# ACMFNet: Asymmetric Convolutional Feature Enhancement and Multiscale Fusion Network for Change Detection

Weipeng Le, Liang Huang, Bo-Hui Tang, *Senior Member, IEEE*, Qiuyuan Tian, and Min Wang

***Abstract*—Existing deep-learning supervised change detection (CD) networks still have room for improvement, as they do not fully utilize multiscale features in their feature extraction, resulting in insufficient feature representation ability and edge blurring problems of the constructed CD networks. In this article, we proposed a very high-resolution (VHR) remote sensing image CD network (ACMFNet) with an asymmetric convolution residual block (ACRB) and multiscale fusion (MSF) to improve the feature representation ability of the CD network and alleviate the edge blurring problem. ACMFNet consists of two main subnetworks: an ACRB feature extraction encoder and an MSF decoder. The ACRB is constructed based on asymmetric convolution and focuses on extraction at the edges of the features and is more robust to rotations, flip distortions, and uneven aspect ratios of the features. In the designed MSF decoder, the fusion feature maps of each level of the decoder are generated by fusing the multiscale encoder feature maps and feature map of the next lowest level of the decoder. MSF contributes to the reconstruction of the edge change area by combining feature information at different scales. It was tested on three public VHR remote sensing image CD datasets, and the proposed method demonstrates the best recall and *F*1-scores, as well as near optimal precision.**

***Index Terms*—Asymmetric convolution, change detection (CD), multiscale fully convolutional Siamese network, multiscale fusion (MSF).**

## I. Introduction

CHANGE detection (CD) refers to the process of identifying changes in a scene from a pair of remote sensing images of the same geographical area acquired at different times [1]. Recently, the number of satellites launched around the world has grown exponentially, and remote sensing satellites and unmanned aerial vehicle platforms can obtain massive numbers of remote sensing images every day [2]. CD is used to effectively mine the information of this massive amount of remote sensing data and has many applications in natural resource monitoring, such as natural disaster assessment [3], land use monitoring [4], urban monitoring [5], and forest monitoring [6]. Multitemporal very high-resolution (VHR) remote sensing images provide the possibility of monitoring land cover changes at a fine scale. However, CD based on VHR remote sensing images is still challenging for the following reasons: 1) the complexity of objects in a remote sensing scene; and 2) different imaging conditions, such as sensor characteristics and illumination changes. Therefore, developing a CD model to detect real changes is of great importance.

As an effective means of extracting VHR remote sensing image features, deep-learning methods offer a new avenue for CD based on VHR remote sensing images. Deep learning can learn high-level features from various remote sensing images automatically, unlike traditional methods that rely on artificial features. Convolutional neural networks (CNNs) [7] can extract abstract and robust features and are extensively used in the field of CD. Based on CNNs, supervised deep-learning CD can be mainly categorized into three types: difference map-based methods, image patch classification-based methods, and semantic segmentation-based methods. Difference map-based methods first perform CNN feature extraction on remote sensing images from different times to obtain high-level spatial features and then compute the difference map. Finally, thresholding, clustering, and other methods are used to segment and extract the change regions in the difference map. Zhan et al. [8] extracted high-level spatial features by a deep convolutional Siamese network. Zhang et al. [9] optimized a deep cascaded semantic supervision network by using an improved triplet loss function, which enhanced the intraclass separability and interclass discrepancy, and obtained a binary change map by threshold segmentation. However, these methods all need to use thresholds to segment the difference map into change and nonchange regions, which require different threshold selection strategies for different datasets and scenarios. Improper threshold selection often has a great impact on CD results. Image patch classification-based methods transform the CD problem into a pixel classification problem, which classifies the center pixel of the image patch. Daudt et al. [10] proposed two Siamese early fusion CD network frameworks to predict the center pixel of an image patch. Rahman et al. [11] constructed a Siamese neural network, which used a Siamese VGG16 architecture to extract deep features of image patch pairs. Overall, based on image

Manuscript received 21 November 2023; revised 28 January 2024; accepted 25 February 2024. Date of publication 7 March 2024; date of current version 25 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42361054, in part by the Yunnan Fundamental Research Projects under Grant 202201AT070164, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2023JJ60561, and in part by the "Xingdian" talent support program project. *(Corresponding author: Liang Huang.)*

The authors are with the School of Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650093, China (e-mail: leweipeng@stu.kust.edu.cn; kmhuangliang@kust.edu.cn; tangbh@kust.edu.cn; tianqykm@163.com; 20212101059@stu.kust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3372386

patch CD methods, the similarity of image patches is judged by the shared Siamese network process, which needs to construct image patches for each pixel of the image. However, this often leads to considerable spatial information redundancy.

The most widely used CNN supervised CD network framework is based on semantic segmentation methods, which are more efficient than image patch classification methods. The idea is to use CNN to perform pixel-level image segmentation and distinguish the change regions from the background. Currently, most CD models based on semantic segmentation methods have evolved from semantic segmentation encoder—decoder architectures, such as Unet [12], UNet++ [13], and feature pyramid networks (FPNs) [14]. CD models based on semantic segmentation methods can be divided into two types. The first type of method cascades the bitemporal images and inputs them into a single-branch encoder–decoder semantic segmentation architecture for pixel-level segmentation. For example, Peng et al. [15] proposed an end-to-end CD method based on UNet++, which cascaded the registered bitemporal images as the input of the UNet++ network and obtained the final change map by multiside output fusion. However, based on image cascading methods, image cascading into the network for training often results in spatial feature entanglement and mismatch problems. The other method is to construct a Siamese encoder–decoder CD architecture. To extract multiscale features and perform effective feature fusion, Chen et al. [16] fully utilized the spatiotemporal information in the image, combined UNet with an attention mechanism, and designed a new Siamese CNN for CD. Zheng et al. [17], using the FPN architecture, constructed a fully convolutional multitask difference enhancement Siamese CD network. Lv et al. [18] proposed that a multiscale information attentional module was embedded in the backbone of UNet to achieve a multiscale information fusion task of bitemporal images. Lv et al. [19] designed a novel neural network with a spatial–spectral attention mechanism and multiscale dilation convolution modules. Li et al. [20] proposed an attention-guided multilayer feature aggregation network. However, the feature extraction and multiscale utilization of these CD networks still need to be further optimized. There are many methods used to enhance feature extraction based on deep learning CD, such as depth separable convolution [21], pruning [22], and the ghost module [23]. Inspired by the good contour-keeping advantages of horizontal and vertical convolution and its strong robustness to rotated or flipped distorted data [24], this article introduces asymmetric convolution into CD. Spatial detail information is a necessary condition for accurately detecting changes, and how to strengthen the exploration of spatial information is a key issue [25]. The underlying encoder features contain more detailed information (i.e., texture and color) and provide a more direct and instructive representation than the deeper features [26]. The downsampling operation of the CD network will inevitably result in the loss of spatial details. The use of intermediate features to build feature context has been shown to aid model representation learning [27]. In [28], [29], and [30], the CD network uses skip connections to mitigate the loss of spatial detail. Zhang et al. [31] used a dense skip connections mechanism to indirectly fuse features at different scales. The multiscale fusion method used in this article fuses different level and scale feature maps. Compared to other fusion methods, the fusion method in this article utilizes more shallow features to enhance the detection of target detail information.

We design a new deep CNN CD model consisting of two parts: feature extraction and multiscale fusion (MSF). This article makes the following main contributions.

1) To improve the feature extraction ability of the CD network architecture encoder and obtain multiscale features with more discriminative change information ability, we design an asymmetric convolution residual block (ACRB), which replaces the standard $3 \times 3$ convolution in the residual block with asymmetric convolution, to enhance the feature expression ability of the residual block and effectively improve the performance of the network model without increasing inference time and overhead.
2) To reconstruct the edge change area, we construct a multiscale feature fusion decoder that fuses different scale feature information.
3) The ACMFNet network model constructed in this article can be used as a baseline for supervised CD models, and it is compared with other CD models. It was tested on three public VHR remote sensing image CD datasets, and the proposed method demonstrated the best recall and $F1$-scores, as well as near-optimal precision.

The rest of this article is organized as follows. Section II is an introduction to the related work. Section III presents the proposed network architecture. Section IV presents the setup and results of all experiments. Section V is the discussion. Finally, Section VI concludes this article.

