Two-Level Feature Fusion Network for Remote Sensing Image Change Detection

Mingyao Feng, Ruifan Zhang, Hao Wang, Yikun Liu, and Gongping Yang

***Abstract*—With the advancement of satellite technology, the application space of change detection (CD) in remote sensing images is continuously expanding. However, the development of satellite remote sensing technology is still ongoing, and limited resolution and complex ground object information remain significant challenges in the field of CD. Recent CD networks generally utilize multifeature fusion to make full use of detailed information at different scales. However, most networks have limited capabilities in handling large-scale feature maps, leading to an impact on the effectiveness in detecting detailed information. In this article, we propose a two-level feature fusion CD network that enhances the semantic information contained in large-scale difference feature through a combination of convolutional neural network and transformer-based feature fusion structures. Leveraging a simple backbone network (ResNet-18) to extract dual-temporal feature maps, our model achieves better performance to mainstream state-of-the-art networks. On the LEVIR-CD, WHU-CD, and SYSU-CD datasets, we obtain F1 scores of 92.03%/92.73%/83.25%, intersection over union of 85.24%/86.45%/71.31%, and Kappa coefficient (κ) of 91.61%/92.45%/78.26%, respectively.**

***Index Terms*—Change detection (CD), dilated convolution, feature fusion, remote sensing image, semantic information.**

I. Introduction

CHANGE detection (CD) based on remote sensing technology is the process of utilizing two or more images captured at different times of the same geographical location to identify and recognize differences in land features [1]. The development of remote sensing technology has captured the interest of many researchers and has been applied in various fields, including disaster monitoring [2], [3], resource survey [4], [5], [6], urban planning [7], [8], and more.

For CD tasks, remote sensing images often exhibit complex backgrounds with features, such as mountains, rivers, forests, and roads, making the learning of background information challenging. Simultaneously, due to differences in capture times, factors such as lighting and climate can lead to differences in the distribution of images. Considering the focus on building CD tasks, where the target is buildings that vary in morphology, factors such as changes in shadow patterns due to building height and differences in roof colors may also have a negative impact on the accuracy of CD result.

In the past few decades, considerable efforts have been invested in developing various CD methods, including traditional approaches [9], [10] and those based on deep learning [11]. Nowadays, owing to the powerful feature extraction capabilities of convolutional neural networks (CNNs), CD methods based on CNNs have become the predominant method in the field of remote sensing image CD and have demonstrated excellent performance. Simultaneously, with the success of transformers [12], originally from the field of natural language processing (NLP), the vision transformer (ViT) [13] introduced its structure into the field of computer vision (CV), its powerful contextual modeling ability has effected interest in various CV field. Networks structured with the transformer architecture have become prevalent in the CV domain [14], [15], [16], [17]. Transformer has also been introduced to the field of remote sensing image CD and has achieved promising results [18], [19], [20], [21], [22].

Due to the fact that satellite technology is still in the development stage and the high costs of the manual annotation, the public datasets mainly used in the field of remote sensing image CD often lead to a negative insufficient in dataset accuracy and image quantity. These issues result in a more significant impact of factors, such as lighting and shadows in the image affecting the model's judgment of the distribution of pixel features to a greater extent, ultimately affecting the score effect of the model. Current CD tasks predominantly involve the use of siamese networks structured with the U-Net architecture [23]. These networks employ a shared parameter backbone network and designed modules to model the feature maps for obtaining semantic feature maps of dual-temporal images as an encoder structure. The decoder structure process two semantic feature maps to derive the temporal change information, and fuse features in different scales during the upsampling phase. Existing methods primarily focus on the feature extraction stage of dual-temporal remote sensing images, aiming to design siamese networks with the shared weights and strong representational capabilities to extract high-level semantic features of dual-temporal remote sensing images [24], [25], [26], [27]. For the decoder stage, most models use some traditional methods such as concatenation or making a difference [28] to calculate differential feature maps, and the final prediction map is obtained through the upsample operation to the differential feature maps [29], [30], [31], [32]. In this process, the large-scale feature map obtained from the

Manuscript received 25 January 2024; revised 22 March 2024; accepted 7 April 2024. Date of publication 10 April 2024; date of current version 22 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62176139, and in part by the Major Basic Research Project of Natural Science Foundation of Shandong Province under Grant ZR2021ZD15. *(Corresponding author: Gongping Yang.)*

The authors are with the School of Software, Shandong University, Jinan 250101, China (e-mail: 1304912860@qq.com; 736381914@qq.com; 1742496405@qq.com; liuyk@163.com; gpyang@sdu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3387093

middle layer of the network contains fewer semantic features and can only use to supplement some detailed information of small receptive fields.

To address the aforementioned issues and enhance the semantic information representation capability of large-scale feature maps, we propose a two-level feature fusion CD network architecture. This architecture aims to strengthen the model's CD capability by improving the accuracy of the temporal change information at different scales while fully integrating semantic information at various scales. Considering that the main function of CNN is to learn feature representations of input images through neural networks, dual temporal remote sensing images come from different distribution spaces, and the feature information of dual temporal feature maps obtained using siamese networks belongs to different distributions. By subtracting absolute values, the generated feature maps can belong to the same distribution. To model the large-scale feature maps in the model and enrich their semantic information representation, our model use a primary-level feature fusion module (PFFM) to fuse semantic information from feature maps of different scales. After obtaining the temporal change information through subtractions, mixed convolutions (Mixed-conv) is applied to the large-scale feature maps to significantly increase the receptive field with lower computational cost and enhance the semantic information representation. Finally, an advanced feature fusion transformer (AFF Transformer) is employed to further fuse large-scale feature maps with the high-level semantic information of small-scale feature maps, augmenting the semantic information of large-scale feature maps and enhancing the segmentation effectiveness of change regions.

The contributions of our work are as follows.

1) We propose a two-level feature fusion model for CD in remote sensing images. The network employs different feature fusion approaches during the stages of obtaining semantic feature maps of dual-temporal images and processing the temporal change information of feature maps. This is done to enable low-level, high-resolution feature maps to capture more advanced semantic information, thereby enhance the detection effect of the model on change regions.
2) To enhance the receptive field of large-scale feature maps while preserving local dependencies between pixels, we design a Mixed-conv composed of a combination of dilated convolution and regular convolution to enhance the temporal change information. This convolutional approach allows large-scale feature maps to maintain a higher receptive field range and capture detailed semantic information, thereby increasing the semantic information representation.
3) We propose a feature fusion cross-transformer structure to handle multiscale feature maps. This structure facilitates the global modeling of difference features through global attention, while simultaneously allowing the integration of advanced-level semantic information from small-scale feature maps into large-scale feature maps.

The rest of this article is as follows. Section II provides a brief overview of recent related work. Section III details the proposed methodology. Section IV presents a series of experimental results and analyzes the model. Finally, Section V concludes this article.

II. Related Work

A. Remote Sensing Image CD Based on Deep Learning

Different from earlier methods primarily utilizing traditional approaches for CD tasks [33], [34], [35], [36], [37], [38], [39], contemporary remote sensing image CD predominantly employs deep learning methods [40]. These models often use U-Net as the main architecture [29], [30], [31]. In these models, the backbone network first extracts feature maps from dual-temporal remote sensing images. The dual-temporal feature maps are then fused to obtain the temporal change information, and the final prediction results are obtained through an upsampling structure. As CD tasks involve processing two input images captured at different times, the fusion of dual-temporal feature maps can be categorized into two ways based on Daudt et al.'s [28] work: image-level and feature-level. Image-level methods, represented by the FC-EF model [28], directly connect dual-temporal images as a whole input to the semantic segmentation network by concatenation or similar operation to obtain the temporal change information. The network in Peng et al.'s [31] work adopted the image-level method by feeding dual-temporal images into the U-Net++ network. Feature-level methods, more popular than image-level methods, simultaneously feed dual-temporal images into feature extraction networks with the same structure and shared weights. Feature-level models, such as FC-Conc and FC-Diff [28], used concatenation and elementwise subtraction of dual-temporal feature maps, respectively, to obtain the temporal change information. The change maps are then reconstructed through upsampling and convolution modules. Feature-level methods are more favored as they are computationally efficient and avoid the cost of learning the differences in dual-temporal feature maps. Most of the feature-level methods used subtraction [24], [25], [27], [30] while a few used concatenation [41]. In addition, other features and methods are being explored. SUACDNet [42] used the sum, difference, and the feature maps themselves of dual-temporal feature maps. The authors in [43] and [44] proposed algebraic methods to derive difference maps.

