A Lightweight Change Detection Network Based on Feature Interleaved Fusion and Bistage Decoding

Mengmeng Wang, Bai Zhu, Jiacheng Zhang, Jianwei Fan, and Yuanxin Ye, *Member, IEEE*

***Abstract*—Deep learning techniques for change detection have undergone rapid development in the past few years. However, it is still a challenge how to reduce massive network parameters and sufficiently fuse bitemporal image features to improve detection accuracy. Therefore, this work proposes a novel and lightweight network based on feature interleaved fusion and bistage decoding (FFBDNet) for change detection. In the encoding stage, considering the application problems caused by a large number of network parameters, we use the more efficient EfficientNet as the backbone to extract the bitemporal image features based on Siamese architecture. To fuse the bitemporal image features and reduce interference from surrounding objects, we propose a feature interleaved fusion module, which can interleave the shared feature information and the difference variance feature information. During the decoding stage, the fused features are split into two groups, and a novel bistage decoding framework is proposed to generate the accuracy change map gradually. Extensive experiments and ablation studies are validated on three public change detection datasets: WHU-CD, LEVIR-CD, and SYSU-CD datasets. Compared to state-of-the-art methods, the experimental results demonstrate that the proposed FFBDNet produces a better balance between performance and model parameters. Specifically, the F1 values obtained for these three datasets are 93.27%, 91.11%, and 80.10%, respectively, and the model parameters of the network are just 2.85 M.**

***Index Terms*—Bistage decoding, change detection (CD), feature interleaved fusion, lightweight network, remote sensing images.**

I. Introduction

CHANGE detection (CD) is the process of extracting and analyzing ground change information by comparing bitemporal remote sensing images at different times in the same geographical area [1], [2]. During the procedure, a semantic label—such as "0" for "unchanged" or "1" for "changed"—is assigned to each pixel. In short, two different temporal high-spatial-resolution images that have been accurately registered [3], [4], [5] are employed to detect changes on the surface. Acquiring high-spatial-resolution satellite images for CD has become more accessible thanks to breakthroughs in remote sensing imaging techniques over the past few decades. High-spatial-resolution images provide ample information but make the CD task more challenging [6]. In a wide range of fields, including resource surveying [7], urban expansion [8], disaster assessment [9], and urban green ecosystem [10], [11], CD is one of the most important applications of remote sensing images.

Depending on whether they require extracting features manually, current CD methods fall into two broad categories: traditional CD methods and deep-learning-based CD methods [12]. Moreover, based on the adopted basic processing unit, the traditional CD methods can be divided into pixel-based and object-based methods [13]. Pixel-based methods usually directly compare the individual pixels to produce the change result [14]. Researchers have performed a great deal of work on pixel-based methods, such as image differencing [15], principal component analysis [16], and change vector analysis [17]. However, pixel-based methods concentrate on the spectral change of an individual pixel and ignore the spatial context information. As a result, the change maps inevitably exhibit salt-and-pepper noises [18]. Different from pixel-based methods, the fundamental unit employed by object-based methods extends to the entire object. The object-based methods can capture the homogeneous pixels belonging to the same objects using spectral [19], textural [20], and spatial features [21]. Although object-based methods can effectively reduce the "salt-and-pepper" noise, suitable parameters are difficult to choose to extract image objects in segmentation algorithms, which means that the error caused by the segmentation would propagate to the predicted change maps [22]. In addition, these traditional methods tend to rely on handcrafted features, which lack robustness in complex scenarios [23]. Therefore, the accuracy of traditional methods is not satisfactory overall.

Over recent years, deep learning techniques have become possible due to the emergence of big data and the constant advances in the performance of computing devices [12], especially convolutional neural networks (CNNs), which have excellent multilayer feature extraction abilities and an effective end-to-end manner [13]; many researchers have incorporated CNNs into several tasks, such as object detection [24], image registration [25], and CD [26], [27]. Generally, there are two main categories of current deep-learning-based CD methods: patch-based and image-based [14]. The patch-based methods predict the change category of the central pixel using image patches (such as 3×3 and 5×5) as the network input [28]. Gong et al. [29] proposed a novel CD network that takes

Manuscript received 31 August 2023; revised 21 October 2023; accepted 8 December 2023. Date of publication 20 December 2023; date of current version 10 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42271446 and Grant 41971281, and in part by the Natural Science Foundation of Sichuan Province under Grant 2022NSFSC0537. *(Corresponding author: Yuanxin Ye.)*

Mengmeng Wang, Bai Zhu, Jiacheng Zhang, and Yuanxin Ye are with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China (e-mail: wm_gmail@163.com; kevin_zhub@163.com; swjtu_zjc@163.com; yeyuanxin@home.swjtu.edu.cn).

Jianwei Fan is with the School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China (e-mail: fanjw@xynu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3344635

each local neighborhood of a pixel as the network input. Lei et al. [30] proposed a CD method based on stacked denoising autoencoders. This method employs the difference image to estimate the multiscale patch image. To address the problem of choosing the optimal patch size, Wang et al. [31] proposed a feature-regularized mask DeepLab and obtained competitive performance. Although patch-based methods typically do not need much data to train on, there are still some drawbacks. On the one hand, the performance of CD is substantially impacted by the appropriate patch size, which is difficult to establish. Besides, the patch-based methods use a neighboring patch overlap strategy, which takes a long time and makes extensive use of memory [32].

To overcome the above limitations, image-based CD methods have progressively adopted a fully convolutional network to conduct an end-to-end pixelwise prediction [14]. After achieving this milestone, numerous CD methods were proposed [23], [33]. For example, by considering two different image input patterns, Daudt et al. [34] proposed three different CD network models, namely, FC-EF, FC-Siam-diff, and FC-Siam-conc. To further obtain more feature information and produce accurate change maps, Peng et al. [35] designed a multiple side-output fusion module that can fuse multiscale feature maps and proposed an MSOFNet based on UNet++ [36]. To overcome the problem of inaccurate boundary identification, Chen et al. [33] proposed an edge-guided network that focuses on prior information on boundaries and the integrity of change region. To filter background noise and improve the detection accuracy of the change, a feature hierarchical differentiation (FHD) model for CD was proposed by Pei and Zhang [37]. In the FHD model, the dual-branch features are adaptively fused to obtain discriminative feature information and achieve excellent results.

The attention mechanism has the capability to enhance relevant feature information and suppress background noise information by weighting the feature maps [38]. Therefore, numerous researchers have introduced the attention mechanism to CD tasks to obtain the discriminative feature information of bitemporal images [39], [40]. Jiang et al. [32] proposed an attention-guided full-scale feature aggregation network that uses the attention mechanism to alleviate feature redundancy and achieve accurate results. Chen and Shi [39] proposed a spatial–temporal attention network (STANet) for CD by analyzing the spatial–temporal connection and the multiscale attention representation. Similarly, Zhang et al. [40] proposed a deeply supervised image fusion network (DSIFN) for CD after introducing channel and spatial attention to fuse feature information from various domains. Chen et al. [41] designed a method that uses a dual-attention strategy to obtain more discriminating feature information to enhance the model's performance at recognition. Yang et al. [42] proposed a multiscale attention and edge-aware Siamese network for CD; in this network, a multiscale attention module composed of contour channel attention and convolutional block attention is designed to enhance the edges of the changed regions.

In addition, the Transformer has been introduced into the CD field in some existing works due to its remarkable performance in extracting global feature information. For example, Chen et al. [43] proposed a bitemporal image transformer network (BITNet) that expresses the bitemporal images into a few tokens and uses a transformer encoder to model contexts in the compact token-based space-time. Moreover, considering the advantages of CNNs and Transformer in feature extraction, Chu et al. [44] proposed a dual-branch feature guided aggregation network; the spatial position and semantic features are extracted through the CNNs and the Transformer branches, respectively. The adaptive frequency transformer was utilized by Fu et al. [45] to enhance the differential feature information present in bitemporal images, and they proposed a CNN–Transformer network for CD.

The methods mentioned above have yielded beneficial effects. However, there are still some issues. On the one hand, the spatial details and the semantic information within remote sensing images become richer as the spatial resolution rises [6]. As a result, several problems (such as edge details and object integrity) brought on by influencing factors (like seasonal changes, illumination changes, and building shadows) get progressively worse. Therefore, a more effective feature extraction network and feature fusion strategy should be explored for the CD task. On the other hand, there are often lots of network parameters in the most recent deep-learning-based methods from the past. However, the larger parameter will cause a higher level of complexity in the network and other unpredictable issues (such as overfitting) and take a long time in the training stage [46]. Therefore, a more lightweight network is required to meet practical applications.