## II. Related Work

### A. Feature Extraction

The SNN [32] refers to a coupled architecture of two neural networks. The SNN receives two sample data points as input, generates high-level spatial features through a dual-branch neural network and calculates their similarity. The shared-weight dual-branch SNN can highlight similar regions in same-source bitemporal remote sensing images, so using SNN to extract multiscale features of bitemporal remote sensing images is a feasible scheme. Currently, the field of computer vision has many mature basic network backbones, including ResNet [33], VGG [34], and ViT transformer [35]. Combining these basic network backbones with SNNs to construct multiscale feature extraction modules is a common method of supervised deep-learning CD. Chen et al. [36] used VGG16 and ResNet50 as the basic backbone of CD Siamese network feature extraction. Zhang and Shi [37] used a CNN to learn remote sensing image domain feature parameters from remote sensing images based on the VGG16 network backbone. Wang et al. [38] constructed a shared-weight Siamese conjoined network, whose backbone used DeepLabV2 to extract multiscale features. Liu et al. [39] and Chen et al. [40] used the SE-ResNet backbone, extracting multiscale features of bitemporal images. Chen et al. [41] used a transformer encoder to model context in a compact token-based

spatiotemporal domain, effectively simulating context information within the spatiotemporal domain. Zhang et al. [42] used the transformer architecture to construct CD encoding modules. In addition, Zhang et al. [43] used depthwise separable convolution to replace the standard $3 \times 3$ convolution kernel, reducing the number of parameters of CD model training. By applying ghost convolution to a multiscale Siamese CNN network architecture, Lei et al. [44] improved the network performance and further decreased the number of network parameters. The above methods have improved model feature extraction to some extent, but the number of parameters, inference time, and performance still have room for improvement.

### B. Feature Fusion

Different feature layers of the deep-learning CNN architecture show different characteristics. While shallow features retain more positional and detailed information, they lack semantic information. After multiple convolutions downsampling, deep features have stronger semantic information but lose detailed perception ability. Effectively fusing deep features and shallow features is a key factor in improving CD model performance. Zhou et al. [45] used dense connections to fuse shallow and deep layers. Bao et al. [46] used the FPN network architecture to fuse feature information of different network layers and improve CD efficiency. Zheng et al. [47] used UNet as the basic architecture of a CD network and designed a cross-layer block to fuse multiscale features and context information of different levels. Song et al. [48] proposed a Siamese UNet attention mechanism-based CD network to solve the problem of edge detail loss during CD. Fang et al. [49] combined a Siamese network with a UNet++ network architecture to alleviate the problem of neural networks losing positional information in deep layers. Although the above methods improved feature fusion, the feature fusion method still needs to be optimized. In addition to the above multiscale fusion methods. Zhou et al. [50] proposed a novel multiple feature fusion model termed attention multihop graph and multiscale convolutional fusion network. Guo et al. [51] proposed a global spatial feature representation model based on the encoder–decoder structure with channel attention and spatial attention to learn the global spatial features.

## III. Method

### A. Overview

Currently, many deep-learning CD methods have gradually solved the problems of traditional methods, but their ability to deal with complex backgrounds and identify the edges of change regions still needs to be improved. As a result, this article proposes a CD network named ACMFNet to distinguish real change regions from complex backgrounds. Fig. 1 shows the overall flowchart of the proposed ACMFNet model.

The model mainly consists of two parts: an asymmetric convolution residual feature extraction encoder and a multiscale feature fusion decoder. Let us assume that two VHR remote sensing images are taken at different times $T_0$ and $T_1$, with $C$ being the number of channels, and $H$ and $W$ being the length and width of the images, respectively.

1) First, $T_0$ and $T_1$ are input into the asymmetric convolution residual feature extraction encoder to obtain the features of each encoder layer, which can be expressed as $\{f_i^n \mid i = 0, 1\}$, $n = 1, 2, 3, 4, 5$.
2) Then, by the channel stacking operation of feature maps, different time features of the same scale are fused, and the result is expressed as $Em$, $m = 1, 2, 3, 4, 5$. To make full use of different levels of feature information, the encoding layer features of $Em$ are fused by a multiscale feature fusion decoder, and its decoding layer features are expressed as $Dp$, $p = 1, 2, 3, 4$.
3) Finally, through a deep supervision hybrid loss function, network parameters are optimized.

### B. Asymmetric Convolution Residual Feature Extraction Encoder

In CD tasks, the shape, size, and orientation of detection objects present different changes. Based on CD models based on CNNs, most of the convolution basic units are $3 \times 3$ convolutions, but $3 \times 3$ convolutions have difficulty extracting asymmetric features, such as vertical or horizontal edges and textures. However, asymmetric convolutions, such as $1 \times 3$ or $3 \times 1$ convolutions, can better capture the asymmetric features in images. Therefore, in this article, we use asymmetric convolution blocks instead of $3 \times 3$ convolutions in the feature extraction encoder. The asymmetric convolution block used consists of three convolutions of sizes $1 \times 3$, $3 \times 1$, and $3 \times 3$, and each convolution unit performs convolution, batch normalization (BN), and rectified linear unit (ReLU) processes. The three convolution units together act as an asymmetric convolution block to perform convolution operations on the feature map. Due to the additivity of two-dimensional convolution, the asymmetric convolution block is equivalent to adding a $1 \times 3$ or $3 \times 1$ convolution based on a $3 \times 3$ convolution to enhance the extraction of asymmetric features, such as vertical or horizontal edges and textures. The process is shown in Fig. 2.

To enhance the capability of extracting multiscale features, we design an ACRB based on the asymmetric convolution block. Compared with ordinary residual convolution blocks, ACRBs can provide more patterns and directions of feature learning, which enable the model to learn richer and more diverse features, thereby improving the performance of CD models. The ACRB used in this article is shown in Fig. 3. The model consists of two asymmetric convolution blocks, two BNs, two activation functions, and a skip connection. First, the input feature map goes through an asymmetric convolution, and its asymmetric features are extracted. Then, it goes through BN, and the ReLU activation function is applied; afterward, it goes through another layer of asymmetric convolution and BN to extract deeper features. Finally, the output feature map of the previous asymmetric convolution is added to the output feature map of the next asymmetric convolution elementwise, and the feature map, which goes through the ReLU activation function, is output. The calculation process of each level

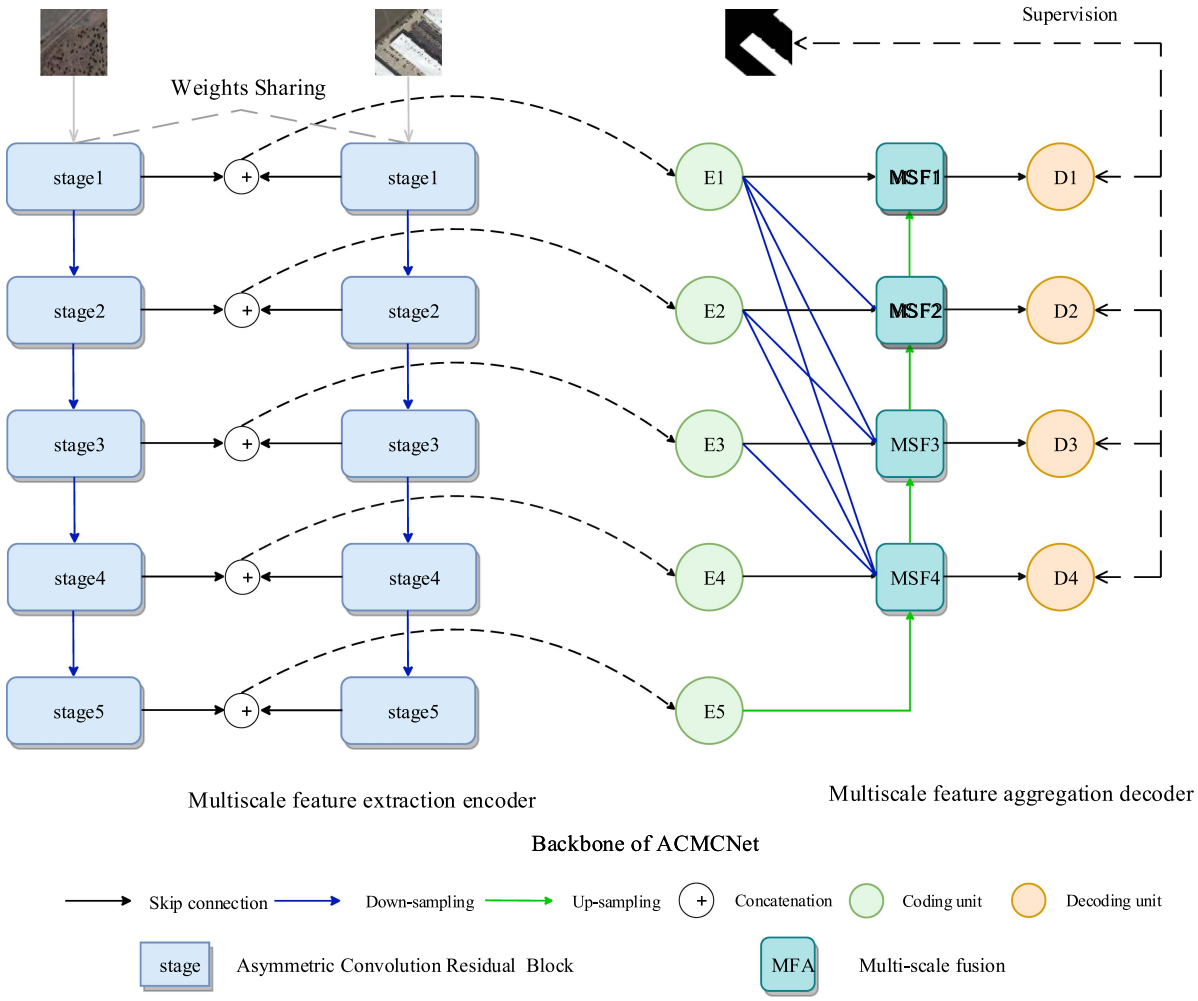


Fig. 1. Proposed overall network architecture of ACMFNet. Black arrows indicate skip connections. Blue arrows indicate downsampling and green arrows indicate upsampling. Concatenation means stacking by channel. The coding unit is the encoder unit after cascading. The decoding unit is the decoder unit. Stage is the asymmetric convolution residual block. MSF is the multiscale feature fusion.

