different metrics for different methods and distinguish the quality of the methods.

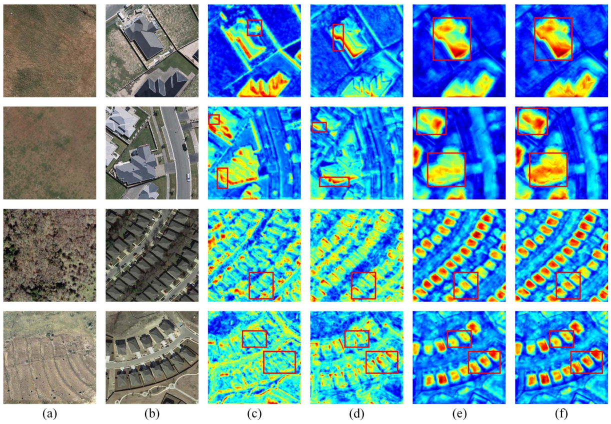

Fig. 9. Visualization mapping of some features of AFNUNet. (a) Image at time1. (b) Image at time2. (c) F_1. (d) F_2. (e) F_3. (f) F_f. Blue indicates lower attention values and red indicates higher attention values.

TABLE VII
SENSITIVITY EXPERIMENTS ON BCD LOSS

λ	WHU-CD		LEVIR-CD		SYSU-CD	
	F1(%)	IoU(%)	F1(%)	IoU(%)	F1(%)	IoU(%)
0	91.42	84.20	90.74	83.05	79.26	65.65
0.2	91.76	84.78	90.81	83.16	79.76	66.34
0.4	91.95	85.10	90.79	83.13	79.65	66.19
0.6	92.07	85.32	90.92	83.35	79.42	65.86
0.8	**92.32**	**85.73**	90.85	83.23	**80.09**	**66.79**
1.0	91.75	84.76	**90.95**	**83.40**	79.48	65.95

The bold entities highlight the best results for different metrics for different methods and distinguish the quality of the methods.

illustrate the working mechanism of AFM in detail. It can be seen from Fig. 9 that F_1, F_2, and F_3 contain different feature information, respectively. F_1 contains a rich set of boundary details. The features of irrelevant changes become less in F_2, but it is still difficult to extract the changed regions. The change features extracted by F_3 lack boundary detail information. With the application of AFM, the network adaptively selects the required boundary detail features and semantic expression features from F_1, F_2, and F_3, which makes the identified changes F_f more discriminative and the boundary more complete.

D. Sensitivity Experiments on BCD Loss

To explore the effect of coefficient λ available in the BCD loss on the training process of the proposed AFNUNet, different values of λ are set on each of the three datasets for experiments. We have presented these results in Table VII. When $\lambda = 0$, the network corresponds to the second baseline “w/o BCD loss” in Section IV-C. The accuracies of all the models using the BCD loss are improved to some extent on all three datasets. When $\lambda = 0.8$, the proposed network achieves the highest $F1$ and IoU on the WHU-CD and SYSU-CD datasets, representing improvements of 0.90% and 1.53%, 0.83% and 1.14%, respectively, as compared to when $\lambda = 0$. The proposed network achieves the highest $F1$ and IoU on the LEVIR-CD dataset when $\lambda = 1.0$, with an improvement of 0.21% and 0.35%, compared to when $\lambda = 0$. This suggests that due to the nature of the dataset itself, the value of λ affects different datasets differently. The WHU-CD and SYSU-CD datasets are more sensitive to the value of λ in the BCD loss.

TABLE VIII
MODEL EFFICIENCY OF DIFFERENT METHODS

Method	Params(M)	FLOPs(G)	It(ms)
FC-EF [24]	**1.35**	**3.63**	**4.75**
FC-Siam-Diff [24]	1.55	5.32	6.69
FC-Siam-Conc [24]	1.35	4.72	6.60
UNet++ [33]	9.16	34.96	8.71
UNet++_MSOF [21]	9.05	34.01	8.03
DDCNN [38]	46.68	177.44	37.87
SNUNet [40]	12.03	54.83	19.08
BIT [30]	12.40	10.63	12.72
MSCANet [55]	16.59	14.80	14.44
SiUNet3+ [45]	27.00	216.72	35.72
AFNUNet	3.34	10.06	7.59

The bold entities highlight the best results for different metrics for different methods and distinguish the quality of the methods.