To solve the above two problems, a lightweight network based on feature interleaved fusion and bistage decoding (FFBDNet) is proposed for the CD task. First, to reduce the parameter redundancy and computational cost of the network, an efficient and lightweight EfficientNet is employed as the backbone to extract multiscale features. Subsequently, to get more discriminative change-related fusion feature, a feature interleaved fusion module (FIFM) is designed to fuse the bitemporal features. Then, the multilevel features are divided into two groups according to the properties of each level feature. Finally, use the proposed bistage decoding network to generate an accurate change map step by step. The main contributions of this work are summarized as follows.

1) A novel lightweight network with an FIFM and bistage decoding is proposed for CD. The proposed network designs a bistage decoding module that divides the full-level features into two groups to generate the accurate change result step by step.
2) An FIFM is proposed to fuse the bitemporal features generated from the Siamese network. The proposed FIFM can achieve more effective feature fusion and generate more distinguishable fusion features by exploiting the difference variance information and the shared information between bitemporal features.
3) Extensive experiments on three challenging CD datasets are conducted to validate the efficiency of the proposed method. The results of quantitative and qualitative studies show that the proposed method has a lower model parameters and outperforms a number of previous state-of-the-art (SOTA) CD methods.

The rest of this article is organized as follows. In Section II, we introduce the proposed method in detail. Section III describes the experiments on different datasets that will be conducted to

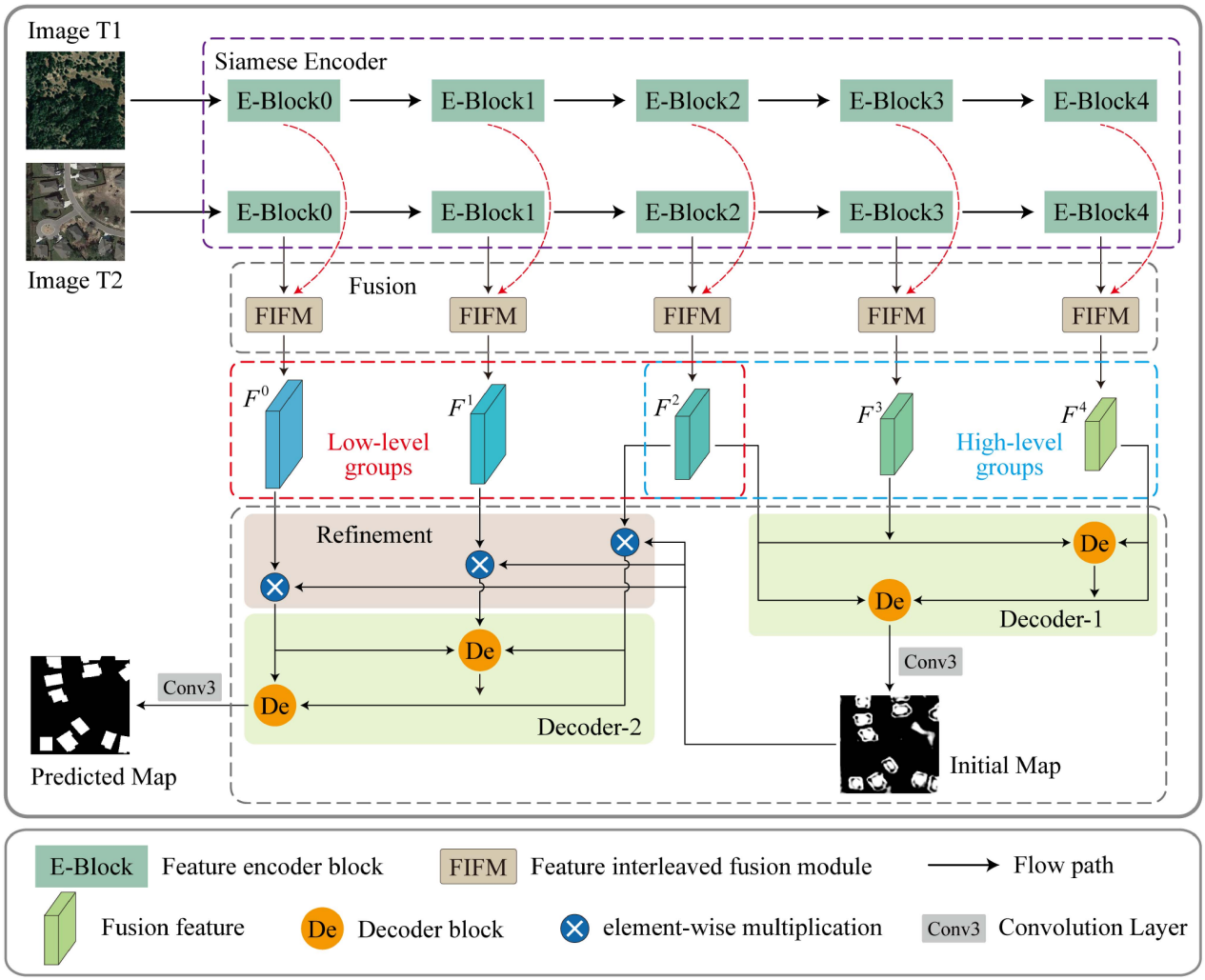

Fig. 1. Framework of the proposed FFBDNet.

validate the proposed FFBDNet. Finally, Section IV concludes this article.

II. Methodology

Fig. 1 shows the framework of the proposed FFBDNet, which is a standard-coded Siamese architecture. First, the Siamese network's encoding branch extracts the multilevel features of the bitemporal images. After that, the bitemporal features are further fused using the designed FIFM. During this process, the FIFM can combine the difference variance feature and the shared feature between bitemporal features to achieve a more effective fusion. In the decoding stage, to alleviate the semantic gap between the low-level and high-level features and improve the accuracy of CD, the fused full-level features are divided into two groups. Subsequently, the two groups' split features are decoded to generate the accurate change map progressively [37].

A. Feature Encoder

In the feature encoder, the input bitemporal images of CD are compatible with the Siamese network structure [47]. Therefore, we use a Siamese network design with two weight-sharing branches to extract multilevel features. As for the backbone, we use EfficientNet-B4 as the feature encoder. EfficientNets are a family of models obtained using neural architecture search to design the new baseline network. Specifically, the EfficientNets use an effective compound coefficient to balance network width, resolution, and depth to obtain better performance. Compared to the Visual Geometry Group Net [48] or Residual Network [49], the EfficientNets show better accuracy and efficiency on computer vision tasks [50].

As shown in the encoding block of Fig. 1, from left to right, a pair of bitemporal images $I_1, I_2 \in \mathbf{R}^{C\times H\times W}$ are used as the input of the Siamese network to extract multilevel features. C, H, and W represent the number of bands, height, and width of the input image, respectively. In this article, only some convolutional stages from the original EfficientNet-B4 have been used to extract features. Specifically, we only use the first five convolutional stages from the original EfficientNet-B4. The different scale feature maps that each convolutional stage makes are represented as layers with different colors in Fig. 1. As a result, each branch of the Siamese network generates five

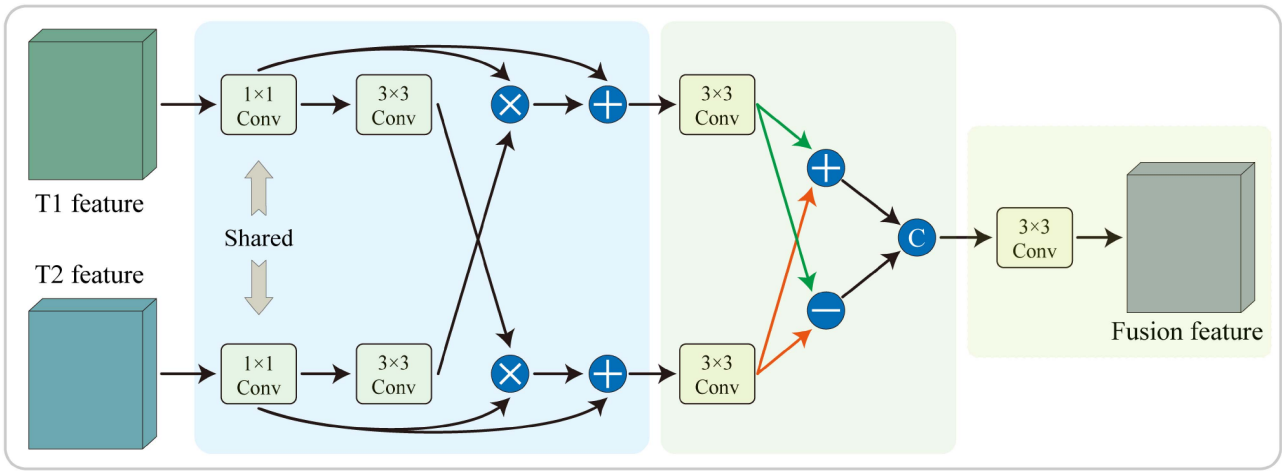

Fig. 2. Structure of the proposed FIFM module.

feature maps, written as $f_1^i, f_2^i, i \in \{0, 1, 2, 3, 4\}$, respectively. Specifically, the size of $f_1^0, f_1^1, f_1^2, f_1^3$, and f_1^4 are 1/2, 1/2, 1/4, 1/8, and 1/16 of the input image size, respectively, and the same for f_2^i. The channels of the five feature maps are 48, 24, 32, 56, and 112, respectively.