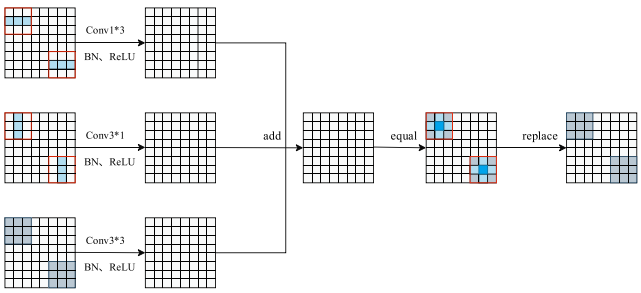


Fig. 2. Asymmetric convolutional block.

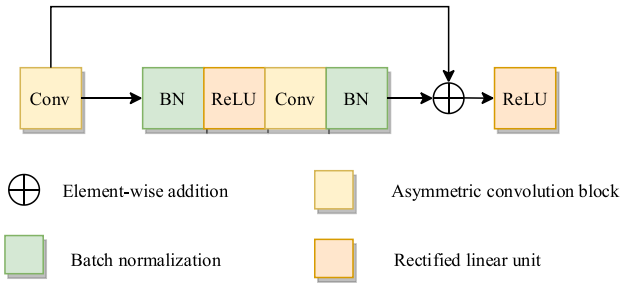


Fig. 3. Structure of ACRB.

of the ACRB in the encoder is expressed by the following formula:

$$\begin{aligned}\mathrm{Conv}_{\mathrm{AC}}\left(f_i^n\right) = {} & R\left(B_N\left(\mathrm{Conv}_{3\times1}\left(f_i^n\right)\right)\right) \\ & + R\left(B_N\left(\mathrm{Conv}_{1\times3}\left(f_i^n\right)\right)\right) \\ & + R\left(B_N\left(\mathrm{Conv}_{3\times3}\left(f_i^n\right)\right)\right) \end{aligned} \quad (1)$$

$$\begin{aligned}\mathrm{output} = {} & R\left(\mathrm{Conv}_{\mathrm{AC}}\left(f_i^n\right)\right. \\ & \left. + B_N\left(\mathrm{Conv}_{\mathrm{AC}}\left(R\left(B_N\left(\mathrm{Conv}_{\mathrm{AC}}\left(f_i^n\right)\right)\right)\right)\right)\right) \end{aligned} \quad (2)$$

where $\mathrm{Conv}_{3\times1}$ is the $3\times1$ convolution kernel, $\mathrm{Conv}_{1\times3}$ is the $1\times3$ convolution kernel, $\mathrm{Conv}_{3\times3}$ is the $3\times3$ convolution kernel, and $\mathrm{Conv}_{\mathrm{AC}}$ is the asymmetric convolution block. $B_N$

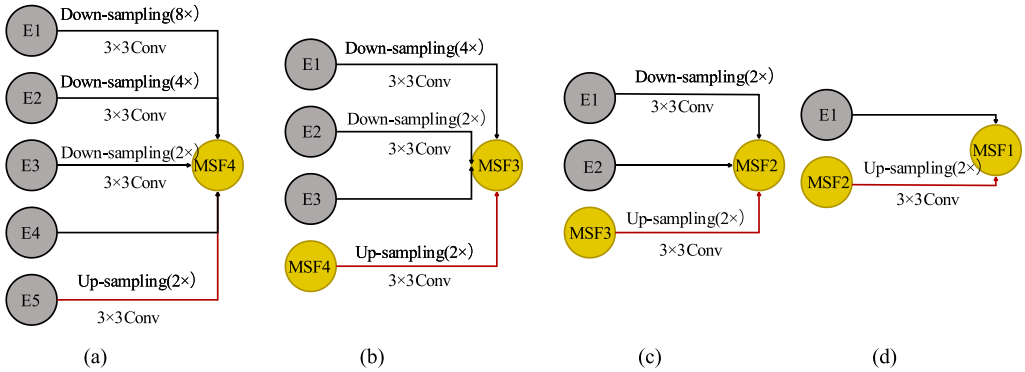


Fig. 4. Multiscale fusion decoder. (a) MSF4. (b) MSF3. (c) MSF2. (d) MSF1.

and $R$ denote BN and ReLU, respectively. $f_i^n\{i = 0, 1\}$ denotes the input feature of the $n$th encoding layer, $n = 1, 2, 3, 4, 5$.

### C. Multiscale Feature Fusion Decoder

Combining high-level semantic information with low-level detailed information, as well as information at different scales, can effectively reconstruct the edge change area. Therefore, designing a multiscale feature fusion network is crucial. We construct a multiscale feature fusion decoder, which aggregates multiscale encoder feature maps and decoder feature maps at the next level to obtain aggregated feature maps at each level of the decoder, thus capturing the detailed and semantic information of different levels of feature maps of the encoder. The CD network decoder constructed in this article is shown in Fig. 4.

First, we obtain the fused features $E_1$, $E_2$, $E_3$, $E_4$, and $E_5$ of each level of the encoder. Then, we input them into the MSF decoder to obtain the MSF features, $\mathrm{MSF}_1$, $\mathrm{MSF}_2$, $\mathrm{MSF}_3$, and $\mathrm{MSF}_4$. Finally, we convert $\mathrm{MSF}_1$, $\mathrm{MSF}_2$, $\mathrm{MSF}_3$, and $\mathrm{MSF}_4$ into two-channel feature maps by $1 \times 1$ convolution and upsample them to the original image size. The output feature maps are $D_1$, $D_2$, $D_3$, and $D_4$. The following describes how $D_1$, $D_2$, $D_3$, and $D_4$ are constructed.

$\mathrm{MSF}_4$ has features derived from the encoder–decoder feature fusion operation, which includes encoder features at the same scale $E_4$, encoder features at smaller scales $E_1$–$E_3$, and decoder features at the next level $E_5$. Each level of the feature map has different operations. In the first step, $E_1$–$E_5$ all need to perform a $3 \times 3$ convolution operation to convert their channel number to 64, while $E_5$ also needs to perform bilinear interpolation upsampling to enlarge its spatial resolution by a factor of 2. $E_1$–$E_3$ need to perform global pooling downsampling to reduce their spatial resolution by factors of 8, 4, and 2, respectively. After these operations, $E_1$–$E_5$ have 64 channels, and their spatial resolutions are consistent. In the next step, we stack the feature maps of $E_1$–$E_5$ along the channel dimension to form a 320-channel feature map. To decrease the number of channels, a $3 \times 3$ convolution is applied, resulting in 64 channels. In the last step, we convert $\mathrm{MSF}_4$ into a two-channel original image size $D_4$ by $1 \times 1$ convolution and eightfold bilinear interpolation upsampling operations. Here is the formula used to calculate the $D_4$ feature map

$$\begin{aligned} D_4 = \mathrm{UP}_8(\mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(P_8(E_1)) \\ \oplus \mathrm{Conv}_{3\times3}(P_4(E_2)) \\ \oplus \mathrm{Conv}_{3\times3}(P_2(E_3)) \oplus \mathrm{Conv}_{3\times3}(E_4) \\ \oplus \mathrm{Conv}_{3\times3}(\mathrm{UP}_2(E_5)))) \end{aligned} \tag{3}$$

where $\mathrm{UP}_8$ represents eightfold bilinear interpolation upsampling operations, $\mathrm{Conv}_{1\times1}$ is $1 \times 1$ convolution, $\mathrm{Conv}_{3\times3}$ is $3 \times 3$ convolution, $P_8$ represents eightfold global pooling downsampling operations, $P_4$ represents quadruple global pooling downsampling operations, $P_2$ represents double global pooling downsampling operations, and $\mathrm{UP}_2$ represents double bilinear interpolation upsampling operations.