Contemporary models are mostly focused on enhancing the feature representation capability of the obtained dual-temporal feature maps primarily through attention mechanisms. STA [45] used a spatial–temporal attention module to capture rich global spatial–temporal relationships among pixels in the entire spatial–temporal space, aiming to acquire more discriminative features. BIT [18] utilized a transformer encoder–decoder structure to enhance the interested change content in dual-temporal feature maps and exclude irrelevant changes. In addition to enhancing the semantic information of dual-temporal feature maps, some networks also enhance the temporal change information through various approaches. In order to improve the boundary integrity and internal compactness of objects in the output change map, IFN [46] integrated multilevel deep features of the original image with difference features through channel attention and spatial attention modules for change map reconstruction. MeGAN [26] utilized a metric-learning-based GAN framework to obtain change prediction information.

B. Multifeature Fusion

The CNN network extracts semantic features from the input through layerwise convolution, with each layer having different receptive fields. Higher layers, due to their larger receptive fields, yield small-scale high-level feature maps with powerful semantic information representation. However, these feature maps have lower resolution, losing many spatial details during the downsampling process, making them suitable for detecting large target objects. On the other hand, lower layers produce large-scale low-level feature maps with higher resolution, containing more spatial details. However, they have smaller receptive fields and poorer semantic information representation capabilities, making them suitable for detecting small target objects [47]. The approach of multiscale feature fusion involves combining features from both high and low layers, absorbing the advantages of different-scale feature maps, enabling the feature maps to possess both semantic and spatial information.

Multiscale feature fusion networks, known for their effectiveness, have found widespread application in the field of remote sensing. Virtually all existing CD networks have embraced the concept of multiscale feature fusion in their network design. Numerous works have focused on enhancing neural network feature extraction by incorporating multilayer feature fusion structures [45], [46], [48], [49], [50]. There are two prevalent methods for multiscale feature fusion. Some kinds of method using the parallel multibranch network structure, exemplified by the structure used in HRNet [51], as seen in the pyramid spatial–temporal attention model used in Chen and Shi's [45] work. This kind of approach combines different scales of spatial–temporal attention contexts to generate multiscale attention features. Other kinds of method using the serial multibranch structure, represented by the upsampling process in the U-Net structure. In this kind of approach, low-level feature maps are fused with high-level feature maps that have undergone upsampling through skip connections, which is the primary method used in the upsampling process. For instance, the neighbor aggregation module used in Li et al.'s [52] work adopts a structure similar to HRNet, allowing each dimension of the feature map to fuse semantic features from adjacent dimensions. Many methods, including [27], [46], [52], and [53], used skip connections to fuse features during the upsampling process. In addition, during the upsampling stage, SEIFNet [54] used an adaptive context fusion module to guide the recovery of low-level features by utilizing contextual information from high-level features. It changes the attention weights of high-level and low-level features through attention operations in the max pooling and average pooling layers, achieving the fusion of high-level and low-level features. It is noteworthy that most of these multifeature fusion modules were primarily used in enhancing the feature representation of deep-layer feature maps or improving the upsampling effect of deep-layer feature maps, with limited emphasis on improving information from large-scale feature maps.

C. Dilated Convolution

Dilated (atrous) convolution, initially proposed for addressing image segmentation challenges [55], was designed to extract multiscale information without altering the resolution of the feature map by controlling the receptive field. In the early segmentation approaches, networks typically underwent a series of downsampling processes to compress images, reducing computational complexity, followed by a series of upsampling processes to restore the original image size. However, this additional upsampling step inevitably led to the loss of a considerable amount of detailed information. The introduction of dilated convolution aims to circumvent the downsampling–upsampling network structure. By introducing gaps between the elements of the convolutional kernel, dilated convolution achieves an enlargement of the receptive field without sacrificing detailed information, thus providing an effective means of multiscale information extraction with lower computational overhead.

However, dilated convolution, due to its computation resembling a checkerboard pattern, generates convolution results in a layer that entirely come from independent sets of the previous layer, lacking mutual dependence and leading to the loss of local information. In addition, the sparse sampling of input signals by dilated convolution results in information obtained from distant convolutions lacking correlation, affecting the learning effectiveness. This issue is referred to as the gridding effect [56]. Since dilated convolution does not utilize all pixel values within its range, and there are unused elements between nonzero elements, it inevitably loses the local dependency characteristics inherent in regular convolution modules. In remote sensing imagery, where pixels encompass factors, such as lighting and shadows, avoiding interference from these elements makes the information from neighboring pixels particularly crucial. Consequently, only a few CD models in recent years have incorporated dilated convolution. Models such as those in [25] and [41] employed four dilated convolutions of different sizes in a parallel structure, forming a Mixed-conv inserted into the backbone convolutional network to extract hierarchical features, increase the depth and width of the network. In Li et al.'s [52] work, a serial structure connected dilated convolutions with different dilation rates, gradually exploring temporal changes in the receptive field from large to small.

D. Cross Attention Transformer Model

The transformer, a deep learning model proposed by Vaswani et al. [12], has gained immense popularity, primarily dominating the field of NLP. Drawing inspiration from the success in NLP, ViT introduced by Dosovitskiy et al. [13] directly apply the standard transformer to the CV domain. While transformers lack certain inductive biases from convolutions, the global receptive field inherent in transformer networks provides powerful contextual modeling capabilities. With sufficient training, transformers have demonstrated outstanding results that surpass those achieved by CNNs. As a result, transformer-based models have rapidly gained prominence in various CV tasks, including image classification, semantic segmentation, object detection, image generation, superresolution, and more. Their effectiveness has positioned transformers as a formidable architecture in the CV domain, expanding their influence beyond their initial success in NLP.

For wide range remote sensing images representing extensive ground areas, effectively addressing the long short-term dependencies can significantly enhance the model's performance. Many models tackle this issue by employing a global attention mechanism. The usage of the global self-attention mechanism in transformer models have quickly become a well-performing framework in the remote sensing image domain. Numerous instances of using transformers in CD tasks have emerged. For instance, the work by BIT [18] pioneered the design of a hybrid model, which concatenated CNN with transformer components, exploring the potential of transformers in CD tasks. The concise structure of the transformer-mixed model demonstrated the effectiveness of transformers. Global semantic relationship modeling in both temporal and spatial dimensions is advantageous for representing semantic changes. Another example is SLDDNet [22], which adopted a CNN-transformer architecture, using a transformer semantic selector to capture global semantic relationships and strengthens local feature information through a pyramid structure of feature stacking. In addition, the ChangeFormer model proposed by Bandara and Patel [19] introduced a transformer-based siamese network structure. It combined hierarchical transformer encoders with MLP decoders to extract coarse and fine features of dual-temporal image for CD.