B. Feature Interleaved Fusion Module

After bitemporal feature encoding, how to fuse the bitemporal features to obtain the representative features for CD is the core issue [51]. To merge the bitemporal image features, current networks usually use the fusion strategy called "concatenation" or "subtraction" [34], [52]. Even though these fusion strategies help improve the accuracy of CD results, more effective fusion methods are needed to explore. To address the fusion problem of bitemporal features, this article developed a novel FIFM that can generate more discriminative fusion features. Unlike current fusion modules, which focus only on single fusion strategies, the proposed FIFM follows a three-step procedure. First, the bitemporal image features are enhanced across interleave fusion. Next, it extracts the shared representation features and difference variation features from the enhanced bitemporal features. Finally, the two distinct features are concatenated to generate discriminative fusion features for CD.

The proposed FIFM module's structure is shown in Fig. 2; consider the bitemporal image features $f_1^i \in \mathbf{R}^{C_i \times H_i \times W_i}$ and $f_2^i \in \mathbf{R}^{C_i \times H_i \times W_i}$. In this example, C_i, H_i, and W_i represent the channel numbers, height, and width of the ith feature level, respectively. First, a shared 1×1 convolutional layer is used to project f_1^i and f_2^i into a single feature space. Subsequently, the two convolution features are fed into a 3×3 convolutional layer. In this way, two feature maps are obtained, which can be used to enhance the feature maps. Formally, the two feature attention maps are described as follows:

$$\begin{cases} w_1^i = \delta(Conv_3(Conv_1(f_1^i))) \\ w_2^i = \delta(Conv_3(Conv_1(f_2^i))) \end{cases} \tag{1}$$

where w_1^i and w_2^i are the feature attention maps, δ is the sigmoid function, $Conv_1$ is a convolution layer with a kernel size of 1×1, and $Conv_3$ denotes a 3×3 convolution layer . Subsequently, the feature attention maps w_1^i and w_2^i can be used to enhance their corresponding feature maps f_1^i and f_2^i. Moreover, a residual connection approach is utilized to merge the improved features with their original features, ensuring the preservation of information pertaining to the latter. The cross-enhanced features of bitemporal images can be written as

$$\begin{cases} \bar{f}_1^i = f_1^i + f_1^i \otimes w_2^i \\ \bar{f}_2^i = f_2^i + f_2^i \otimes w_1^i \end{cases} \tag{2}$$

where $\bar{f}_1^i$ and $\bar{f}_2^i$ represent the cross-enhanced features, $\otimes$ denotes elementwise multiplication operations, and w_1^i and w_2^i are feature attention maps. After getting the cross-enhanced features $\bar{f}_1^i$ and $\bar{f}_2^i$, it is crucial to fuse the bitemporal features with a robust strategy. Every different fusion method has its benefits and drawbacks. For example, suppose that the shared feature information in the bitemporal features is significant, but the difference information is small. In that case, it is reasonable to infer that the corresponding areas are unchanged. On the contrary, if the difference information is significant, but the shared information is small, we may infer that the corresponding regions in the bitemporal images have changed. Therefore, shared feature information and difference feature information have been integrated into this work. In particular, according to the two cross-enhanced features $\bar{f}_1^i$ and $\bar{f}_2^i$, the shared features $\bar{f}_{\text{sh}}^i$ and difference variance features $\bar{f}_{\text{di}}^i$ can be computed by

$$\begin{cases} \bar{f}_{\text{sh}}^i = \bar{f}_1^i + \bar{f}_2^i \\ \bar{f}_{\text{di}}^i = abs(\bar{f}_1^i - \bar{f}_2^i) \end{cases} \tag{3}$$

where $\bar{f}_{\text{sh}}^i$ and $\bar{f}_{\text{di}}^i$ are shared features and difference features, respectively. abs represents the absolute difference between the two cross-enhanced features. Subsequently, in the channel dimension, $\bar{f}_{\text{sh}}^i$ and $\bar{f}_{\text{di}}^i$ are further concatenated. At last, the final fused features are generated by feeding the concatenated features into a 3×3 convolutional layer. This procedure can be denoted as

$$F^i = Conv_3([\bar{f}_{\text{sh}}^i; \bar{f}_{\text{di}}^i]) \tag{4}$$

where F^i denotes the fused features at the ith feature level, $Conv_3$ is the 3×3 convolution, which is then followed by a BN and a ReLU function, and $[;]$ is the channel dimension's concatenation operation.

C. Bistage Decoding

Previous studies [32], [53] have shown that aggregate full-scale feature information based on the UNet3+ network can effectively improve the performance of results. However, these methods based on UNet3+ that directly aggregate the full-scale features would introduce some issues. The semantic difference between low-level and high-level features has become commonly accepted [54]. This difference is particularly pronounced when attempting to aggregate features on a full scale, often leading to discrepancies and confusion within the network. In particular, the high-level features have rich semantic information that helps find areas that have changed, while the low-level features have spatial information that helps refine edge details [55]. On the other hand, the full-scale skip connection is characterized by excessive network parameters, which could be more problematic.

To alleviate these problems, we propose a bistage decoding strategy that can alleviate the confusion problems caused by the semantic gap. As shown in Fig. 1, the proposed bistage decoding module comprises two progressive decoding stages. Specifically, the fused full-level features are divided into two groups: high-level groups are labeled $C_h (C_h = F^2, F^3, F^4)$ and low-level groups are labeled $C_l (C_l = F^0, F^1, F^2)$. It should be noted that the feature F^2 is contained in two different groups simultaneously. The reason is that the feature F^2 is generated from the middle convolution layers. In other words, we argue that the feature F^2 has both spatially detailed information and discriminative semantic information, which can alleviate the semantic gap caused by the direct use of the full-scale skip connection. Subsequently, the whole network is trained in two stages using the features of two groups.

In the first decoding stage, the three features (i.e., F^2, F^3, F^4) of the high-level groups are progressively integrated to obtain the initial change result P_1. Specifically, the proposed decoding stage is built similarly to UNet3+. First, features F^2 and F^4 are resampled to the same scale as the feature F^3. This entails a downsampling for F^2 and, vice versa, an upsampling for F^4. A feature set containing the same number of channels as F^3 is generated through a 3×3 convolution operation. Subsequently, the three same resolution feature maps in the feature set are concatenated. Then, the fusion feature F_D^3 is obtained by a 3×3 convolution. The entire process is formulated as follows:

$$Cat_f = [Conv_3(D(F^2)); Conv_3(F^3); Conv_3(U(F^4))] \quad (5)$$

$$F_D^3 = ReLu(BN(Conv_3(Cat_f))) \quad (6)$$

where D and U represent the downsampling and upsampling, respectively. $Conv_3$ denotes a 3×3 convolution function, $[;]$ is the concatenation operation, BN is the batch normalization, and $ReLu$ denotes rectified linear unit function. By analogy, by replacing F^3 with F_D^3, the fusion feature F_D^2 can be obtained through the same operation. By using this approach, the resulting fusion feature maps not only encompass multilevel feature information but also exhibit a reduced semantic gap and require fewer network parameters. The reason for this is that each feature group, differently from UNet3+, contains only three levels of features. After F_D^2 is obtained, a 3×3 convolution with a channel size of 1 followed by a sigmoid function is employed to obtain the initial change map P_1, which is formulated as

$$P_1 = \delta(Conv_3(F_D^2)) \quad (7)$$

where P_1 is the initial change map, δ denotes the sigmoid function, and $Conv_3$ is a 3×3 convolution with a channel size of 1. Then, the obtained initial change map P_1 is used as a feature attention map to refine the three low-level feature maps, which are denoted as

$$\bar{F}^i = F^i \otimes P_1 \quad (8)$$

where $F^i, i \in 0, 1, 2$, is the feature in the low-level group, $\otimes$ denotes the elementwise multiplication operation, and $\bar{F}^i$ represents the refined features. In the second decoding stage, the three refined features in the low-level group undergo the same decoding process as in the first stage, resulting in the generation of the final change result P_2.