Similarly, the fusion formula for $D_3$, $D_2$, and $D_1$ is as follows:

$$\begin{aligned} D_3 = \mathrm{UP}_4(\mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(P_4(E_1)) \oplus (\mathrm{Conv}_{3\times3}(P_2(E_2)) \\ \oplus \mathrm{Conv}_{3\times3}(E_3) \oplus \mathrm{Conv}_{3\times3}(\mathrm{UP}_2(\mathrm{MSF}_4)))) \end{aligned} \tag{4}$$

$$\begin{aligned} D_2 = \mathrm{UP}_2(\mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(P_2(E_1)) \oplus \mathrm{Conv}_{3\times3}(E_2) \\ \oplus \mathrm{Conv}_{3\times3}(\mathrm{UP}_2(\mathrm{MSF}_3)))) \end{aligned} \tag{5}$$

$$D_1 = \mathrm{Conv}_{1\times1}\left(\mathrm{Conv}_{3\times3}(E_1) \oplus \mathrm{Conv}_{3\times3}(\mathrm{UP}_2(\mathrm{MSF}_2))\right) \tag{6}$$

where $\mathrm{UP}_4$ represents quadruple bilinear interpolation upsampling operations.

### D. Loss Function

In CD tasks, the change regions are much smaller than the unchanged regions, which causes a class imbalance problem in the model training process. Therefore, in this article, we use a hybrid loss function composed of dice loss [52] and weighted cross-entropy loss [53] to mitigate the impact of class imbalance. In the proposed ACMFNet network, the final decoder layer outputs feature $D_1$–$D_4$. We use a deep supervision [15], [54], [55] strategy to optimize the loss for each output layer feature. Each layer has a weight parameter $W_i$ ($i = 1, 2, 3, 4$). Therefore, we set $W_i$ to (1, 1, 1, 1). The total loss function is defined as

follows:

$$L_{\text{total}} = \sum_{i}^{4} w_i l_{\text{side}}^i \tag{7}$$

where $l_{\text{side}}^i$ ($i = 1, 2, 3, 4$) denotes the hybrid loss function used for each output, which is defined as follows:

$$l_{\text{wce}} = \frac{1}{H \times W} \sum_{j=1}^{H \times W} \text{weight}[c] \cdot \left( \log \left( \frac{\exp(\hat{y}[j][c])}{\sum_{l=0}^{1} \exp(\hat{y}[j][l])} \right) \right) \tag{8}$$

$$l_{\text{dice}} = 1 - \frac{2 \cdot Y \cdot \text{soft}\max(\hat{Y})}{Y + \text{soft}\max(\hat{Y})} \tag{9}$$

$$l_{\text{side}}^i = l_{\text{wce}}^i + l_{\text{dice}}^i \tag{10}$$

where $l_{\text{wce}}^i$ ($i = 1, 2, 3, 4$) denotes the weighted cross-entropy loss. Weight is the weight parameter, $c$ is either 0 or 1, indicating unchanged pixels and changed pixels, respectively, and $\hat{Y}$ represents the change map of each layer, denoted by a set $\hat{Y} = \{\hat{y}_j, j = 1, 2, \ldots, H \times W\}$. $\hat{y}_i$ represents a binary element of $\hat{Y}$. $l_{\text{dice}}^i$ ($i = 1, 2, 3, 4$) is the dice loss, and $Y$ represents the ground truth.

## IV. Experimental Results and Discussions

### A. Dataset Descriptions

To thoroughly evaluate the performance of the ACMFNet model, we conducted both quantitative and qualitative assessments on three publicly available datasets: LEVIR–CD [56], WHU–CD [57], and GZ–CD [58]. The detection of changes in urban buildings is the main objective of the LEVIR–CD dataset. It includes 637 pairs of VHR remote sensing images, each with a spatial resolution of 0.5 m, a spatial size of 1024 × 1024 pixels, and three spatial channels. In addition to the real change regions, the dataset also includes many pseudochanges caused by seasonality and illumination. We further cropped the 1024 × 1024-pixel images into 256 × 256-pixel patches and removed the unchanged image pairs. The resulting sets were 3107 pairs for training, 433 pairs for validation, and 923 pairs for testing. WHU–CD is an urban disaster building aerial image CD dataset. The dataset is composed of a pair of VHR remote sensing images, which have a spatial resolution of 0.075 m and a spatial size of 32 507 × 15 354 pixels. Similarly, we cut the original image into 256 × 256-pixel patches and removed the unchanged image pairs. Next, we partitioned the cropped WHU–CD dataset into a training set with 1403 pairs, a validation set with 173 pairs, and a test set with 173 pairs. GZ–CD is a Guangzhou suburban building CD dataset. The dataset contains 19 pairs of VHR images with a spatial resolution of 0.55 m and a channel number of 3. Their sizes range from 1006 × 1168 pixels to 4936 × 5224 pixels. Likewise, we crop the images into 256 × 256 pixels. After removing the unchanged image pairs, we divide the dataset into a training set/757 pairs, a validation set/94 pairs, and a test set/94 pair. The image pairs and labels of the three datasets are shown in Fig. 5.


Fig. 5. (a) are LEVIR-CD dataset image pairs and labels. (b) are WHU-CD dataset image pairs and labels. (c) GZ–CD dataset image pairs and labels.

### B. Comparative Methods

We evaluated the performance and effectiveness of the proposed ACMFNet model by comparing it with several recent supervised CD models. The comparative models are FC–EF [28], FC-Siam-conc [28], FC-Siam-diff [28], CDNet [59], SNUNet [49], DSIFN [54], BIT [41], L-Unet [60], and WNnt [61]. The following is a brief description of these methods.

1) FC–EF is a fully convolutional CD network based on UNet feature encoding and decoding. It uses a single-branch network to cascade images for change region detection.
2) FC-Siam-conc is a CD network framework based on a Siamese fully convolutional architecture. Its encoder branches perform feature concatenation at the same scale and then stack them with the decoder layer at the same scale through skip connections.
3) FC-Siam-diff is also a Siamese fully convolutional CD network framework, but its encoder branches calculate feature difference maps at the same scale and then stack them with the decoder layer at the same scale through skip connections.
4) CDNet is a deconvolution network for street scene CD.
5) SNUNet is a Siamese network variant that features dense connections both within and across the encoder and decoder layers.
6) DSIFN uses attention modules and deep supervision mechanisms to effectively fuse original image features and image difference features, improving the performance of CD.

TABLE I
ACMFNET NETWORK STRUCTURE PARAMETERS

| Original data size dimension: T0(256,256,3), T1(256,256,3) | | | | | | |
|---|---|---|---|---|---|---|
| Encode | Siamese network branch 1 | (ACRB,3,32) | (ACRB,32,64) | (ACRB,64,128) | (ACRB,128,256) | (ACRB,256,512) |
| | Siamese network branch 2 | (ACRB,3,32) | (ACRB,32,64) | (ACRB,64,128) | (ACRB,128,256) | (ACRB,256,512) |
| Decode | Multiscale feature fusion | – | ($MSF_4$,320,64) | ($MSF_3$,256,64) | ($MSF_2$,192,64) | ($MSF_1$,128,64) |

7) BIT represents bitemporal images as a small number of semantic tokens and uses a transformer encoder to model context in a compact token-based spatiotemporal domain, effectively simulating context information within the spatiotemporal domain.
8) L-Unet is a deep multitask learning framework able to couple semantic segmentation and CD using fully convolutional long short-term memory (LSTM) networks. UNet-like architecture that models the temporal relationship of spatial feature representations using integrated fully convolutional LSTM blocks on top of every encoding level.
9) WNnt is a new W-shaped dual-Siamese branch hierarchical network for HRRS image CD named W-shaped hierarchical network. WNet first incorporates a Siamese CNN and a Siamese transformer into a dual-branch encoder to extract multilevel local fine-grained features and global long-range contextual dependencies.

### C. Experimental Details

The PyTorch framework was used to implement our model, which was trained on an NVIDIA GeForce RTX 3090 GPU with a memory of 24 GB. The input images were 256 × 256 pixels with three spatial channels, and the model used a hybrid loss function comprising dice loss and weighted cross-entropy loss. We set the learning rate to $10^{-3}$ and the batch size to 8 and optimized the parameters using the AdamW. optimizer. The model was trained for 100 epochs, and the learning rate was adjusted by a decay factor of 0.5 every 8 iterations. We adopted precision, recall, and $F1$-score as the quantitative measures to evaluate our experiments.

Table I gives the structural parameters and description of the ACMFNet network. The meanings of the parameters are as follows: (256, 256, 3) means that the size of the feature map is 256 × 256 and the number of channels of the feature map is 3. (ACRB, 3, 32) means that the ACRB transforms a feature map with input channel number 3 into a feature map with output channel number 32. ($MSF_1$, 128, 64) indicates that the multilayer features are aggregated into a 128-channel feature map, and then the 128-channel feature map is downscaled to a 64-channel map afterward.