The application of transformer models with cross-attention mechanisms first appeared in Vaswani et al.'s [12] work as part of the decoder structure, primarily used in NLP for tasks, such as machine translation, text recognition, and image captioning. These models can fuse information from multiple sources or handle cross-modal data, enabling them to better capture dependencies between different types of information. In the CV domain, there are also several applications of transformer models with cross-attention mechanisms. For instance, CrossViT [57] investigated the effectiveness of incorporating the idea of feature fusion at different scales into transformers. In the field of CD in remote sensing images, BIT was among the first to use this cross-attention mechanism to reintegrate advanced-level semantic information extracted from the feature maps back into the feature maps. BIT employ the approach reduces the computational cost of the transformer module by extracting limited-length high-level semantic information. Another work, DCAT [58] modeled the changes in relationships between patches by connecting a series of hierarchical double-cross-attention blocks to extract multiscale features.

III. Model Structure

A. Overview

Our two-level feature fusion model's process is illustrated in the Fig. 1. Distinguishing itself from existing networks, our model achieves more accurate semantic feature extraction in the process of obtaining dual-temporal feature maps. Our model employs the simplest pretrained residual network structure, ResNet-18 [59], as the backbone network, using the first four layers' outputs as different-scale feature maps. These maps are then preliminarily fused through the primary-level feature fusion model (PFFM). The PFFM's process is illustrated in the Fig. 2. Subsequently, similar to most networks, the dual-temporal feature maps of the same dimension are subtracted and the absolute value is taken, resulting in the temporal change information representing different scales.

Due to the simple structure of the backbone and the PFFM, the network extracts less temporal information for dual-temporal images. To enhance the change information in large-scale feature maps, we design a Mixed-conv that runs in parallel with dilated convolution and ordinary convolution, significantly increasing the receptive field while preserving local dependencies. Considering that small-scale feature maps in the network can best express the distribution information of temporal information, they represent a valuable form of advanced semantic features. Consequently, we fuse change maps of different scales with small-scale feature maps using an AFF Transformer. This allows the semantic information of large-scale feature maps and the advanced semantic information of small-scale feature maps to form a mapping relationship, enabling large-scale feature maps to incorporate more high-dimensional semantic information at the global level. This achieves the fusion of advanced multiscale feature maps. Finally, we adopt a step by step upsampling approach. After all, different-scale feature maps are fed into the upsampling module, which generates pixel-level prediction maps as the overall output of the network through simple concatenation and convolution.

B. Primary-Level Feature Fusion Module

Considering that ResNet-18 extracts limited semantic information, we preliminarily fuse feature maps of different scales through PFFM. For remote sensing images, the semantic information contained in feature maps of different scales is crucial for the subsequent CD process. Larger scale feature maps contain more detailed information, while smaller scale feature maps possess higher dimensional semantic information. Inspired by HRNet, we use a parallel network structure to fuse feature maps of different scales, exchanging information between feature maps of different scales to learn sufficiently rich features of various scales. Subsequently, we convert feature maps of each scale to the same number of channels, reducing redundant information in smaller scale feature maps to decrease computational costs during subsequent feature fusion, and enhancing the semantic information that larger scale feature maps can store.

1) For the feature map F, we pass it through a 3×3 convolutional block to further extract semantic features while adjusting the channel number to C_0. This ensures that the resolution of the new feature map F' is the same as F, and that it undergoes the same number of convolutions as feature maps of other scales.
2) For the large-scale feature map F_+ with higher resolution, we pass it through a 3×3 convolutional block to extract semantic features while adjusting the channel number to C_0. Subsequently, we perform downsampling using MaxPooling to obtain a new feature map F'_+ with the same resolution as F.

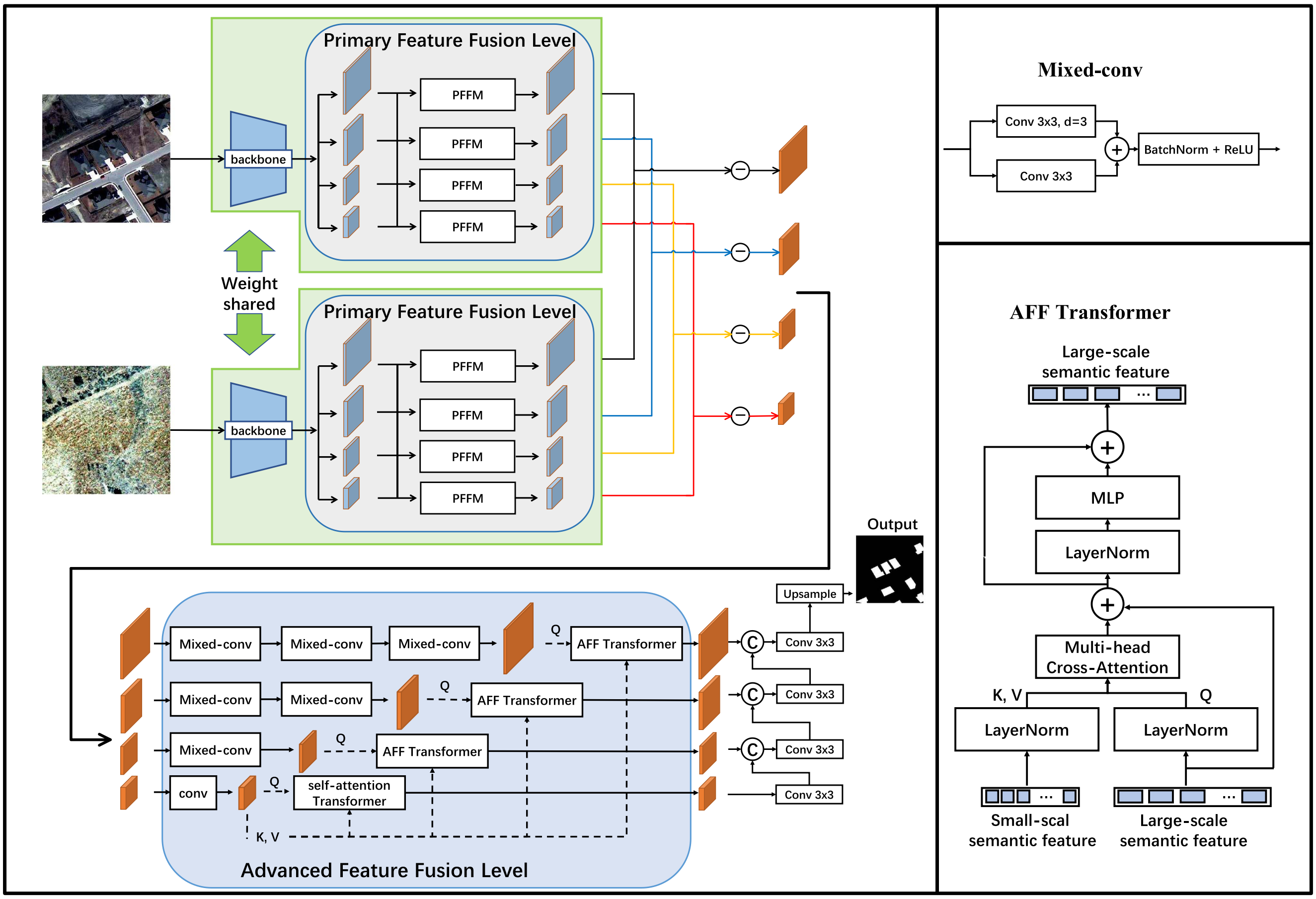

Fig. 1. Visualization of the overall workflow of our two-level feature fusion model. A shared-weight ResNet-18 is utilized to extract four different-scale feature maps for a pair of dual-temporal feature maps. PFFM is employed to merge semantic information across various scales. Mixed-conv is applied to significantly increase the receptive field while retaining local dependencies among neighbor pixels. The AFF Transformer as the advanced level feature fusion model use to integrate deep semantic information from the smallest scale feature map.