D. Loss Function

CD can be interpreted as a task of binary classification with two labels: "unchanged" and "changed." Given that the binary cross-entropy (BCE) function is widely employed in binary classification tasks and shows good performance, the BCE function is used as the loss function as follows:

$$\mathcal{L}_{\text{bce}}(t, p) = -\frac{1}{N}\sum_{i=1}^{N}[t_i \log(p_i) + (1 - t_i)\log(1 - p_i)] \quad (9)$$

where N represents the total pixel numbers, and t and p represent the truth label and the predicted change map, respectively. $t_n \in \{0, 1\}$ is the value of position n in t (i.e., $t_i = 0$ and $t_i = 1$ represent unchanged and changed categories, respectively). p_n and $1 - p_n$ denote the predicted probabilities of changed categories and unchanged categories, respectively; $p_n \in [0, 1]$.

III. Experiments and Analysis

We first provide a detailed illustration of the three challenging datasets and the evaluation metrics. Then, we present an overview of eight SOTA CD methods, followed by a description of the relevant experimental setting. Subsequently, we provide a comparison and analysis of the results obtained from a series of experiments. Finally, ablation experiments are conducted to validate the efficacy of the FIFM and the bistage decoding module.

A. Datasets

1) *WHU-CD [56]:* The WHU-CD dataset includes portions of the region in New Zealand. The bitemporal images

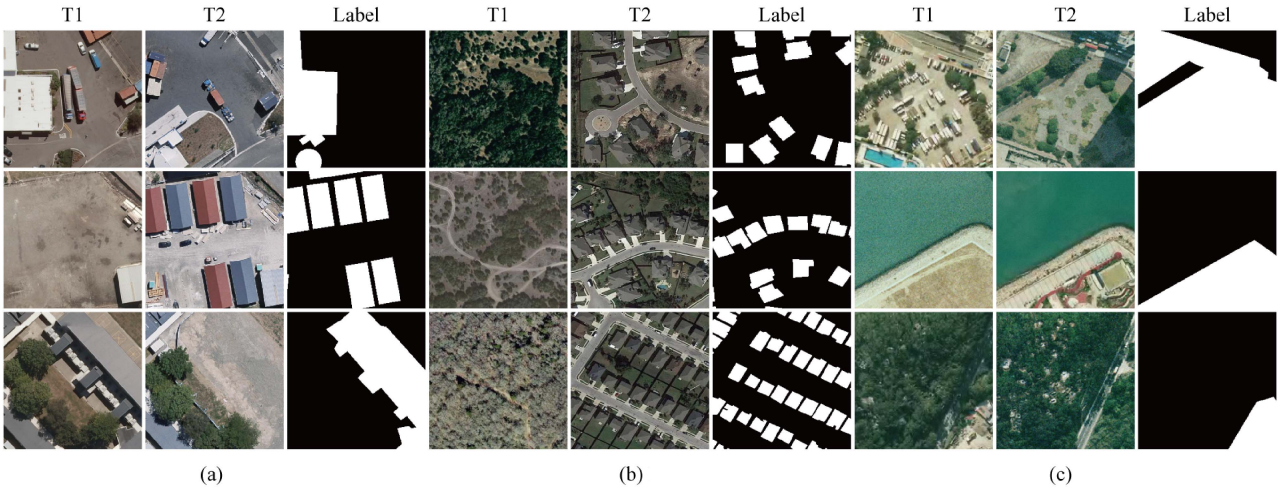

Fig. 3. Sample images from three datasets, where "T1" and "T2" represent bitemporal images and "Label" represent truth map. (a) WHU-CD. (b) LEVIR-CD. (c) SYSU-CD.

were captured immediately after the magnitude 6.3 earthquake in 2011 and again after it was reconstructed in 2016. This dataset with a spatial resolution of 0.2 m and 32 507 × 15 354 pixel resolutions. Like previous studies [26], [44], the original bitemporal images are divided into 256 × 256 patch size without overlapping and obtained 6096, 762, and 762 image pairs for training, validation, and testing, respectively. Fig. 3(a) shows the sample images of the WHU-CD dataset.

2) *LEVIR-CD [39]:* The LEVIR-CD dataset is collected from Google Earth within a time span between 2002 and 2018. It comprises 637 image pairs with a 0.5-m spatial and 1024 × 1024 pixel resolutions. The LEVIR-CD dataset focuses on building with different types of changes. It contains numerous pseudo-changes due to light and season, making it a challenging dataset for CD. Chen and Shi [39] provided a standard dataset division for the LEVIR-CD dataset. The original images are cropped into patch size. The number of image pairs is 7120, 1024, and 2048 for the training, validation, and test, respectively. Fig. 3(b) shows the sample images of the LEVIR-CD dataset.

3) *SYSU-CD [57]:* Shi et al. [57] have recently released the challenging dataset known as SYSU-CD. This dataset consists of 20 000 pairs of images with 0.5-m spatial resolution and 256 × 256 spatial size. In contrast to WHU-CD and LEVIR-CD, which only focus on building CD, the SYSU-CD dataset contains multiple change types, such as roads, buildings, ships, and croplands. According to the official set, the number of pairs for training, validation, and testing is 12 000, 4000, and 4000, respectively. Some sample images of the SYSU-CD dataset are shown in Fig. 3(c).

B. Evaluation Metrics

To conduct a comprehensive analysis of the performance of the proposed method, four popular evaluation metrics have been adopted, including precision, recall, F1, and intersection over union (IoU). To be more specific, precision refers to the ratio of changed pixels that are successfully detected in contrast to the total number of changed pixels that are detected. A higher precision value indicates a lower commission error. The ratio of changed pixels correctly detected to total ground truth pixels is expressed as recall. The omission error is lower when the recall value is larger. F1 metric is a comprehensive evaluation of the model's performance that can be calculated from the harmonic average of precision and recall. IoU is the ratio between the overlap area of the predicted result and the ground truth and their union. The metrics described above are formulated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{TN}} \tag{11}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{13}$$

where TP stands for the total number of true positives, FP for the total number of false positives, and FN for the total number of false negatives.

C. Comparison Methods

To evaluate the effectiveness of the proposed method, eight SOTA methods are selected to make a comparison, including FC-EF [34], FC-Siam-diff [34], FC-Siam-conc [34], STANet [39], DSIFN [40], SNUNet [52], BITNet [43], and LightCDNet [58]. The details are listed as follows.

1) *FC-EF:* This model is a UNet-based single-stream method. First, the channel dimension is used to concatenate the bitemporal images. Subsequently, the model receives the concatenated images as input to obtain the change map.

2) *FC-Siam-diff:* The Siamese network has introduced this model. In particular, the bitemporal features obtained from

the Siamese network first undergo a subtraction operation and are then put into the decoder to produce the result.

3) *FC-Siam-conc:* Similar to FC-Siam-diff, the FC-Siam-conc extracts bitemporal features using the Siamese network. In contrast, the FC-Siam-conc employs concatenation operations to fuse the bitemporal features.
4) *STANet:* The STANet presents a spatial–temporal module that focuses on capturing the spatial–temporal information between any points on the space-time continuum to obtain more discriminatory features for CD.
5) *DSIFN:* The DSIFN proposes a deeply supervised difference discrimination network (DDN), which can be improved by introducing change map losses directly to intermediate network layers. Furthermore, the DSIFN introduces a hybrid attention mechanism that combines spatial and channel attention.
6) *SNUNet-32:* The SNUNet is composed of the NestedUNet and the Siamese network. The SNUNet utilizes the dense skip connection strategy to alleviate the loss of information. Considering its efficiency and accuracy, the channel number of the SNUNet is set to 32 in this article.
7) *BITNet:* The BITNet is a novel method that combines CNNs and Transformer. The BITNet expresses bitemporal images as a small number of semantic tokens and model contexts in a token-based space-time.
8) *LightCDNet:* The LightCDNet is a lightweight Siamese network for CD. This method improves the representation of change information by introducing a multitemporal feature fusion combining two-stream features.