### D. Analysis and Discussion of Experimental Results on the LEVIR–CD Dataset

The quantitative evaluation metrics and visualization results of the ACMFNet model on the LEVIR-CD dataset are displayed in Table II and Fig. 6. Compared to the other models, the FC–EF, FC-Siam-conc, and FC-Siam-diff models had poor performance in the three evaluation metrics. The resulting images also had many false positives and false negatives, and the object edges showed irregular shapes. This may be due to the simplicity of these three models, which could not fully extract features at different scales and distinguish real change regions from complex

TABLE II
QUANTITATIVE EXPERIMENTAL RESULTS ON THE LEVIR–CD DATASET

| Methods | Precision | Recall | F1-score |
|---|---|---|---|
| FC—EF | 78.32% | 66.12% | 71.71% |
| FC-Siam-conc | 85.39% | 78.27% | 81.67% |
| FC-Siam-diff | 84.63% | 75.45% | 79.78% |
| CDNet | 89.93% | 86.85% | 88.36% |
| SNUNet | 88.18% | 88.12% | 88.15% |
| DSIFN | **94.56%** | 80.82% | 87.15% |
| BIT | 89.21% | 89.05% | 89.13% |
| L-UNet | 91.43% | 88.42% | 89.90% |
| WNet | 91.66% | 89.63% | 90.60% |
| **ACMFNet** | 90.71% | **90.92%** | **90.82%** |

TABLE III
QUANTITATIVE EXPERIMENTAL RESULTS ON THE WHU–CD DATASET

| Methods | Precision | Recall | F1-score |
|---|---|---|---|
| FC—EF | 83.95% | 81.83% | 82.88% |
| FC-Siam-conc | 82.85% | 85.13% | 83.97% |
| FC-Siam-diff | 82.73% | 81.89% | 82.31% |
| CDNet | 82.72% | 88.42% | 85.47% |
| SNUNet | 90.85% | 89.87% | 90.36% |
| DSIFN | **95.85%** | 81.80% | 88.27% |
| BIT | 85.39% | 91.18% | 88.19% |
| L-UNet | 72.74% | 89.53% | 80.27% |
| WNet | 92.12% | 90.14% | 91.09% |
| **ACMFNet** | 91.53% | **91.21%** | **91.37%** |

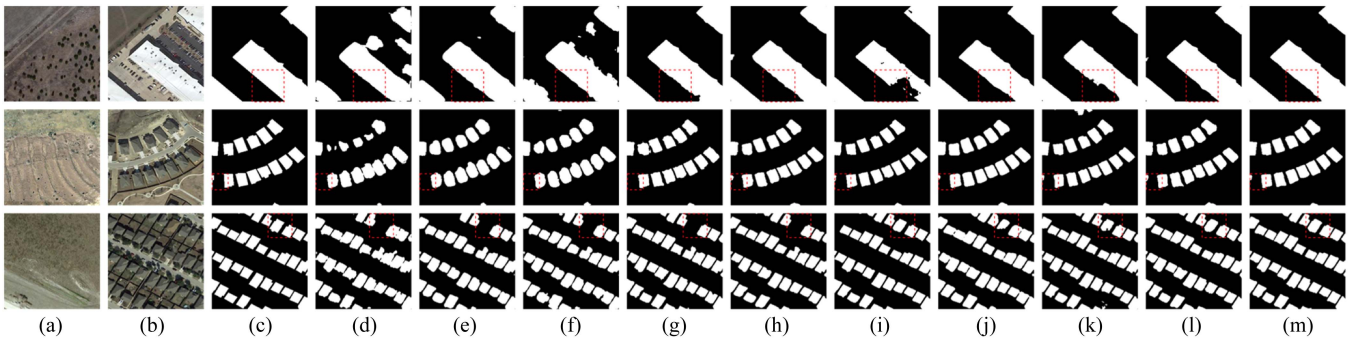


Fig. 6. Visualization results of the LEVIR–CD dataset. (a) Prechange image. (b) Postchange image. (c) Ground truth label. (d) FC–EF result. (e) FC-Siam-conc result. (f) FC-Siam-diff result. (g) CDNet result. (h) SNUNet result. (i) DSIFN result. (j) BIT result. (k) L-Unet result. (l) WNet result. (m) ACMFNet result.

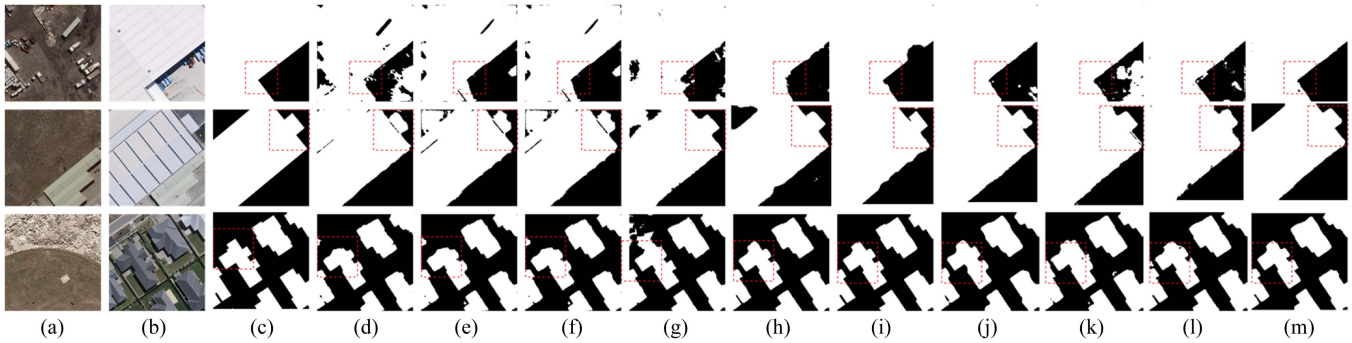


Fig. 7. Visualization results of the WHU–CD dataset. (a) Prechange image. (b) Postchange image. (c) Ground truth label. (d) FC–EF result. (e) FC-Siam-conc result. (f) FC-Siam-diff result. (g) CDNet result. (h) SNUNet result. (i) DSIFN result. (j) BIT result. (k) L-Unet result. (l) WNet result. (m) ACMFNet result.

objects. The CDNet model achieved a significant improvement in precision, with an $F1$-score of 88.36%. The predicted result images also had no obvious irregular shapes at the boundaries. However, as shown by the red boxes in the second row of CD result images Fig. 6(d)–(g), all four CD models mentioned above failed to effectively extract small objects at the image boundaries. This is because these models cannot effectively fuse features at different scales, leading to the loss of small objects. As shown by the third row of pre and postchange images in Fig. 6(a) and (b), there was a changed building that had a different spectral feature from other buildings. The red boxes in the third row of CD result images in Fig. 6(d)–(h) show that in Table II, the first five models could not fully identify the changes in objects with different spectral differences. The DSIFN and BIT models were able to effectively identify changes in objects with different spectral differences. The DSIFN model used attention modules and deep supervision mechanisms to effectively fuse original image features and image difference features, while the BIT model used the transformer to effectively simulate contextual information within the spatiotemporal domain. Although the DSIFN model achieved the best precision of 94.56%, its recall was low. As shown by the first row of Fig. 6(i), there were many false negatives in the DSIFN model result image. From Table II, it can also be seen that the $F1$-score of the ACMCNet model is 0.92% and 0.22% higher than L-Unet and WNet, respectively. As shown in Table II, the proposed ACMFNet model achieved a suboptimal precision of 90.71% and optimal recall and $F1$-scores of 90.92% and 90.82%, respectively. As shown in Fig. 6(k), the proposed model detected most change regions, such as some small buildings and object edge areas. This is because the ACMFNet model used asymmetric convolution to extract vertical or horizontal edges and textures and other asymmetric features, enhancing the ability to extract edge features of the network. It also fully fuses features at different scales, thus effectively detecting subtle changes. Moreover, it utilized shallow features to alleviate edge blurring problems.