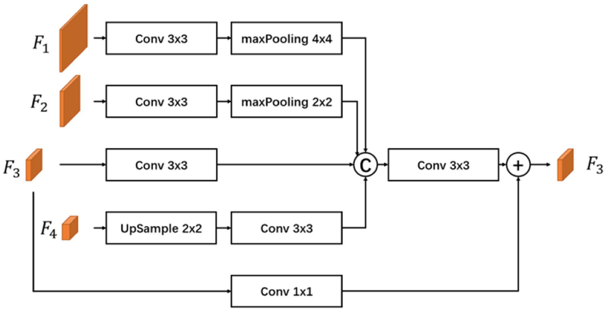

Fig. 2. Primary-level feature fusion module.

3) For the small-scale feature map F_- with lower resolution, we use an interpolation upsampling module to obtain the same resolution as F. Subsequently, we pass it through a 3 × 3 convolutional block to reduce the loss of information accuracy caused by the interpolation during the upsampling process, while adjusting the channel number to C_0. This yields a new feature map F'_- with the same resolution as F.

4) Finally, we concatenate the new feature maps with the same resolution, adjust the channel number to C through a 3 × 3 convolution, obtaining the fused feature F_{new} at that scale. Using residual learning, we add F_{new} to the original-scale feature map F, which has had its channel number normalized through a 1 × 1 convolution. This results in the module's final output F_{out} at that scale.

Where C_0 represents the same number of channels to which various scale feature maps are adjusted in the intermediate process, and C represents the same number of channels for different scale feature maps in the final output. Taking the feature map F_3 as an example, the process of PFFM is as follows:

$$\mathbf{F}_1' = \text{maxPooling}(\text{conv}_{3\times3}(\mathbf{F}_1)) \tag{1}$$

$$\mathbf{F}_2' = \text{maxPooling}(\text{conv}_{3\times3}(\mathbf{F}_2)) \tag{2}$$

$$\mathbf{F}_3' = \text{conv}_{3\times3}(\mathbf{F}_3) \tag{3}$$

$$\mathbf{F}_4' = \text{conv}_{3\times3}(\text{UpSample}(\mathbf{F}_4)) \tag{4}$$

$$\mathbf{F}_3^{\text{out}} = \text{conv}_{1\times1}(\mathbf{F}_3) + \text{conv}_{3\times3}(\text{Cat}(\mathbf{F}_1', \mathbf{F}_2', \mathbf{F}_3', \mathbf{F}_4')) \tag{5}$$

where $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$, and $\mathbf{F}_4$ are different scale feature maps, $\mathbf{F}_3^{\text{out}}$ is the output of the feature map $\mathbf{F}_3$, $\text{maxPooling}(\cdot)$ represents downsampling pooling operation by maxpooling operation, $\text{UpSample}(\cdot)$ represents the upsampling process using bilinear interpolation, and $\text{conv}_{3\times3}(\cdot)$ in this place represents a 3×3 convolutional layer with batch normalization and ReLU activation, and $\mathrm{Cat}(\cdot)$ denotes the concatenation operation.

In the above-mentioned process, we adjust the channel number to the same value C_0 for each scale feature map, ensuring that different-scale feature maps have the same semantic weight during concatenation. It is worth noting that, to reduce the computational cost and improve the learning efficiency of the subsequent model, the channel number C after multifeature fusion for each scale feature map is the same. This ensures that pixels in the large-scale feature map contain more semantic information and that the semantic information in the small-scale feature map is more compact.

C. Mixed-conv Blocks

Although multiscale feature fusion can effectively combine high-level semantic information from small-scale feature maps with detailed information from large-scale feature maps, the large-scale feature map passes through fewer layers, containing lower level details. The semantic information obtained from the small-scale feature map will introduce errors during the upsampling process. In addition, the subtraction and absolute difference operations for derive the temporal change information from the dual-temporal feature map result in a loss of semantic information, making it challenging for large-scale feature maps to contain sufficient semantic information. To supplement the semantic information of large-scale feature maps and efficiently increase the receptive field, increasing the receptive field is a simple and effective method. However, as the scale of the feature map increases, the loss of semantic information becomes more severe, requiring a substantial increase in the receptive field, leading to an exponential growth in computational demands. To address this, we explore the application of dilated convolutions in the field of remote sensing imagery. Our goal is to increase the receptive field of the feature map to a certain extent. Since our objective is relatively straightforward, the common improvements for dilated convolutions are too complex for our needs. Therefore, we opt for a simple approach named Mixed-conv by incorporating a standard convolution in parallel with the dilated convolution. Different from the Mixed-conv used in the related works using four different dilation rate, we only use one 3×3 dilated convolution with a dilation rate of 3 and one standard 3×3 convolution. The Mixed-conv reduces the impact of the gridding effect, simultaneously increasing the receptive field of the large-scale feature map without sacrificing local dependencies between pixels.

The formula for each Mixed-conv is as follows:

$$\mathrm{MixedConv}(\mathbf{F}) = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\mathbf{F}) + \mathrm{Conv}_{3\times3}^{d=3}(\mathbf{F}))) \tag{6}$$

where MixedConv denotes the Mixed-conv, $\mathrm{ReLU}(\cdot)$ represents the ReLU activation function, BN denotes the batch normalization layer, $\mathrm{Conv}_{3\times3}(\cdot)$ in this place is a standard 3×3 convolution, $\mathrm{Conv}_{3\times3}^{d=3}(\cdot)$ is a 3×3 dilated convolution with a dilation rate of 3.

For larger scale feature maps requiring a greater increase in receptive field, our network allows the larger scale feature maps to pass through more Mixed-conv blocks. In subsequent experiments, we demonstrate the effectiveness of this simple parallel Mixed-conv approach in enhancing the semantic information for large-scale feature maps. In addition, for small-scale feature maps that do not use Mixed-conv modules, due to their small resolution, dilated convolutions may skip a large amount of relevant semantic information. As an alternative, we let them pass through a 3×3 convolution block, allowing the pixels in the feature map to focus only on the semantic information of their neighboring pixels, thereby reducing the impact of semantic information loss during the subtraction process of feature maps to some extent.

D. Advanced-Level Feature Fusion Transformer

A small-scale feature map is the temporal change information obtained through a complete CNN backbone and a primary-level FFM. It is a high-level semantic feature that contains rich high-level semantic information. This semantic feature has a large receptive field, deep feature information, and also contains some detail information of large-scale feature maps, which can well represent the semantic information of differential features. After enhancing the semantic information of large-scale feature maps, we use an AFF Transformer structure to further fuse the advanced semantic features in small-scale feature maps. In order to achieve the fusion of features at different scales, unlike the usual transformer structure, our AFF transformer does not focus on global long and short-term memory. Instead, it utilizes the transformer structure to adaptively learn the corresponding relationships between different positions. The input K vectors, Q vectors, and V vectors required by the transformer are all modified to the transpose of the original content, focusing on the correlation information of features at different scales, and the feature information of each region is weighted based on the importance of each feature.