D. Implementation Details

Our experiment uses an NVIDIA Geforce RTX 3080Ti with 12 GB of memory with the PyTorch framework for model building and training. AdamW is used as the optimizer, and its initial learning rate is stated as 1e-3, while its weight decay is stated as 1e-4. Each of the methods receives training on the three datasets for a total of 100 epochs. In addition, due to the GPU's limited physical memory, the batch size has been fixed to 8. To ensure fair comparisons, we use their released code and default parameters to compare all the methods in the same experiment environment. In addition, following each training epoch, validation is conducted, and the best validation model is evaluated on the test sets.

E. Experimental Results

1) On the WHU-CD Dataset: The quantitative results on the WHU-CD dataset are displayed in Table I. It is evident that precision, recall, F1, and IoU of the proposed FFBDNet are 93.60%, 92.95%, 93.27%, and 87.39%, respectively, which outperform the other SOTA methods in all of the metrics. According to the results of the comparison methods, the results of the three FC-based methods are relatively low. Specifically, the FC-Siam-conc yields the lowest F1 and IoU with a value of 61.96% and 44.89%, respectively. Although the three comparison methods of the STANet, DSIFN, and SNUNet are better than the three FC-based methods, F1 of these three methods is lower than 89.00%. Moreover, the two comparison methods of the BITNet and LightCDNet obtain an F1 of over 91.00%. LightCDNet achieves the best score among these methods, with F1 and IoU of 91.50% and 84.30%, respectively. Compared to the second-ranked LightCDNet, the proposed FFBDNet improves F1 and IoU by approximately 1.77% and 3.09%, respectively. The quantitative analysis demonstrates that the proposed FFBDNet outperforms the other SOTA methods.

TABLE I
QUANTITATIVE EVALUATION OF EXPERIMENTAL RESULTS OBTAINED FROM DIFFERENT METHODS ON THE WHU-CD DATASET

Methods	Precision (%)	Recall (%)	F1 (%)	IoU (%)
FC-EF	79.33	74.58	76.88	62.45
FC-Siam-diff	67.55	63.21	65.31	48.75
FC-Siam-conc	48.58	85.49	61.96	44.89
STANet	86.11	88.14	87.11	77.17
DSIFN	85.89	91.31	88.52	79.40
SNUNet	82.63	90.33	86.31	75.92
BITNet	92.71	89.83	91.25	83.91
LightCDNet	92.00	91.00	91.50	84.30
FFBDNet	93.60	92.95	93.27	87.39

Red is used to highlight the best result, and blue is used to highlight the second-best result.

TABLE II
QUANTITATIVE EVALUATION OF EXPERIMENTAL RESULTS OBTAINED FROM DIFFERENT METHODS ON THE LEVIR-CD DATASET

Methods	Precision (%)	Recall (%)	F1 (%)	IoU (%)
FC-EF	85.87	82.22	83.35	72.43
FC-Siam-diff	88.59	80.72	85.37	74.48
FC-Siam-conc	86.18	85.12	87.58	77.91
STANet	83.81	91.00	87.30	77.40
DSIFN	87.30	88.57	88.42	78.09
SNUNet	90.55	89.28	89.91	81.67
BITNet	89.24	89.37	89.31	80.68
LightCDNet	91.30	88.00	89.60	81.20
FFBDNet	92.28	89.98	91.11	83.67

Red is used to highlight the best result, and blue is used to highlight the second-best result.

Fig. 4 shows the visualization prediction maps of the various methods on the WHU-CD dataset. To have better readability, TP (white), FP (red), FN (blue), and TN (black) do a different color to represent each. Among them, FC-EF, FC-Siam-diff, and FC-Siam-conc have yielded the worst results, making it difficult for these three methods to detect building edge shadows effectively. The results of the STANet are unsatisfactory in the sense that they show a heavily jagged edge. A large number of misclassified unchanged pixels and salt-and-pepper noise are present in the SNUNet's output. Although BITNet and LightCDNet perform better than the previously mentioned methods, they still produce a significant number of false positives at the edges of buildings. In contrast, our FFBDNet can maintain the shape of changed regions more accurately, as shown by both the visualization results and the quantitative analysis in Table I.

2) On the LEVIR-CD Dataset: Table II shows that the proposed FFBDNet achieves optimal performance, with precision, F1, and IoU values of 92.28%, 91.11%, and 83.67%, respectively. Compared with the second-ranked SNUNet, the proposed FFBDNet has enhanced F1 and IoU by approximately 1.2%

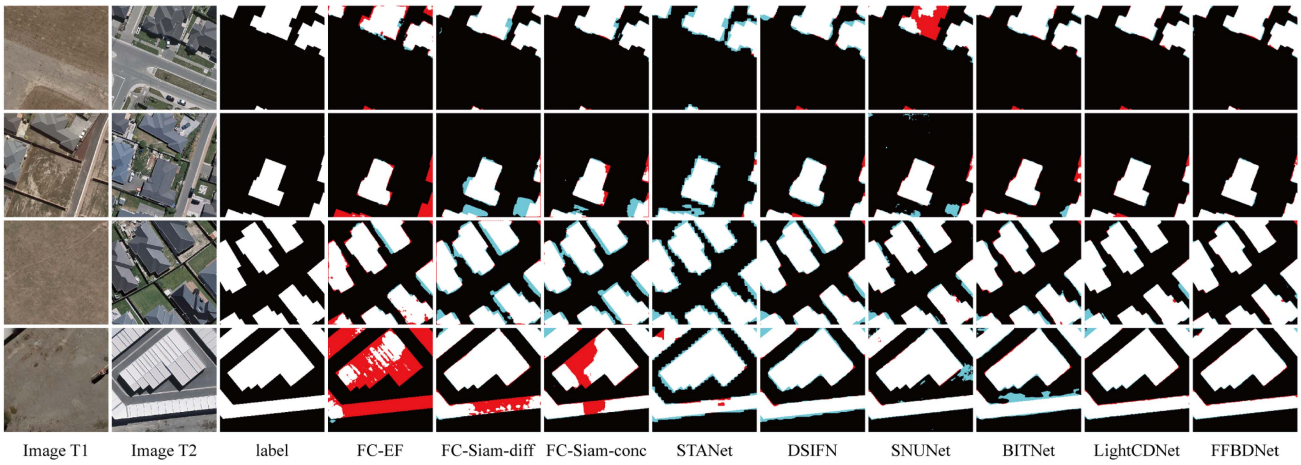

Fig. 4. Visual comparison of the different methods on the WHU-CD dataset.

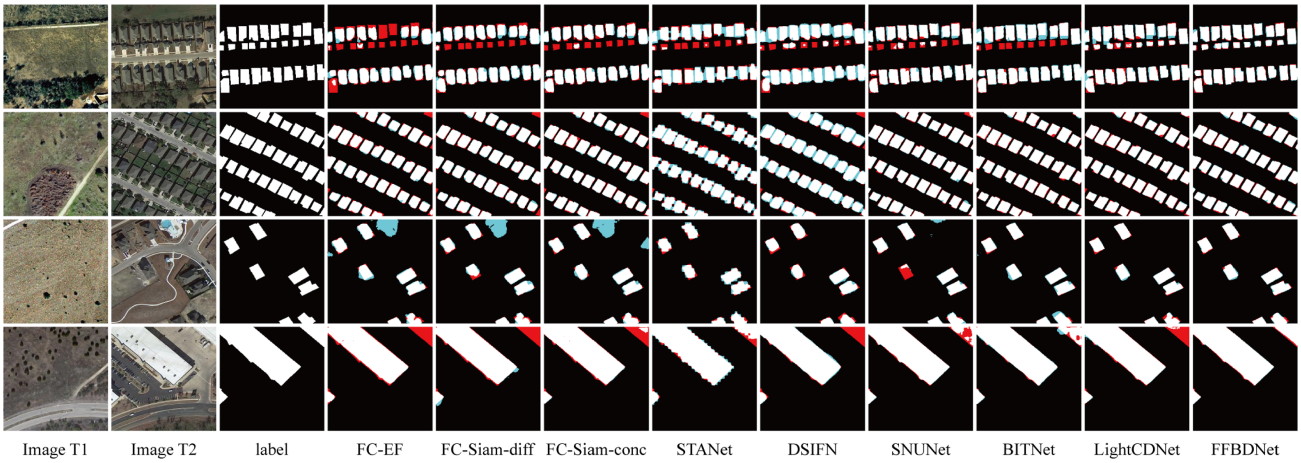

Fig. 5. Visual comparison of the different methods on the LEVIR-CD dataset.

and 2.0%, respectively. While the STANet has the highest recall value at 91.00%, it is only 0.02% higher than the recall value achieved by our proposed FFBDNet. Based on the quantitative results mentioned above, the superior performance of the proposed FFBDNet can be attributed to two factors: the proposed FIFM, which extracts more discriminative fusion features for CD, and the designed bistage decoding module, which generates accurate maps.