### *E. Analysis and Discussion of Experimental Results on the WHU–CD Dataset*

Fig. 7 and Table III show the visualization results and quantitative evaluation metrics of the ACMFNet model on the WHU–CD dataset. The FC–EF, FC-Siam-conc, FC-Siam-diff, and CDNet models achieved $F1$-scores of 82.88%, 83.97%, 82.31%, and 85.47%, respectively. As shown in Fig. 7(d)–(g), these four models had many false positives and false negatives in their change result images. The object edges also showed irregular shapes. This may be due to the simplicity of these four models, which could not fully extract features at different scales and distinguish real change regions from complex objects. The SNUNet, DSIFN, and BIT models achieved significant improvements in the $F1$-score compared to the first four models in Table III. Their $F1$-scores were 90.36%, 88.27%, and 88.19%, respectively. As shown in Fig. 7(h)–(j), these three models also detected the most

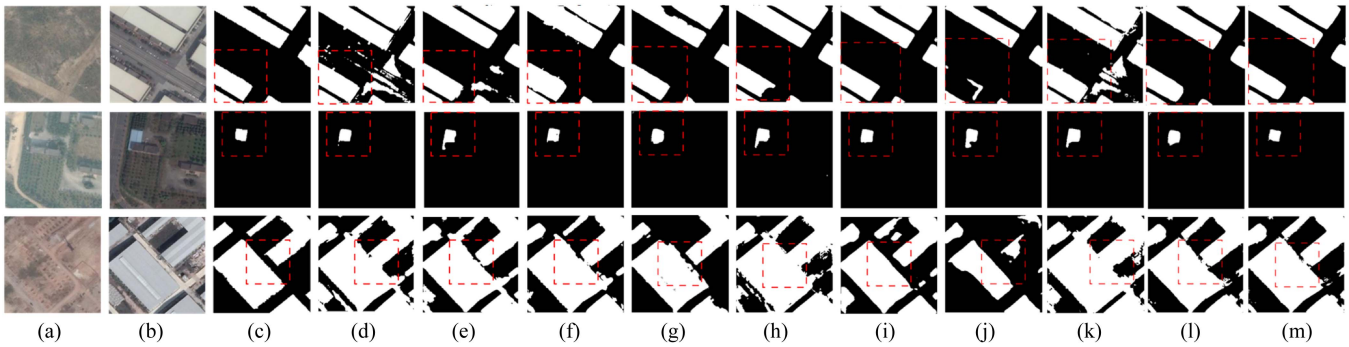


Fig. 8. Visualization results of the GZ–CD dataset. (a) Prechange image. (b) Postchange image. (c) Ground truth label. (d) FC–EF result. (e) FC-Siam-conc result. (f) FC-Siam-diff result. (g) CDNet result. (h) SNUNet result. (i) DSIFN result. (j) BIT result. (k) L-Unet result. (l) WNet result. (m) ACMFNet result.

TABLE IV
QUANTITATIVE EXPERIMENTAL RESULTS ON THE GZ–CD DATASET

| Methods | Precision | Recall | F1-score |
|---|---|---|---|
| FC–EF | 76.76% | 79.30% | 78.01% |
| FC-Siam-conc | 75.19% | 83.02% | 78.91% |
| FC-Siam-diff | 84.19% | 73.23% | 78.33% |
| CDNet | **93.64%** | 64.81% | 76.60% |
| SNUNet | 83.25% | 78.56% | 80.83% |
| DSIFN | 84.83% | 80.51% | 82.61% |
| BIT | 93.17% | 74.48% | 82.78% |
| L-UNet | 81.40% | 71.92% | 76.37% |
| WNet | 87.35% | 83.18% | 84.91% |
| ACMFNet | 85.23% | **86.90%** | **86.06%** |

change regions, but the resulting images also showed some blurring and incompleteness at the boundaries. The SNUNet model achieved suboptimal precision, but as shown by the red border in the first row of Fig. 7(h), it also failed to fully detect the change region at the corner of the boundary. The DSIFN model achieved the best precision, but its recall was low. As shown in the first row of Fig. 7(i), there was a large range of missing boundaries in the DSIFN model result image. The BIT model detected the change region at the corner of the boundary, and the boundary was also more complete, but its precision was low, at 85.39%. As shown in the first row of Fig. 7(j), there were some false positives in the BIT model result image. On the $F1$-score metric, ACMCNet has 11.1% and 0.28% higher accuracy than L-Unet and WNet models. Table III shows that the proposed ACMFNet model achieved an optimal recall and $F1$-score of 91.21% and 91.37%, respectively, and a suboptimal precision of 91.53%. As shown in Fig. 7(k), the proposed model detected the most change regions. Compared to other model result images, the object boundary was more regular and complete. This proves that asymmetric convolution feature extraction and multiscale feature fusion can help effectively process object boundary information.

TABLE V
ABLATION EXPERIMENTS ON THE LEVIR–CD DATASET YIELDED QUANTITATIVE EXPERIMENTAL RESULTS

| Methods | Parameters (MB) | Precision | Recall | F1-score |
|---|---|---|---|---|
| Base | **5.41** | 86.52% | 87.03% | 86.78% |
| Base + MSF | 10.99 | 89.20% | 87.42% | 88.30% |
| Base + ACRB + MSF | 15.20 | **90.71%** | **90.92%** | **90.82%** |

### *F. Analysis and Discussion of Experimental Results on the GZ–CD Dataset*

The quantitative and qualitative results of the proposed model on the GZ–CD dataset are shown in Table IV and Fig. 8. The experimental results of precision, recall, and $F1$-score are 85.23%, 86.90%, and 86.06%, respectively. Compared with other experimental models, the proposed method achieves the best accuracy in recall and $F1$-score, with accuracies of 86.90% and 86.06%, respectively. As seen from Table IV, FC–EF, FC-Siam-diff, FC-Siam-conc, and CDnet have $F1$ scores of 78.01%, 78.91%, 78.33%, and 76.60%, respectively. The $F1$ scores of SNUnet, DSIFN, and BIT are 80.83%, 82.61%, and 82.78%, respectively. The $F1$-scores of L-Unet and WNet are 76.37% and 84.91%. Its visualization is shown in Fig. 8. The model proposed in this article can basically detect most of the change regions, and it has fewer false detection regions than other models, and its edge regions are also well reconstructed.

### *H. Ablation Experiment*

To evaluate the effectiveness of the ACRB module and the MSF proposed in this article, we performed ablation experiments on the LEVIR–CD dataset. We used precision, recall, and $F1$-score as quantitative metrics to evaluate the CD model after adding each module. The number of parameters (MB) of the model after applying each module was also computed. We used Siam_unet as the experimental baseline (base). The experimental results of adding each module are shown in Table V. The visualization results are shown in Fig. 9.