To calculate the similarity information of features between feature maps of different scales, these feature maps need to have the same number of semantic features, that is, the same number of regions. We crop different scale feature maps of various scales into 16×16 regions, with each region transformed into a semantic word vector of length $(C \times P_h \times P_w)$, from $\mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{(C\cdot P_h\cdot P_w)\times16\times16}$, where C is the number of channels in the original feature map, and $P_h = \frac{H}{16}$, $P_w = \frac{W}{16}$ are the height and width of each region, ensuring that each feature map obtains the same number of semantic features. Due to cropping feature maps of different scales into the same number of blocks to obtain semantic features, blocks located at the same position on the change feature map will represent the same region. In the process of cross attention, we treat large-scale feature maps as query vectors and small-scale feature maps as key vectors and value vectors. By performing dot product operations between query vectors and key vectors, we calculate the attention weights between large-scale and small-scale features, and use the softmax function to obtain the similarity correlation attention map between large-scale feature maps and small-scale feature maps. Then, the obtained attention map use to weight and sum the small-scale feature map, calculate the region information that the large-scale feature map focuses on, and obtain the final AFF Transformer output. The AFF Transformer cross attention and each layer of multihead cross attention are defined as follows:

$$CA(\mathbf{Q},\mathbf{K},\mathbf{V}) = \sigma\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{7}$$

$$\text{Out}(\mathbf{F}^L,\mathbf{F}^S) = (\text{Cat}(\text{head}_1,\ldots,\text{head}_h)\mathbf{W}^O)^T \tag{8}$$

$$\text{where}\quad \text{head}_j = CA((\mathbf{F}^L)^T\mathbf{W}_j^q,(\mathbf{F}^S)^T\mathbf{W}_j^k,(\mathbf{F}^S)^T\mathbf{W}_j^v) \tag{9}$$

where $\mathbf{W}_j^q,\mathbf{W}_j^k,\mathbf{W}_j^v \in \mathbb{R}^{P_hP_w\times d}$, $\mathbf{W}^O \in \mathbb{R}^{hd\times P_hP_w}$ are parameter matrices obtained from linear mappings, $\sigma(\cdot)$ denotes the softmax function operated on the channel dimension of small-scale feature map, h is the number of attention heads, d is the spatial dimension of three linear projection layers, $\mathbf{F}^L$ represents the transformed large-scale feature map, $\mathbf{F}^S$ represents the transformed small-scale feature map. The output of the AFF Transformer module will be transformed back to the same scale of the large-scale feature map.

Similarly, the AFF Transformer structure can also act as self-attention on minimum scale feature maps, enhancing features with high correlation and reducing features with low correlation, learning the correlation information between features through self-attention, and strengthening feature information that is more suitable for the minimum scale. Considering that AFF Transformer may cause the feature map to lose some low correlation feature information, the minimum scale feature map first performs AFF Transformer with other scale feature maps, while the minimum scale change feature map that passes through the transformer structure only participates in the subsequent upsampling stage. It is worth noting that due to the limited number of channels in feature maps of different scales, transformers using this cross attention structure reduce a significant amount of computational requirements compared to self-attention transformers.

E. Loss Function

For loss optimization, we utilize a hybrid loss function combining binary cross-entropy loss and dice loss, proposed by A2-Net [52]. In addition to the overall output of the network, after each concatenation operation in the upsampling stage at the network's end, a 1 × 1 convolution module is applied to generate a pixel-level prediction map at that scale. These prediction map are then upsampled and compared with the ground truth to calculate the loss, which are added to the total loss. This process is employed to further optimize the parameters of the model at that scale and smaller scales. The specific formula for the loss function is as follows:

$$L_{\text{bce}_i}(p_i,gt) = p_i\cdot\log gt + (1-p_i)\log(1-gt) \tag{10}$$

$$L_{\text{dice}_i}(p_i,gt) = 1-\frac{2\cdot p_i\cdot gt}{||p_i||+||gt||} \tag{11}$$

$$L = \sum_{i=1}^{4}(L_{\text{bce}_i}(p_i,gt)+L_{\text{dice}_i}(p_i,gt)) \tag{12}$$

where p_i represents the predicted map after upsampling for the ith dimension, gt is the ground truth, and $||\cdot||$ denotes L1 regularization.

IV. Experimental Results and Analysis

A. Experimental Setup

1) Experimental Details: Our experiments are implemented using PyTorch [60], and all experiments are conducted on a system equipped with an Intel Xeon Gold 5218 CPU (2.30 GHz) and a GeForce RTX 3090 GPU (24 GB memory). The experiments are carried out on three commonly used datasets. During training, we apply standard data augmentation techniques, such as flipping, cropping, scaling, and Gaussian blur. The model is optimized using the adaptive moment estimation [61] optimizer, with a batch size of 8, an initial learning rate of 1.25e-4, momentum of (0.9, 0.99), and weight decay of 1e-4. The learning rate decay followed a polynomial decay strategy, as given by the formula:

$$\text{lr}_{\text{new}} = lr\cdot\left(1-\frac{\text{cur\_epoch}}{\text{max\_epoch}+1}\right)^{0.9} \tag{13}$$

where lr represents the initial learning rate, lr_{new} is the updated learning rate calculated using the formula, cur_epoch denotes the current epoch number, and max_epoch is the total number of epochs for training.

2) Evaluation Indicators: We adopt five commonly used evaluation metrics, with F1-score as the primary metric, calculated using precision and recall. In addition, we employed precision (Pre), recall (Rec), intersection over union (IoU), and Kappa coefficient (κ) as evaluation metrics, with the following formulas:

$$\text{Pre} = \frac{\text{TP}}{\text{TP+FP}} \tag{14}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP+FN}} \tag{15}$$

$$\text{F1} = \frac{2\text{Pre}\cdot\text{Rec}}{\text{Pre+Rec}} \tag{16}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP+FN+FP}} \tag{17}$$

$$\text{OA} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{18}$$

$$P_e = \frac{(\text{TN+FN})\cdot(\text{TN+FP})+(\text{FP+TP})\cdot(\text{FN+TP})}{(\text{TP+FP+TN+FN})^2} \tag{19}$$

$$\kappa = \frac{\text{OA}-P_e}{1-P_e} \tag{20}$$

where TP, TN, FP, FN represent the quantities of true positive, true negative, false positive, and false negative, respectively.

B. Datasets

In this experiment, we conduct tests using three datasets, and the details of each dataset are provided as follows.

1) Learning, vision, and remote sensing (LEVIR-CD) [45]: LEVIR-CD is a widely used public large-scale dataset for building CD. It comprises 637 pairs of CD images, each with a size of 1024 × 1024 and a spatial resolution of 0.5 m. Following the dataset processing described in the literature, we divide it into three parts: training, validation, and testing sets. The images are cropped into nonoverlapping blocks of size 256 × 256, resulting in 7120 pairs for training, 1024 for validation, and 2048 for testing. This dataset is among the most extensively utilized for building CD tasks.

2) WHU building dataset (WHU-CD) [62]: WHU-CD is a large and accurate open-source dataset of aerial and satellite images provided by Wuhan University. This aerial dataset contains over 220 000 independent buildings in Christchurch, New Zealand, covering rural, residential, cultural, and industrial areas. The dataset consists of a pair of aerial images with dimensions of 32 507 × 15 354 pixels and a spatial resolution of 0.075 m. As the authors have not provide a specific dataset segmentation plan, we crop the images into nonoverlapping blocks of size 256 × 256. We randomly split the dataset into three parts, comprising 5205 pairs for training, 743 for validation, and 1486 for testing, ensuring that the three parts have a similar distribution. While this dataset has fewer instances compared to the LEVIR-CD dataset, it boasts a higher resolution.

3) SYSU-CD [63]: The SYSU-CD dataset comprises 20 000 pairs of dual-temporal remote sensing images, each with a size of 256 × 256 pixels and a spatial resolution of 0.5 m. It includes various types of complex scene changes, such as road expansion, new urban construction, vegetation changes, suburban expansion, and preconstruction groundwork. The dataset has been divided into three parts for training, validation, and testing, with 12 000, 4000, and 4000 pairs, respectively. Due to the presence of vegetation changes and other variations in the dataset, the detection of vegetation changes in the test set may be significantly affected by factors, such as lighting, shadows, and seasons. Moreover, the annotation accuracy of this dataset is relatively lower than that of the other two datasets. It contains some overlapping regions, and the extraction of dual-temporal feature maps can have a substantial impact on the model's performance. In this experiment, the SYSU-CD dataset is primarily use to evaluate the model's performance in challenging environments compared to other models.