E. Model Efficiency

Parameters (Params), floating point operations (FLOPs), and inference time (It) are employed as measures of the efficiency of all comparison methods. Params, FLOPs, and It denote the total number of parameters that the model needs to learn during training and the computational cost and time complexity of the model, respectively.

Given a pair of images of size $1 \times 3 \times 256 \times 256$, Table VIII shows Params, FLOPs, and It of all compared methods. The three U-Net-based networks, FC-EF, FC-Siam-Diff, and FC-Siam-Conc, although their structures are simple and high efficiency, combined with the previous performance on the three datasets, apparently do not meet the requirements for accurate identification of changes. Due to a large number of feature transmissions and 3×3 standard convolutions, UNet++, UNet++_MSOF, DDCNN, SNUNet, and SiUNet3+ have a high number of Params and FLOPs and slow It. BIT and MSCANet, which are based on a hybrid CNN-transformer architecture, use efficient decoding strategies, but their ResNet18-based backbone limits their efficiency. In addition, SNUNet applies transposed convolution in the upsampling stage, which introduces more computational cost and increases time complexity. DDCNN and SiUNet3+ are the least efficient, with a last decoder layer of 1024 channels, twice as wide as the other methods in terms of network width.

It is evident from Table VIII that AFNUNet has the lowest number of parameters, computational cost, and time complexity compared to other UNet++, transformer, and UNet3+-based CD methods. This is mostly attributed to the following reasons.

AFNUNet only uses a 3×3 standard convolution at the beginning of each encoder layer and uses an inverted bottleneck structure to improve efficiency. In the decoding stage, AFNUNet uses depthwise convolution instead of standard convolution to effectively reduce Params, FLOPs, and It. Moreover, AFNUNet uses the $Upsample$ method, which does not introduce additional parameters and allows fast upsampling.

V. Conclusion

In this article, we propose an AFNUNet to effectively and efficiently capture the differences in bi-temporal optical RS images. It achieves a fusion of different scale features with low consumption by improving the NestedUNet of the encoder and decoder. Since the final same-scale features extracted by NestedUNet contain different details and semantic information, the network adaptively selects change features from the channel and spatial dimensions by the AFM, thus obtaining changed targets with more refined boundaries. In addition, we introduce the BCD for balancing the effect of changed and unchanged samples for enhancing the accuracy of the network to extract changed information. Experimental results show that AFNUNet achieves better performance on both the building CD datasets WHU-CD and LEVIR-CD as well as the SYSU-CD dataset containing multiple changes in type. We will investigate the CD methods with a broader range of applications in the future and improve the proposed methods to weakly supervised or unsupervised CD methods for satisfying the demands of more diverse scenarios.

References

[1] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

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

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

[4] 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.

[5] B. Desclée, P. Bogaert, and P. Defourny, "Forest change detection by statistical object-based method," *Remote Sens. Environ.*, vol. 102, no. 1/2, pp. 1–11, May 2006.

[6] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS-J. Photogrammetry Remote Sens.*, vol. 80, no. 2, pp. 91–106, Jun. 2013.

[7] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sens. Environ.*, vol. 199, pp. 241–255, Sep. 2017.

[8] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[9] C. Zhang, S. Wei, S. Ji, and M. Lu, "Detecting large-scale urban land cover changes from very high resolution remote sensing images using CNN-based classification," *ISPRS Int. J. Geo- Inf.*, vol. 8, no. 4, Apr. 2019, Art. no. 189.

[10] C. Marin, F. Bovolo, and L. Bruzzone, "Building change detection in multitemporal very high resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2664–2682, May 2015.

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

[12] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.

[13] M. Volpi, D. Tuia, F. Bovolo, M. Kanevski, and L. Bruzzone, "Supervised change detection in VHR images using contextual information and support vector machines," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 20, pp. 77–85, Feb. 2013.

[14] C. Benedek and T. Sziranyi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.

[15] G. Cao, L. Zhou, and Y. Li, "A new change-detection method in high-resolution remote sensing images based on a conditional random field model," *Int. J. Remote Sens.*, vol. 37, no. 5, pp. 1173–1189, Jan. 2016.

[16] P. Lv, Y. Zhong, J. Zhao, and L. Zhang, "Unsupervised change detection based on hybrid conditional random field model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 4002–4015, Jul. 2018.