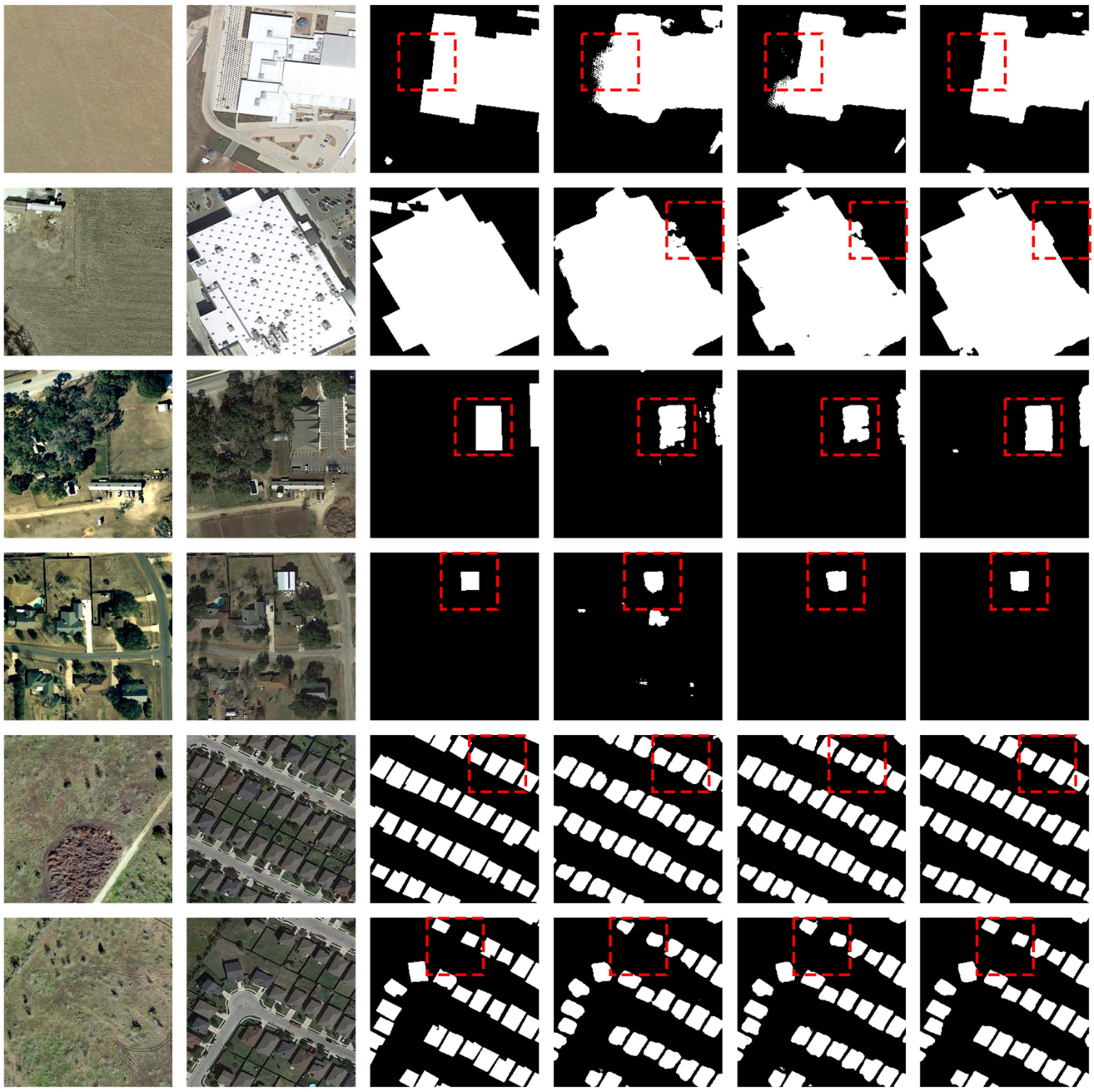


Fig. 9. Visualization results of ablation experiments. (a) shows the prechange image. (b) shows the postchange image. (c) shows the ground truth. (d) shows the base result. (e) shows the base + MSF result. (f) shows the base + ACRB + MSF result.

As presented in Table V, the base model obtained a precision of 86.52%, a recall of 87.03%, and an $F1$-score of 86.78%. The $F1$-score of the network increased by 1.52% after adding the MSF compared to the base model. After adding the ACRB module, the network improved its $F1$-score by 4.04% compared to the base model. This proves the effectiveness of the added MSF and ACRB modules. As shown in Fig. 9, the CD result image (d) of the base model had many false positives, and the object boundaries were also irregular. After adding the MSF, the false positive regions were greatly reduced. After further adding the ACRB module, the object boundaries were clearly more complete and more regular than those of the base model.

Table VI shows the results of the added MSF ablation experiments. Among them, $MSF_4$, $MSF_3$, $MSF_2$, and $MSF_1$ are fused as shown in Fig. 4(a)–(d). Base model whose fusion is UNet skip connection, while $MSF_4$, $MSF_3$, $MSF_2$, and $MSF_1$ are fused feature maps with different scales of multiple layers of the encoder. The experimental results show that the performance of the model can be effectively enhanced by using multilayer

TABLE VI
RESULTS OF THE MSF ABLATION EXPERIMENT

| Model | Precision | Recall | F1-Score |
|---|---|---|---|
| Base | 86.52% | 87.03% | 86.78% |
| Base + $MSF_4$ | 88.94% | 86.76% | 87.68% |
| Base + $MSF_4$ + $MSF_3$ | 89.58% | 87.03% | 88.03% |
| Base + $MSF_4$ + $MSF_3$ + $MSF_2$ + $MSF_1$ | 89.20% | 87.42% | 88.30% |

TABLE VII
QUANTITATIVE EXPERIMENTAL RESULTS OF BOUNDARY IOU ON THE LEVIR–CD DATASET

| Methods | Mean_ boundary IoU (%) | Small boundary IoU (%) | Large boundary IoU (%) | Dense boundary IoU (%) |
|---|---|---|---|---|
| Base | 57.29% | 49.95% | 48.48% | 84.21% |
| Base + MSF | 61.51% | 69.78% | 58.85% | 86.20% |
| Base + ACRB + MSF | **67.90%** | **85.12%** | **70.71%** | **88.15%** |

fusion. After using MSF feature fusion, the accuracy of its experimental results achieved an accuracy improvement of 1.52% on the $F1$-score score.

## V. DISCUSSION

### A. Asymmetric Convolutional Residual Block

In this article, an ACRB is constructed based on AC convolution, the feature extraction is focused on the edge of the object, and it is more robust to the object with rotation, flip distortion, and uneven aspect ratio. The effect of edge extraction in this article is the result of asymmetric convolution residuals and MSF. MSF makes use of shallow detailed information, which helps to reconstruct the edge details. After using MSF, asymmetric convolutional residuals are used to further enhance the extraction of edge features and enhance the performance of the network. We will use quantitative and qualitative experimental methods to measure the extraction effect of the model on the edge of ground objects. Among them, we introduce the evaluation index of the boundary IoU [62] to measure the quality of the partition boundary. The quantitative evaluation is shown in Table VII. At the same time, to visualize the effect of edge detection more intuitively, we display the visual results on small-scale, large-scale, and dense objects. The visualization results are shown in Fig. 9. The calculation formula of the boundary IoU is shown as follows:

$$\text{BoundaryIou} = \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cap (P_d \cap P)|} \tag{11}$$

where boundary regions $G_d$ and $P_d$ are the sets of all pixels within $d$ pixel distances from the ground truth and prediction contours, respectively. The new measure evaluates only mask pixels that are within pixel distance $d$ from the contours. A simpler version with IoU calculated directly for boundary regions $G_d$ and $P_d$ loses information about sharp contour corners that are smoothed by considering all pixels within distance $d$ from the contours. This experiment $d$ is set to 7.

Mean boundary IoU is the value of the average boundary IoU for all buildings in the test dataset. The small boundary IoU indicates the value of the boundary IoU of a small building. The large boundary IoU indicates the value of the boundary IoU of a large building. Dense boundary IoU indicates the value of the boundary IoU of a dense building. Except for mean boundary IoU is calculated for the average of all buildings in the test dataset. The small boundary IoU, large boundary IoU, and dense boundary IoU selected some buildings for calculation. Table VII shows that after using MSF, compared to base, its mean boundary IoU, small boundary IoU, large boundary IoU, and dense boundary IoU achieve 4.22%, 19.83%, 10.37%, and 1.99% accuracy improvement, respectively. This further shows that the edge area is reconstructed to a certain extent after using multiple scales. After continuing to use the ACRB, compared with the base, its mean boundary IoU, small boundary IoU, large boundary IoU, and dense boundary IoU achieve 10.61%, 35.17%, 22.23%, and 3.94% accuracy improvement, respectively. This illustrates that the use of ACRBs can enhance the extraction of edge features and thus reconstruct the edge regions to some extent. For small buildings, our model has the best effect on handling boundaries, while for dense buildings, our model has a poor effect on handling edges. From Fig. 9, we can see that object boundaries are improved to some extent after using MSF and ACRB.

### B. MSF

The fusion strategy used in this article is a multilevel feature fusion strategy, which aims to utilize different levels of underlying features to improve the performance of CD. First, the image is fed into the feature extraction network to get the feature mappings ($E_1$, $E_2$, $E_3$, $E_4$, $E_5$) at different levels. Then, the feature maps ($D_1$, $D_2$, $D_3$, $D_4$) are obtained by multilevel feature fusion strategy, and the core idea of the strategy is to fuse the different dimensional channel feature maps to select the important feature information. It effectively improves the fusion of semantic features and early features, which makes the detailed information of the shallow convolutional layer fully enhanced. Compared with the skip connection of the UNet network model, the multiscale feature fusion decoder used in this article combines the feature maps of different scales of the encoder to utilize the feature information of different scales. Its MSF fusion mode is shown in Fig. 4. In this article, ablation experiments are also performed to demonstrate the effect of our proposed multiscale feature fusion decoder. In Table V, after using MSF, its model improves the $F1$-score evaluation index by 1.52%.

### C. Contrasting With Other Models

The comparative experimental models used in this article are FC–EF, FC-Siam-diff, FC-Siam-conc, CDNet, L-UNet, SNUNet, DSIFN, BIT, and WNet. Analyzing the model architectures, FC–EF, FC-Siam-diff, FC-Siam-conc, and CDNet use a simple convolutional architecture. L-UNet builds on the UNet architecture and uses fully convolutional LSTM blocks to construct the temporal relevance of spatial features. SNUNet

uses a combination of the Siamese network and UNet++. DSIFN adopts a deep supervision strategy and uses channel attention and spatial attention mechanisms to focus on change regions. BIT and WNet use the transformer approach. BIT adopts a combination of CNN and transformer to model the spatiotemporal features in the deep layer and obtain global features. WNet incorporates a Siamese CNN and a Siamese transformer into a dual-branch encoder to extract multilevel local fine-grained features and global long-range contextual dependencies. ACMCNet design an ACRB, which replaces the standard $3 \times 3$ convolution in the residual block with asymmetric convolution, to enhance the feature expression ability of the residual block and effectively improve the performance of the network model without increasing inference time and overhead. The MSF method was also adopted to utilize the encoder multilayer different scale feature information for the reconstruction of the detail information of the change region. Although the ACMCNet model proposed in this article is more effective and can alleviate the edge problem to some extent. However, there are still some problems, compared with FC–EF, FC-Siam-diff, FC-Siam-conc, and CDNet network model, ACMCNet network parameter number is larger, and the running time is longer. And compared to the BIT and WNet models constructed by the transformer method, the feature extraction encoder constructed by ACMCNet is unable to adequately global feature information for long-range spatiotemporal modeling.