C. Comparison With SOTA

We compare our proposed method with several state-of-the-art approaches, including three convolution-based methods: FC-EF [28], FC-Siam-Di [28], FC-Siam-Conc [28]; two attention-based methods: STA [45] and IFN [46]; a transformer-based method BIT [18]; and three recent methods: A2-Net [52] utilizing a lightweight network, SLDDNet [22] incorporating a CNN-transformer hybrid encoder and SEIFNet [54] enhancing the exploration of time differences and the utilization of multiscale features. We conduct experiments using the parameters suggested in the respective papers for each of these methods as follows.

1) FC-EF [28] proposed an image-level fusion method, where the concatenated dual-temporal images were fed into a fully convolutional network (FCN) to extract semantic features.
2) FC-Diff [28] was one of the first siamese extensions of FCNs. It employed siamese FCNs to extract multilevel features and fused the temporal information through difference of features during the step-by-step upsampling process.
3) FC-Conc [28] was one of the first siamese extensions of FCNs. It employed siamese FCNs to extract multilevel features and fused the temporal information through concatenation of features during the step-by-step upsampling process.
4) STA [45] introduced a metric-based siamese FCN-based method, integrating temporal–spatial attention mechanisms to expand more features.
5) IFN [46] proposed a deep supervision image fusion network. It used attention modules to merge deep features from the original images and change image features in the CD network, helping the reconstruction of the the temporal change information.
6) BIT [18] presented a transformer-based method, incorporating transformer into the CD task for improved context modeling of dual-temporal images.
7) A2-Net [52] introduced a new lightweight model (with only 3.78 M parameters and 6.02 G FLOPs). It used supervised attention to progressively merge multilevel features, employing a coarse-to-fine strategy to identify change information.
8) SLDDNet [22] introduced a novel CNN-transformer hybrid encoder for dual-temporal image feature extraction. By parallelizing pyramid structure feature stacking with the transformer semantic selector, it addressed the limitation of existing methods only use a single CNN or transformer architecture for feature extraction.
9) SEIFNet [54] proposed a spatiotemporal enhancement and interval fusion network. A spatiotemporal difference enhancement module with a dual branch structure was designed to obtain the changing features of dual temporal images, and the interlayer features were integrated through an adaptive context fusion module to better reconstruct the detailed information of objects.

We employ publicly available code to invoke the model code for the aforementioned CD networks. The hyperparameters and optimizer are set according to the parameters mentioned in the papers, the default parameters in the publicly available code, and the priority of parameters used in our model.

1) Quantitative Evaluation: Table I reports the overall performance scores on three datasets and presents the model parameters (Params) size and floating point operations (FLOPs) for these methods. Our model's performance in Pre. and Rec. metrics on three datasets is not as outstanding as other advanced models. However, our research focuses on improving F1 scores because it comprehensively calculates Pre. and Rec., It is a more comprehensive evaluation indicator, and by finding a balance point between two indicators, we can obtain more

TABLE I
COMPARISON RESULTS FOR THREE CD TESTSETS

	Params.	FLOPs.	LEVIR-CD					WHU-CD					SYSU-CD				
Methods	(M)	(G)	Pre.	Rec.	F1	IoU	κ	Pre.	Rec.	F1	IoU	κ	Pre.	Rec.	F1	IoU	κ
FC-EF	1.35	3.57	85.23	76.83	80.81	67.80	79.84	75.56	76.03	75.80	61.03	74.85	78.55	71.48	74.85	59.81	67.56
FC-diff	1.35	4.72	92.18	84.72	88.29	79.04	87.69	84.34	79.32	81.75	69.14	81.07	**90.54**	57.29	70.18	54.06	63.51
FC-conc	1.55	5.32	91.54	85.47	88.40	79.21	87.80	81.05	78.53	79.77	66.35	78.99	**85.70**	72.93	78.80	65.02	72.93
STA	33.82	13.16	73.83	**95.78**	83.38	71.50	82.37	82.80	90.23	86.36	75.99	85.80	80.95	78.26	79.58	66.09	73.42
IFN	35.73	82.26	90.96	89.61	90.28	82.28	89.76	91.14	85.33	88.14	78.79	87.69	79.19	**81.45**	80.29	67.08	74.10
BIT	3.50	10.61	92.32	89.45	90.86	83.26	90.38	**94.16**	88.22	91.09	83.64	90.75	82.42	77.87	80.08	66.78	74.16
A2-Net	3.78	3.05	**92.88**	90.14	**91.49**	**84.32**	**91.04**	93.06	**91.91**	**92.48**	**86.01**	**92.19**	84.74	80.83	82.74	70.56	**77.57**
SLDDNet	2.76	5.22	92.19	90.70	91.44	84.23	90.98	90.32	84.13	87.12	77.17	86.63	81.41	80.38	80.89	67.92	75.05
SEIFNet	8.37	27.91	92.02	89.28	90.63	82.86	90.13	92.25	91.64	91.94	85.09	91.63	82.92	**82.71**	**82.82**	**70.67**	77.52
Ours	8.95	25.74	**92.71**	**91.37**	**92.03**	**85.24**	**91.61**	**93.24**	**92.24**	**92.73**	**86.45**	**92.45**	85.50	81.12	**83.25**	**71.31**	**78.26**

The highest score is highlighted in red bold, and the second-highest score is highlighted in blue bold. All scores are represented as percentages (%).

robust and reliable results. Compared with other methods, our model consistently demonstrates superior performance, with F1 scores 0.54%/0.25%/043% higher than the second highest score on the three datasets, IoU higher by 0.92%/0.44%/0.67%, and κ coefficients higher by 0.57%/0.26%/0.69%, indicating the effectiveness of the proposed model. The A2-Net had a low number of Params and FLOPs, and achieved the highest Pre. score in the LEVIR-CD dataset. However, our model has stronger feature expression ability through the two-level feature fusion module, resulting in better experimental performance on other scores in all three datasets. The same phenomenon also occurs in lightweight networks BIT and SLDDNet that use the transformer structure. Compared to BIT that does not use a multifeature fusion structure and SLDDNet that directly uses skip connections, our model improves its feature expression ability by increasing the semantic expression ability of feature maps at different scales at the cost of computational cost. SEIFNet has a better Rec. score in the challenging SYSU-CD dataset, and has a similar number of Params and FLOPs as our model. However, due to the improved feature representation ability of large-scale feature maps through Mixed-conv and AFF Transformer, our model as a whole achieved better scores.

2) Qualitative Evaluation: Fig. 3 illustrates the visual comparisons of these methods on three datasets. For better visualization, we use different colors to represent true positives (white), false positives (red), true negatives (black), and false negatives (green). The white regions indicate correctly detected changed areas, red areas represent unchanged regions incorrectly identified as changed, green denotes changed regions that were not recognized, and black regions denote correctly identified unchanged areas. The results suggest that the proposed method exhibits superiority in the following aspects.

It can be observed that our model achieves a high recall score while maintaining a good precision score. Our model can better learn the distribution information representing the dataset. As shown in Fig. 3 (12), our model better identifies the difference between green buildings and background in preimage and obtains more accurate detection results. As shown in Fig. 3 (10) and (11), compare to other models, our model also better identifies changes in building and vegetation information. As shown in Fig. 3 (6), our model successfully identifies the difference between buildings and background roads in postimage. At the same time, our model can effectively recognize complex terrain information by improving the feature representation ability of large-scale feature maps. As shown in Fig. 3 (3) and (5), our model effectively identifies the difference between the building and shadow information in the image, and successfully detects the area of building changes. In Fig. 3 (7), it accurately identifies the shadows between buildings of different heights. Finally, due to the improvement of feature representation ability of our model in large-scale feature maps, our model also has a certain improvement in detecting the edges of changing regions. Both Fig. 3 (1) and (8) demonstrate better edge segmentation performance.