[17] C. Zhang, G. Li, and W. Cui, "High-resolution remote sensing image change detection by statistical-object-based method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2440–2447, Jul. 2018.

[18] L. Huang, G. Zhang, and Y. Li, "An object-based change detection approach by integrating intensity and texture differences," in *Proc. 2nd Int. Asia Conf. Inform. Control, Autom. Robot.*, 2010, vol. 3, pp. 258–261.

[19] A. Lefebvre, T. Corpetti, and L. Hubert-Moy, "Object-oriented approach and texture analysis for change detection in very high resolution images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2008, vol. 4, pp. 663–666.

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

[21] 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, Jun. 2019, Art. no. 1382.

[22] A. Chouhan, A. Sur, and D. Chutia, "DRMNet: Difference image reconstruction enhanced multiresolution network for optical change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4014–4026, 2022.

[23] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

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

[25] 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.

[26] H. Lee et al., "Local similarity siamese network for urban land change detection on remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4139–4149, 2021.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[28] 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.

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

[30] 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.

[31] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, vol. 9351, pp. 234–241.

[33] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, 2018, vol. 11045, pp. 3–11.

[34] H. Huang et al., “UNet3+: A full-scale connected UNet for medical image segmentation,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 1055–1059.

[35] 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.

[36] E. W. Beals, “Bray-curtis ordination: An effective strategy for analysis of multivariate ecological data,” *Adv. Ecological Res.*, vol. 14. pp. 1–55, 1984.

[37] S. Sun, L. Mu, L. Wang, and P. Liu, “L-UNet: An LSTM network for remote sensing image change detection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8004505.

[38] 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.

[39] 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.

[40] 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.

[41] 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, 2022, Art. no. 3510805.

[42] M. Liu, J. Huang, L. Ma, L. Wan, J. Guo, and D. Yao, “A spatial-temporal-channel attention UNet for high resolution remote sensing image change detection,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4344–4347.

[43] H. Li, F. Zhu, X. Zheng, M. Liu, and G. Chen, “MSCDUNet: A deep learning framework for built-up area change detection integrating multispectral, SAR, and VHR data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5163–5176, 2022.

[44] Y. Du, R. Zhong, Q. Li, and F. Zhang, “TransUNet++SAR: Change detection with deep learning about architectural ensemble in SAR images,” *Remote Sens.*, vol. 15, no. 1, p. 6, Dec. 2022.

[45] B. Zhao, P. Tang, X. Luo, L. Li, and S. Bai, “SiUNet3+-CD: A full-scale connected siamese network for change detection of VHR images,” *Eur J. Remote Sens.*, vol. 55, no. 1, pp. 232–250, Mar. 2022.

[46] J. Mo, S. Seong, J. Oh, and J. Choi, “SAUNet3+CD: A siamese-attentive UNet3 for change detection in remote sensing images,” *IEEE Access.*, vol. 10, pp. 101434–101444, 2022.

[47] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[48] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020 s,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11976–11986.

[49] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.

[50] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “ECA-Net: Efficient channel attention for deep convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.

[51] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[52] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, “A tutorial on the cross-entropy method,” *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[53] 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, May 2020, Art. no. 1662.

[54] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, “A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.

[55] M. Liu, Z. Chai, H. Deng, and R. Liu, “A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[56] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, “Changedetection.net: A new change detection benchmark dataset,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 1–8.

[57] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.

Junwei Li (Member, IEEE) received the M.S. and Ph.D. degrees in control science and engineering from Northwestern Polytechnical University, Xi’an, China, in 2009 and 2013, respectively.

He is currently an Associate Professor with the School of Artificial Intelligence, Henan University, Zhengzhou, Henan, China. His research interests include information fusion, pattern recognition, and deep learning.

Shijie Li received the B.S. degree in computer science and technology from the School of Computer and Information Engineering, Henan University, Kaifeng, China, in 2020. He is currently working toward the M.S. degree in computer technology with the School of Artificial Intelligence, Henan University, Zhengzhou, China.

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

Feng Wang received the Ph.D. degree in pattern recognition and intelligent systems from the School of Automation, Northwestern Polytechnical University, Xi’an, China, in 2019.

He is currently a Lecturer with Weinan Normal University, Weinan, China. So far, he has authored or coauthored more than nine papers, in which eight papers were indexed by EI. His research interests include remote sensing image change detection, deep learning, and image fusion.