## VI. Conclusion

This article proposes a CD network that uses asymmetric convolution feature enhancement and MSF. Extensive comparative and ablation experiments demonstrate the effectiveness of the proposed CD model, and it is also proven that asymmetric convolution feature extraction and multiscale feature fusion can effectively handle object boundary information. In the comparative experiments on the LEVIR–CD, WHU–CD, and GZ–CD datasets, the proposed model achieved optimal recall and $F1$-score accuracy and suboptimal precision accuracy compared to other advanced CD models. In the ablation experiments, adding the MSF and ACRB modules improved the accuracy of the CD model by 4.04% compared to the base model, verifying he effectiveness of the two modules. However, our method still leaves some areas for improvement. For example, the utility of our method in real-world scenarios needs to be improved, and its inference time needs to be shortened. In the future, we will reduce the number of parameters in the model. To further reduce the number of data labels used by the model and improve the applicability of the model, we will build a self-supervised CD model based on the ACMFNet model to achieve the transition from supervised to self-supervised learning.

## References


[1] G. Cheng et al., "Change detection methods for remote sensing in the last decade: A comprehensive review," May 2023, *arXiv:2305.05813*.

[2] Y. Huang, Z. X. Chen, T. Yu, X. Z. Huang, and X. F. Gu, "Agricultural remote sensing Big Data: Management and applications," *J. Integrative Agriculture*, vol. 17, no. 9, pp. 1915–1931, Sep. 2018.

[3] C. Wu et al., "Building damage detection using U-Net with attention mechanism from pre- and post-disaster remote sensing datasets," *Remote Sens.*, vol. 13, no. 5, Feb. 2021, Art. no. 905.

[4] R. D. Johnson and E. S. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *Int. J. Remote Sens.*, vol. 19, no. 3, pp. 411–426, Jan. 1998.

[5] H. Luo, C. Liu, C. Wu, C. Wu, and X. Guo, "Urban change detection based on Dempster–Shafer theory for multitemporal very high-resolution imagery," *Remote Sens.*, vol. 10, no. 7, Jun. 2018, Art. no. 980.

[6] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5407–5423, Sep. 2017.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017.

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

[9] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

[10] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.

[11] F. Rahman, B. Vasu, J. V. Cor, J. Kerekes, and A. Savakis, "Siamese network with multi-level features for patch-based change detection in satellite imagery," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2018, pp. 958–962.

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

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[14] T.-Y. Lin et al., "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[15] 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.

[16] T. Chen, Z. Lu, Y. Yang, Y. Zhang, B. Du, and A. Plaza, "A Siamese network based U-net for change detection in high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2357–2369, Mar. 2022.

[17] 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, Aug. 2022, Art. no. 3775.

[18] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multiscale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Apr. 2023, Art. no. 2501805.

[19] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 4412712.

[20] M. Li, L. Lei, Y. Tang, Y. Sun, and G. Kuang, "An attention-guided multilayer feature aggregation network for remote sensing image scene classification," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3113.

[21] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[22] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," *Proc. Adv. neural inf. proces. syst.*, vol. 2015, pp. 1135–1143, 2015.

[23] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1580–1589.

[24] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1911–1920.

[25] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Symmetrical feature propagation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5536912.

[26] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "RGB-induced feature modulation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5512611.
[27] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 7252–7263, Nov. 2022.
[28] 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.
[29] Y. Zhang, L. Fu, Y. Li, and Y. Zhang, "HDFNet: Hierarchical dynamic fusion network for change detection in optical aerial images," *Remote Sens.*, vol. 13, no. 8, pp. 1440–1461, 2021.
[30] R. Liu, D. Jiang, L. Zhang, and Z. Zhang, "Deep depthwise separable convolutional network for change detection in optical aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1109–1118, Mar. 2020.
[31] X. Zhang et al., "DifUnet ++: A satellite images change detection network based on Unet++ and differential pyramid," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 8006605.
[32] S. Zagoruyko and N. komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4353–4361.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and D. Weissenborn, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.
[36] 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, Nov. 2021.
[37] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.
[38] Z. Wang, C. Peng, Y. Zhang, N. Wang, and L. Luo, "Fully convolutional Siamese networks based change detection for optical aerial images with focal contrastive loss," *Neurocomputing*, vol. 457, pp. 155–167, Oct. 2021.
[39] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[40] P. Chen et al., "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 101–119, May 2022.
[41] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607514.
[42] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5224713.
[43] K. Zhang, I. M. Bello, Y. Su, J. Wang, and I. Maryam, "Multiscale depthwise separable convolution based network for high-resolution image segmentation," *Int. J. Remote Sens.*, vol. 43, no. 18, pp. 6624–6643, Nov. 2022.
[44] 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, Dec. 2022, Art. no. 4507013.
[45] H. Zhou, M. Song, and K. Sun, "A full-scale feature fusion Siamese network for remote sensing change detection," *Electronics*, vol. 12, no. 1, Dec. 2022, Art. no. 35.
[46] T. Bao, C. Fu, T. Fang, and D. Du, "PPCNET: A combined patch-level and pixel-level end-to-end deep network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. vol. 17, no. 10, pp. 1797–1801, Oct. 2020.
[47] Z. Zheng et al., "CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 247–267, May 2021.
[48] L. Song, M. Xia, J. Jin, M. Qian, and Y. Zhang, "SUACDNet: Attentional change detection network based on Siamese U-shaped structure," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102597.
[49] 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, Feb. 2022, Art. no. 8007805.
[50] H. Zhou, F. Luo, H. Zhuang, Z. Weng, X. Gong, and Z. Lin, "Attention Multihop graph and multiscale convolutional fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5508614.
[51] T. Guo, R. Wang, F. Luo, X. Gong, L. Zhang, and X. Gao, "Dual-view spectral and global spatial feature fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5512913.
[52] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. - Int. Conf. 3D Vis.*, 2016, pp. 565–571.
[53] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.
[54] 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, Aug. 2020.
[55] K. Li, Z. Li, and S. Fang, "Siamese NestedUNet networks for change detection of high-resolution satellite image," in *Proc. 1st Int. Conf. Control Robot. Intell. Syst.*, 2021, pp. 42–48.
[56] 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.
[57] 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.
[58] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.
[59] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Streetview change detection with deconvolutional networks," *Auton. Robots.*, vol. 42, no. 7, pp. 1301–1322, 2018.
[60] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, Sep. 2021.
[61] X. Tang, T. Zhang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "WNet: W-shaped hierarchical network for remote-sensing image change detection," *IEEE Trans. Geosci. Remote Sens*, vol. 61, Jul. 2023, Art. no. 5615814.
[62] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15329–15337.

**Weipeng Le** received the B.S. degree in remote sensing science and technology from the School of Surveying, Mapping and Spatial Information Engineering, East China University of Technology, Fuzhou, China, in 2020. He is currently working toward the master's degree in surveying and mapping with the Kunming University of Science and Technology, Kunming, China.

His research interests include change detection and deep learning.

**Liang Huang** received the Ph.D. degree in geodetection and information technology from the Kunming University of Science and Technology, Kunming, China, in 2015.

He is currently working as an Associate Professor with the Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming, China. He has authored more than 80 research papers published in various journals and conferences. His research interests include change detection, semantic segmentation, and object detection.

**Bo-Hui Tang** (Senior Member, IEEE) received the B.S. degree in cartography and GIS from the Wuhan University, Wuhan, China, in 1999, the M.S. degree in cartography and GIS from the China Remote Sensing Satellite Ground Station, Chinese Academy of Sciences, Beijing, China, in 2004, and the Ph.D. degree in cartography and GIS from the Institute of Geographic Sciences and Resources, Chinese Academy of Sciences, in 2007.

He is currently a Professor with the Kunming University of Science and Technology, Kunming, China. His research interests include remote sensing quantitative inversion of surface parameters and remote sensing estimation methods of net surface radiation.

**Qiuyuan Tian** received the B.S. degree in geographic information science from the Tianjin Chengjian University, Tianjin, China, in 2021. She is currently toward the master's degree in surveying and mapping with the Kunming University of Science and Technology, Kunming, China.

Her research interests include change detection and deep learning.

**Min Wang** received the B.S. degree in remote sensing science and technology from the Mianyang Teachers' College, Mianyang, China, in 2021. She is currently working toward the master's degree in photogrammetry and remote sensing with the Kunming University of Science and Technology, Kunming, China.

Her research interests include change detection and deep learning.