D. Ablation Studies

To validate the effectiveness of the proposed modules in our network, we conduct a comprehensive ablation study on three datasets.

1) PFFMs: We conduct an ablation study on PFFM by removing it from the model to assess its impact. The goal is to verify if the dual-temporal feature maps pass through the PFFM could produce better results for subsequent operations on the feature maps. In the Table II, we observe a significant decrease in F1 scores on all three datasets when PFFM is removed. We attribute this decrease to two potential reasons: first, the use of PFFM effectively enhances the semantic information contained in the large-scale feature maps, resulting in better performance in subsequent modules. Second, relying solely on ResNet-18 as the backbone may limit the effectiveness of operations on feature maps due to the limited semantic information it contains. The experiments demonstrate the effectiveness of using PFFM for feature map fusion.

2) Mixed-Conv Blocks: To validate the effectiveness of Mixed-convs, we conduct experiments on the parts of the model that used Mixed-convs, replacing them with different configurations, including removal of Mixed-convs, using only standard convolutions, using only dilated convolutions, and using the Mixed-convs. In the Table II, we observe a significant decrease in F1 scores on all three datasets when Mixed-convs are removed. Experiments using standard convolutions as a replacement achieve higher scores, indicating that the information loss during the process of subtracting and taking the absolute value of dual-temporal feature maps is nonnegligible, and convolutions bring local dependencies that can alleviate the negative impact of this information loss to some extent. It is noteworthy that the model using only dilated convolutions performs poorly on

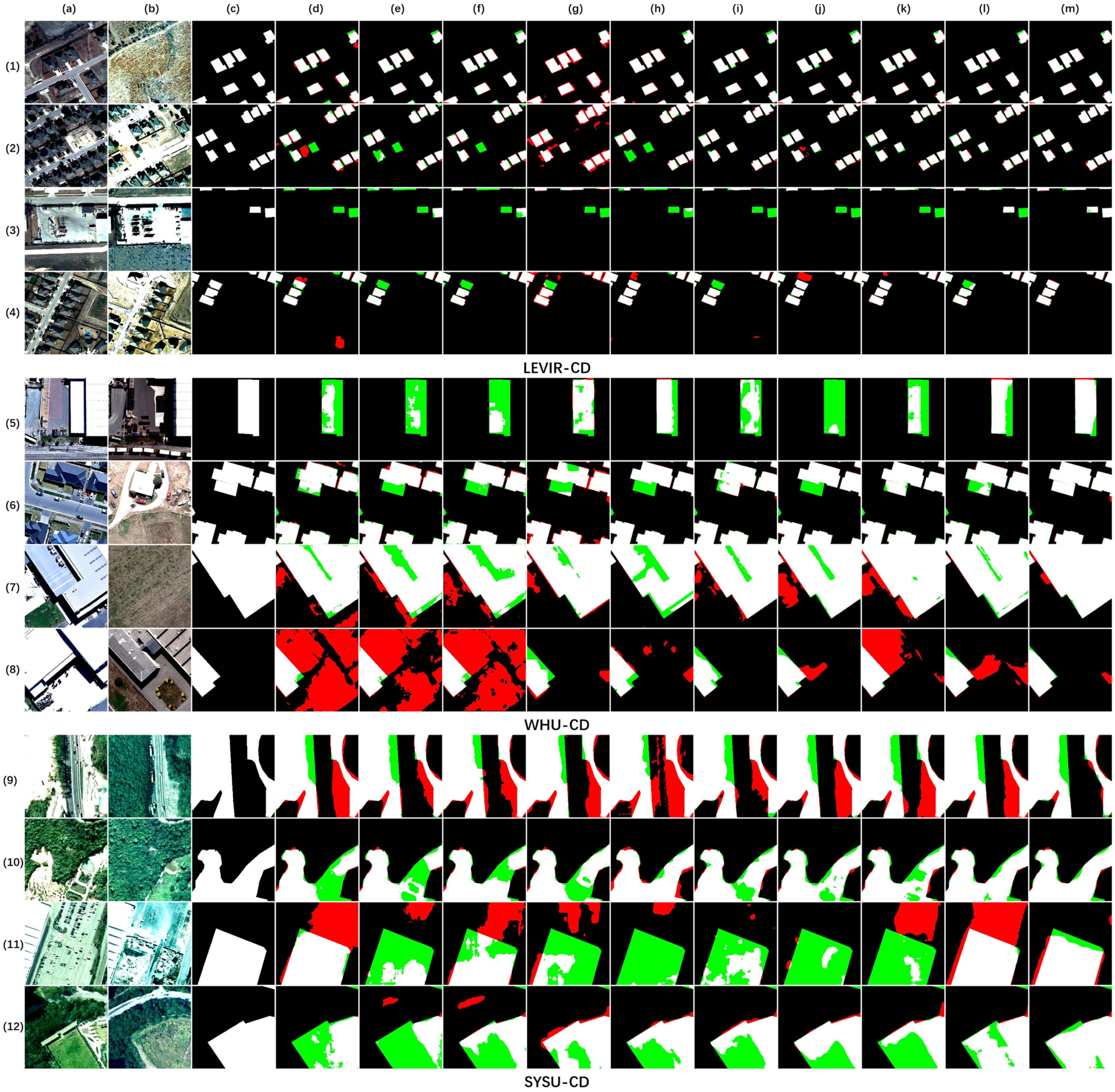

Fig. 3. Visualization results of different methods on the LEVIR-CD, WHU-CD, and SYSU-CD datasets. Various colors use to illustrate the visual outcomes, where white represents true positives, black denotes true negatives, red signifies false positives, and green indicates false negatives. Lines (1) to (12) depict the prediction results of all compared methods on different samples. (a)Preimages. (b) Postimages. (c) Ground truth. (d) FC-EF. (e) FC-Diff. (f) FC-Conc. (g) STANet. (h) IFN. (i) BIT. (j) A2-Net. (k) SLDDNet. (l) SEIFNet. (m) Ours.

TABLE II

Differential Experiments Were Conducted on Three Remote Sensing Image CD Datasets to Quantitatively Compare Pre., Rec., F1 Scores, IoU, and κ Under Different Configurations

	LEVIR-CD					WHU-CD					SYSU-CD				
Methods	Pre.	Rec.	F1	IoU	Kappa	Pre.	Rec.	F1	IoU	Kappa	Pre.	Rec.	F1	IoU	Kappa
with out PFFM	92.81	89.93	91.35	84.08	90.89	94.06	90.46	92.22	85.57	91.93	84.48	80.60	82.50	70.21	77.26
with out Mixed-conv	92.56	90.55	91.55	84.41	91.10	92.90	91.13	92.01	85.20	91.70	84.19	80.64	82.37	70.03	77.09
use standard convolution	92.41	91.24	91.82	84.88	91.38	94.20	91.21	92.68	86.36	92.40	84.76	81.29	82.99	70.92	77.88
use dilated convolution	92.74	90.41	91.56	84.43	91.11	93.13	90.89	92.00	85.18	91.69	83.04	84.13	83.58	71.80	78.47
no transformer	92.64	90.74	91.68	84.64	91.24	93.44	91.22	92.32	85.73	92.02	84.48	80.33	82.35	70.00	77.08
self-attention transformer	92.49	91.06	91.77	84.79	91.33	92.38	91.53	91.96	85.11	91.64	87.44	79.47	83.27	71.33	78.42
cross-attention transformer	92.83	90.93	91.87	84.96	91.44	92.75	92.61	92.68	86.36	92.40	86.20	79.84	82.90	70.80	77.89
full structure	92.71	91.37	92.03	85.24	91.61	93.24	92.24	92.73	86.45	92.45	85.5	81.12	83.25	71.31	78.26

All scores are represented as percentages (%).