As shown in Fig. 5, the LEVIR-CD dataset contains more minor and more numerous changed buildings compared to the WHU-CD dataset, and the visualization results are generated using different methods. In Case I, the changed buildings are small and densely adjacent. The edge details of the results of the STANet and DSIFN are stuck together. The results of the SNUNet and BITNet are improved to some extent, but they suffer more false positives and false negatives than the proposed FFBDNet. In Case III, the appearance of similar colors between the dense buildings and the environment causes severe disturbance. Except for LightCDNet, all the compared methods have difficulty detecting small changed buildings accurately. Besides, LightCDNet only partially detects the changed buildings, leaving ample missed areas. Compared to other methods, the proposed FFBDNet effectively detects minor changed buildings, has fewer false positives, and captures edge details more accurately. The results from all cases demonstrate the superiority of the proposed FFBDNet.

3) On the SYSU-CD Dataset: Table III shows the quantitative results of different methods on the SYSU-CD dataset. It can be seen that the proposed FFBDNet achieves better results than the other SOTA methods in terms of precision, F1, and IoU, except for recall. Among these comparison methods, the three FC-based methods achieve relatively low results. The BITNet and LightCDNet perform relatively better than other comparison methods, but F1 of these two methods is lower than 79.00%. Specifically, the LightCDNet achieves the relatively best performance among these methods, with F1 and IoU of 78.75% and 66.50%, respectively. Compared with the second-ranked LightCDNet, the proposed FFBDNet obtains the highest F1 and IoU of 80.10% and 66.81%, respectively, 1.35% and 1.83% higher than those of the LightCDNet, respectively. It is essential to point out that although the STANet achieves the best recall with 82.73%, its F1 and IoU are relatively lower. In

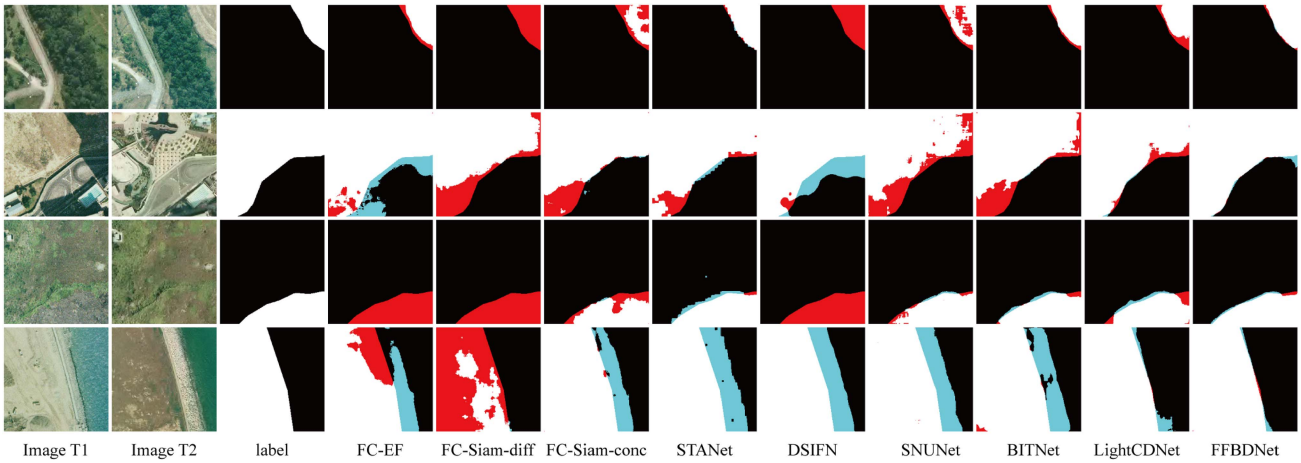

Fig. 6. Visual comparison of the different methods on the SYSU-CD dataset.

TABLE III
QUANTITATIVE EVALUATION OF EXPERIMENTAL RESULTS OBTAINED FROM DIFFERENT METHODS ON THE SYSU-CD DATASET

Methods	Precision (%)	Recall (%)	F1 (%)	IoU (%)
FC-EF	80.16	70.69	75.13	60.17
FC-Siam-diff	78.34	66.13	70.17	55.11
FC-Siam-conc	83.54	69.61	75.94	61.21
STANet	73.33	82.73	77.75	63.59
DSIFN	79.32	73.85	77.46	62.94
SNUNet	82.16	71.33	76.36	61.76
BITNet	80.40	77.09	78.72	64.90
LightCDNet	83.01	74.90	78.75	64.98
FFBDNet	84.48	76.15	80.10	66.81

Red is used to highlight the best result, and blue is used to highlight the second-best result.

TABLE IV
MODEL COMPLEXITY COMPARISONS ON THE WHU-CD DATASET

Methods	Params (M)	FLOPs (G)	F1 (%)
FC-EF	1.35	3.58	76.88
FC-Siam-diff	1.35	4.73	65.31
FC-Siam-conc	1.55	5.33	61.96
STANet	16.89	6.43	87.11
DSIFN	50.46	50.77	88.52
SNUNet	12.03	54.83	86.31
BITNet	3.04	4.35	91.25
LightCDNet	10.75	21.54	91.50
FFBDNet	2.85	7.81	**93.27**

contrast to precision and recall, F1 and IoU are comprehensive evaluations of the performance of the network. Therefore, it is evident that the proposed FFBDNet performs better than other SOTA methods.

The experimental results of a variety of methods are visualized in Fig. 6. It is evident that the three FC-based methods obtain the relatively worst visualization results, which both have a lot of false detection and negative detection. The reason may be that the three FC-based methods' network architecture and fusion mode is simple, which means they cannot deal with datasets with multiple change types. Although the two comparison methods of the SNUNet and BITNet are better than the three FC-based methods, these two methods are still not satisfactory. Among these comparison methods, the LightCDNet achieves relatively better results. However, it can be seen that the proposed FFBDNet obtains the best visual results on multiple change types. Specifically, in the three different cases, the proposed FFBDNet not only has less false detection but also has a smoother edge of changed regions. In contrast, the other comparison SOTA methods have many false positives and negatives. The visual results demonstrate that the proposed FFBDNet has the best performance, which is consistent with the quantitative analysis in Table III.

4) Model Complexity: We further evaluate the model complexity of the proposed FFBDNet on the WHU-CD dataset from two different perspectives: the number of parameters (Params) and the number of floating-point operations (FLOPs). The values of Params and FLOPs are directly correlated with the complexity of a network.

The Params and FLOPs of all compared methods are displayed in Table IV. For an intuitive visualization, the scatterplot of all compared methods is shown in Fig. 7. The FC-EF, FC-Siam-diff, and FC-Siam-conc networks have fewer Params and FLOPs due to the network's straightforward design. However, their results are not acceptable, so they should not be used in practice. The DSIFN, SNUNet, and LightCDNet have not only larger Params and FLOPs but also lower performance than the proposed FFBDNet. As for the proposed FFBDNet, its Params are slightly larger than those of the FC-EF, FC-Siam-diff, and FC-Siam-conc. However, the proposed FFBDNet achieves the best performance on two challenging datasets. In sum, the results of the model analyses show that the proposed FFBDNet offers a better balance between the parameters of the model and its performance.

F. Ablation Experiments

A series of ablation experiments are conducted on the three datasets to evaluate the performance of the proposed FIFM and

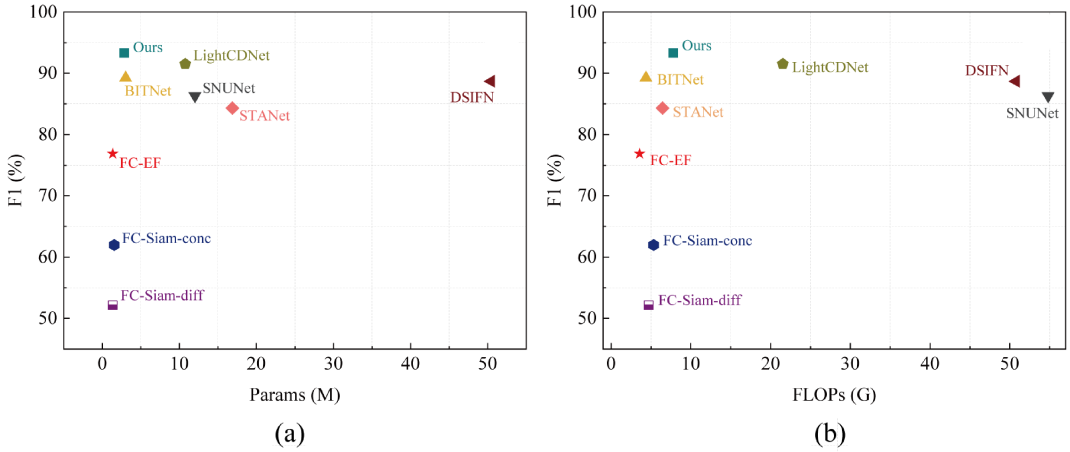

Fig. 7. Scatterplot of different methods' performance. (a) Number of parameters. (b) Floating point of operations.

TABLE V
ABLATION EXPERIMENTAL OF "FIFM" ON THE THREE DATASETS

Methods	WHU-CD		LEVIR-CD		SYSU-CD		Params (M)
	F1(%)	IoU(%)	F1(%)	IoU(%)	F1(%)	IoU(%)	
Baseline + Diff + BID	90.52	82.68	90.04	81.92	78.37	64.43	1.75
Baseline + Conc + BID	91.22	83.85	90.55	82.39	78.92	65.19	3.00
Baseline + DDN + BID	90.76	83.09	90.14	82.06	79.58	65.93	2.80
Baseline + FIFM + BID	**93.27**	**87.39**	**91.11**	**83.67**	**80.10**	**66.81**	2.85

The best score is marked in bold.

the bistage decoding module. All of the ablation experiments use the same training strategies to guarantee accurate comparisons.

1) Ablation Experiment of FIFM: To test the effectiveness of the proposed FIFM, we replace the FIFM module with other fusion strategies like difference fusion [34], concatenation fusion [34], and DDN fusion [40]. First, we define the Siamese encoder based on EfficientNet-B4 without any other module as the "Baseline." The difference fusion mode is represented as "Diff," the concatenation fusion mode is defined as "Conc," the bistage decoding module is represented as "BID," and the DDN fusion is denoted as "DDN."

The results of the ablation investigations that are conducted on the FIFM module are displayed in Table V. It is clear that the "Diff" fusion mode produces the poorest results on all three datasets. The reason may be that the simple "Diff" fusion mode cannot obtain sufficient fusion information, similar to the FC-Siam-diff. The "DDN" fusion mode achieves the second-ranked scores on the SYSU-CD dataset but has worse results than the "Conc" fusion mode on the WHU-CD and LEVIR-CD datasets. Specifically, compared to "DDN," the improvements in F1 of "Conc" are 0.11% and 0.58% on the WHU-CD and LEVIR-CD, respectively. The reason might be that the "Conc" fusion mode can capture more feature information for CD. The proposed "FIFM" fusion mode obtains the best performance among these fusion modes. Specifically, compared to the "Conc" fusion mode, the proposed "FIFM" improves F1 by approximately 2.05% and 0.56% on the WHU-CD and LEVIR-CD datasets. On the SYSU-CD dataset, compared to the ranked-second "DDN" fusion mode, the proposed "FIFM" achieves improvements of 0.52% in F1. The reason is that the proposed FIFM module can simultaneously exploit the difference variation feature and the shared representation feature. Fig. 8 shows the visualization of the ablation experiment of the FIFM. It is evident that compared to other fusion modes, the results of the proposed "FIFM" fusion mode have fewer false negatives.

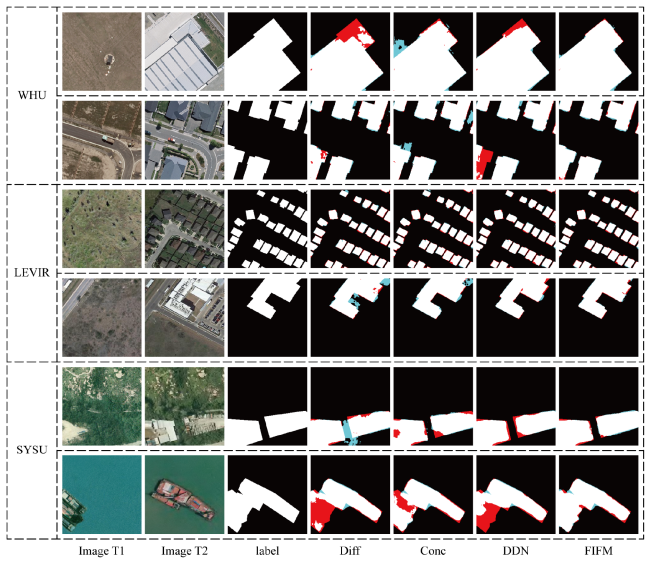

Fig. 8. Visual results of ablation experiment of the FIFM.

2) Ablation Experiment of Bistage Decoding: To test the effectiveness of the proposed bistage decoding module, we

TABLE VI
ABLATION EXPERIMENTAL OF "BISTAGE DECODING" ON THE THREE DATASETS

Methods	WHU-CD		LEVIR-CD		SYSU-CD		Params (M)
	F1(%)	IoU(%)	F1(%)	IoU(%)	F1(%)	IoU(%)	
Baseline + FIFM + FSD	91.89	85.00	90.52	82.65	79.74	66.32	3.87
Baseline + FIFM + BID	**93.27**	**87.39**	**91.11**	**83.67**	**80.10**	**66.81**	2.85

The best score is marked in bold.

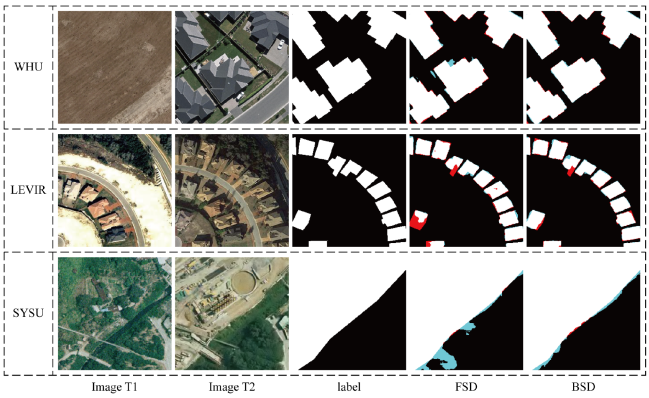

Fig. 9. Visual results of ablation experiment of bistage decoding strategy.

replaced it with the full-scale skip connections that are used in the traditional UNet3+ network. For convenience, "FSD" is the abbreviation used to denote full-scale skip connection decoding.

The quantitative results of the ablation experiment between the bistage decoding module and FSD are shown in Table VI. It is clear that the "BID" decoding mode not only achieves optimal results on all three datasets but also has fewer parameters. Specifically, when compared to the "FSD" decoding mode on the WHU-CD dataset, the improvements reached by the BID are approximately 1.38% and 2.39% of F1 and IoU, respectively. On the LEVIR-CD dataset, the "BID" decoding mode achieves improvements of approximately 0.59% of F1 and 1.02% of IoU. On the SYSU-CD dataset, the proposed "BID" decoding mode has enhanced F1 and IoU by approximately 0.36% and 0.49%, respectively. In addition, the proposed "BID" mode has fewer parameters compared with the "FSD" decoding mode. Fig. 9 shows the visualization results. It is clear that the "BID" mode generates more accurate edges than the "FSD" mode and that its detection results are closer to the label. Therefore, we can conclude that the proposed bistage decoding module is effective.

IV. CONCLUSION

In this article, a lightweight CD network with a novel FIFM and bistage decoding was proposed. In the feature encoder stage, the proposed FFBDNet used the EfficientNet-B4 to extract bitemporal image features more efficiently. Then, the bitemporal features were fused using the designed FIFM to obtain more discriminative fusion features. Finally, a novel bistage decoding module was proposed in the change map decoding process to alleviate the semantic gap between high- and low-level features. Based on the above contributions, a lightweight network based on FIFM and bistage decoding was proposed for CD. The proposed FFBDNet outperformed eight SOTA CD methods from the extensive experiments conducted on three challenging public datasets. The F1 values obtained for these three datasets were 93.27%, 91.11%, and 80.10% on the WHU-CD, LEVIR-CD, and SYSU-CD datasets, respectively. In addition, the network surpassed eight SOTA methods in terms of model parameters, and the ablation experiment established the effectiveness of the designed FIFM and bistage decoding module. It is worth noting that although the proposed FFBDNet achieves the best performance compared to other SOTA methods on all three datasets, all the methods are supervised-based with extensive labeled data. In the future, we will focus on weakly supervised and unsupervised CD algorithms to save labor-intensive and time-consuming annotated image labels.

ACKNOWLEDGMENT

The authors would like to thank everyone who has contributed datasets and fundamental research models to the public. The authors also appreciate the editors and anonymous reviewers for their valuable comments, which greatly improved the quality of this article.

REFERENCES

[1] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.
[2] Z. Lv et al., "Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective," *Proc. IEEE*, vol. 110, no. 12, pp. 1976–1991, Dec. 2022.
[3] Y. Ye, B. Zhu, T. Tang, C. Yang, Q. Xu, and G. Zhang, "A robust multimodal remote sensing image registration method and system using steerable filters with first-and second-order gradients," *ISPRS J. Photogrammetry Remote Sens.*, vol. 188, pp. 331–350, 2022.
[4] B. Zhu, C. Yang, J. Dai, J. Fan, Y. Qin, and Y. Ye, "R2Fd2: Fast and robust matching of multimodal remote sensing images via repeatable feature detector and rotation-invariant feature descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606115.
[5] Y. Ye and L. Shen, "HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 9–16, 2016.
[6] K. Jiang, W. Zhang, J. Liu, F. Liu, and L. Xiao, "Joint variation learning of fusion and difference features for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4709918.
[7] P. P. De Bem, O. A. de Carvalho Jr., R. F. Guimarães, and R. A. T. Gomes, "Change detection of deforestation in the Brazilian Amazon using Landsat data and convolutional neural networks," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 901.
[8] P. Du, S. Liu, P. Gamba, K. Tan, and J. Xia, "Fusion of difference images for change detection over urban areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1076–1086, Aug. 2012.
[9] P. Washaya, T. Balz, and B. Mohamadi, "Coherence change-detection with Sentinel-1 for natural and anthropogenic disaster monitoring in urban areas," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1026.

[10] D. He, Q. Shi, X. Liu, Y. Zhong, G. Xia, and L. Zhang, "Generating annual high resolution land cover products for 28 metropolises in China based on a deep super-resolution mapping network using Landsat imagery," *GISci. Remote Sens.*, vol. 59, no. 1, pp. 2036–2067, 2022.

[11] Q. Shi, M. Liu, A. Marinoni, and X. Liu, "UGS-1 M: Fine-grained urban green space mapping of 31 major cities in China based on the deep learning framework," *Earth Syst. Sci. Data*, vol. 15, no. 2, pp. 555–577, 2023.

[12] H. Jiang et al., "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 7, 2022, Art. no. 1552.

[13] 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, 2021, Art. no. 102597.

[14] Q. Ding, Z. Shao, X. Huang, and O. Altan, "DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, 2021, Art. no. 102591.

[15] A. Mondini, F. Guzzetti, P. Reichenbach, M. Rossi, M. Cardinali, and F. Ardizzone, "Semi-automatic recognition and mapping of rainfall induced shallow landslides using optical satellite images," *Remote Sens. Environ.*, vol. 115, no. 7, pp. 1743–1757, 2011.

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

[17] H. Fang, P. Du, X. Wang, C. Lin, and P. Tang, "Unsupervised change detection based on weighted change vector analysis and improved Markov random field for high spatial resolution imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6002005.

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

[19] K. Tan, X. Jin, A. Plaza, X. Wang, L. Xiao, and P. Du, "Automatic change detection in high-resolution remote sensing images by using a multiple classifier system and spectral–spatial features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3439–3451, Aug. 2016.

[20] Z. Lv, F. Wang, T. Liu, X. Kong, and J. A. Benediktsson, "Novel automatic approach for land cover change detection by using VHR remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8016805.

[21] M. Wang et al., "Exploiting neighbourhood structural features for change detection," *Remote Sens. Lett.*, vol. 14, no. 4, pp. 346–356, 2023.

[22] J. Lei, Y. Gu, W. Xie, Y. Li, and Q. Du, "Boundary extraction constrained siamese network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621613.

[23] Q. Shu, J. Pan, Z. Zhang, and M. Wang, "DPCC-Net: Dual-perspective change contextual network for change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102940.

[24] Y. Ye et al., "An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 516.

[25] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, "A multiscale framework with unsupervised learning for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622215.

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

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

[28] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

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

[30] Y. Lei, X. Liu, J. Shi, C. Lei, and J. Wang, "Multiscale superpixel segmentation with deep features for change detection," *IEEE Access*, vol. 7, pp. 36600–36616, 2019.

[31] Y. Wang et al., "Mask DeepLab: End-to-end image segmentation for change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 104, 2021, Art. no. 102582.

[32] M. Jiang, X. Zhang, Y. Sun, W. Feng, Q. Gan, and Y. Ruan, "AFS-Net: Attention-guided full-scale feature aggregation network for high-resolution remote sensing image change detection," *GISci. Remote Sens.*, vol. 59, no. 1, pp. 1882–1900, 2022.

[33] Z. Chen et al., "Egde-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement," *ISPRS J. Photogrammetry Remote Sens.*, vol. 191, pp. 203–222, 2022.

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

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

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

[37] G. Pei and L. Zhang, "Feature hierarchical differentiation for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6514105.

[38] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

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

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

[41] 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, 2020.

[42] B. Yang, Y. Huang, X. Su, and H. Guo, "MAEANet: Multiscale attention and edge-aware siamese network for building change detection in high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 19, 2022, Art. no. 4895.

[43] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.

[44] S. Chu, P. Li, M. Xia, H. Lin, M. Qian, and Y. Zhang, "DBFGAN: Dual branch feature guided aggregation network for remote sensing image," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, 2023, Art. no. 103141.

[45] Z. Fu, J. Li, Z. Chen, L. Ren, and Z. Hua, "DAFT: Differential feature extraction network based on adaptive frequency transformer for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5061–5076, 2023.

[46] T. Lei, D. Xue, H. Ning, S. Yang, Z. Lv, and A. K. Nandi, "Local and global feature learning with kernel scale-adaptive attention network for VHR remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7308–7322, 2022.

[47] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 539–546.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[49] 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.

[50] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[51] W. Sun et al., "MLR-DBPFN: A multi-scale low rank deep back projection fusion network for anti-noise hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522914.

[52] 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.

[53] X. Xiang, D. Tian, N. Lv, and Q. Yan, "FCDNet: A change detection network based on full-scale skip connections and coordinate attention," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6511605.

[54] 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.

[55] Z. Zheng, Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang, "CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 247–267, 2021.
[56] 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.
[57] 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, 2021, Art. no. 5604816.
[58] H. Yang et al., "A lightweight siamese neural network for building change detection using remote sensing images," *Remote Sens.*, vol. 15, no. 4, 2023, Art. no. 928.

Mengmeng Wang received the B.S. degree in surveying and mapping from Henan Polytechnic University, Jiaozuo, China, in 2017, and the M.S. degree in surveying and mapping from Southwest Jiaotong University, Chengdu, China, in 2020, where he is currently working toward the Ph.D. degree in surveying and mapping science and technology.

His research interests include image processing, deep learning, and change detection.

Bai Zhu received the B.S. degree in remote sensing science and technology from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, in 2019, where he is currently working toward the Ph.D. degree in surveying and mapping science and technology.

His research interests include remote sensing image processing, multimodal image matching, image registration, and feature extraction.

Jiacheng Zhang received the B.S. degree in remote sensing science and technology from the Southwest Jiaotong University, Chengdu, China, in 2021, where he is currently working toward the M.S. degree in remote sensing science and technology with the Faculty of Geosciences and Environmental Engineering.

His research interests include multimodal remote sensing image matching and fusion with related high-level vision tasks.

Jianwei Fan received the B.S. degree in electronic information science and technology from the Henan University of Science and Technology, Luoyang, China, in 2011, and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2017.

He is currently a Lecturer with the School of Computer and Information Technology, Xinyang Normal University, Xinyang, China. His main research interests include remote sensing image processing, image registration, and feature extraction.

Yuanxin Ye (Member, IEEE) received the B.S. degree in remote sensing science and technology from Southwest Jiaotong University, Chengdu, China, in 2008, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013.

He is currently a Professor with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University. His research interests include remote sensing image processing, image registration, change detection, and object detection.

Dr. Ye received the ISPRS Prizes for Best Papers by Young Authors at 23rd International Society for Photogrammetry and Remote Sensing Congress, Prague, Czech Republic, in 2016 and the Best Youth Oral Paper Award at ISPRS Geospatial Week 2017, Wuhan, in 2017.