TABLE III
EFFECT OF THE DEPTH OF TRANSFORMER ON THREE DATASETS TO QUANTITATIVELY COMPARE F1 SCORES, IOU, AND κ

	LEVIR-CD			WHU-CD			SYSU-CD		
Depth	F1	IoU	κ	F1	IoU	κ	F1	IoU	κ
1	91.91	85.03	91.48	92.40	85.88	92.11	83.11	71.09	78.04
2	92.03	85.24	91.61	92.73	86.45	92.45	83.25	71.31	78.26
4	91.87	84.97	91.44	92.47	86.00	92.18	82.77	70.61	77.56

All scores are represented as percentages (%).

TABLE IV
EFFECT OF THE NUMBER OF MIXED-CONV LAYERS ON THREE DATASETS TO QUANTITATIVELY COMPARE F1 SCORES, IOU, AND κ

	LEVIR-CD			WHU-CD			SYSU-CD		
Layers	F1	IoU	κ	F1	IoU	κ	F1	IoU	κ
[1,1,1,1]	91.71	84.70	91.28	92.13	85.41	91.83	82.98	70.91	77.81
[3,2,1,0]	92.03	85.24	91.61	92.73	86.45	92.45	83.25	71.31	78.26

All scores are represented as percentages (%).

the LEVIR-CD and WHU-CD datasets, with scores similar to not using any module, indicating the importance of adding local dependencies to feature maps. However, on the SYSU-CD dataset with less precise labels, the model achieve higher scores than the other two datasets, showing that increasing the receptive field in the presence of less fine-grained labels can effectively enhance the model's representational ability. The model using Mixed-convs demonstrated excellent performance on all three datasets, proving the effectiveness of the simplicity structure. The overall experiments indicate that increasing the receptive field and adding new local dependencies are effective for feature maps with missing semantic information. Increasing the receptive field of large-scale feature maps can effectively enhance the quality of the temporal change information contained in large-scale feature maps, leading to the increase in computation and parameter costs.

3) AFF Transformer: To validate the effectiveness of the AFF Transformer module, we conduct experiments by testing the model without the transformer structure, using self-attention transformer and using cross-attention transformer. In the Table II, we observe a decrease in F1 scores on all three datasets when the Transformer module is removed, indicating the effectiveness of the transformer structure. The model using self-attention transformer also has better performance, but due to the high number of channels in large-scale feature maps, the self-attention transformer model requires a higher amount of additional computation. It should be noted that the model using cross attention transformer shows lower score compared to AFF Transformer on three datasets. We believe that this is because in the process of fusing advanced semantic features, calculating the similarity between features at different scales is more suitable for the model to learn feature fusion compared to spatial similarity. This also indicates that the AFF Transformer module can achieve highly effective results with limited computational complexity.

E. Parameter Analysis

1) Depth of Transformer: For a network utilizing the transformer model, the number of layers in the transformer is a crucial hyperparameter that needs to be tested. Having too many layers in the transformer may increase model complexity, making it difficult to optimize to the best performance. On the other hand, too few layers might limit the model's effectiveness. We test the impact of different numbers of transformer layers for the feature maps of various dimensions.

In the Table III, we observe that, for the LEVIR-CD and SYSU-CD datasets, changing the number of transformer layers from 1 to 2 does not have a significant impact. However, for the WHU-CD dataset, having two layers in the transformer produces noticeably better results. In addition, as the number of transformer layers further increases, the scores start to decrease. We believe that a deeper transformer architecture can better capture information in the feature maps. However, having too many layers also increases training costs, making it challenging for the model to reach optimal performance within a limited training process, resulting in a decrease score. Therefore, we ultimately chose a model with two layers in the transformer.

2) Number of Mixed-Conv Layers for Different Scale Feature Maps: We conduct tests on the number of Mixed-conv layers used for feature maps of various dimensions. The Mixed-conv layers in the Table IV indicates the number of Mixed-conv layers used from the largest to the smallest scale feature maps. Considering the computational cost of Mixed-conv layers, we test the effects of using [1,1,1,1] and [3,2,1,0] Mixed-conv layers.

In the Table IV, we observe that a higher number of Mixed-conv modules result in better scores. Increasing the receptive field for large-scale feature maps has a positive impact on the overall model, providing more semantic information for larger scale feature maps. However, higher number of Mixed-conv layers also lead to a significant increase in computational and parameter costs. While increasing the training time of the model, it inevitably increases the optimization cost of the model.

F. Discussions

While our model exhibits favorable overall performance, there are still challenging issues to address. Our model has higher parameter and computational requirements compared to newer models. This is mainly attributed to the extensive use of convolutional modules for different-scale feature maps, especially large-scale feature maps. Although this effectively enhances the semantic information contained in large-scale feature maps, it leads to increased parameter demands. Our primary research focus going forward is to explore methods to include more high-quality semantic information in large-scale feature maps with fewer parameters and computational demands.

V. CONCLUSION

In this work, we propose an effective two-level feature fusion network for CD in remote sensing images. Our model employs the structurally simple ResNet-18 as the backbone network. Through PFFM, it preliminarily integrates semantic information from different-scale feature maps to generate dual-temporal feature maps. After subtracting and taking the absolute value of the dual-temporal feature maps, Mixed-conv are used for large-scale feature maps to enlarge the receptive field and increase the included semantic information. Finally, the AFF Transformer structure is utilized to further integrate semantic features from small-scale feature maps into large-scale feature maps, resulting in large-scale feature maps containing more deep-level semantic information. Our model successfully achieves improvements in F1 score, IoU, and κ on three datasets (LEVIR-CD, WHU-CD, SYSU-CD). The results indicate that enhancing the semantic information contained in large-scale featuree maps can effectively enhance the feature extraction ability of the model, thereby improving the score.

References

[1] A. Singh, “Review article digital change detection techniques using remotely-sensed data,” *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

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

[3] L. Moya et al., “Detecting urban changes using phase correlation and -based sparse model for early disaster response: A case study of the 2018 Sulawesi Indonesia earthquake-tsunami,” *Remote Sens. Environ.*, vol. 242, 2020, Art. no. 111743.

[4] R. Liu, M. Kuffer, and C. Persello, “The temporal dynamics of slums employing a CNN-based change detection approach,” *Remote Sens.*, vol. 11, no. 23, 2019, Art. no. 2844.

[5] L. Bruzzone and S. B. Serpico, “An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 4, pp. 858–867, Jul. 1997.

[6] P. P. De Bem, O. A. de Carvalho Junior, R. Fontes Guimarães, and R. A. Trancoso 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.

[7] Z. Zhang, G. Vosselman, M. Gerke, D. Tuia, and M. Y. Yang, “Change detection between multimodal remote sensing data using siamese CNN,” 2018, *arXiv:1807.09562*.

[8] J. Chen, H. Liu, J. Hou, M. Yang, and M. Deng, “Improving building change detection in VHR remote sensing imagery by combining coarse location and co-segmentation,” *ISPRS Int. J. Geo- Inf.*, vol. 7, no. 6, p. 213, 2018.

[9] R. Qin, J. Tian, and P. Reinartz, “3D change detection–approaches and applications,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 122, pp. 41–56, 2016.

[10] Y. Ban and O. Yousif, “Change detection techniques: A review,” in *Multitemporal Remote Sensing: Methods and Applications*. Berlin, Germany: Springer, 2016, pp. 19–43.

[11] T. Liu, L. Yang, and D. Lunga, “Change detection using deep learning approach with object-based image analysis,” *Remote Sens. Environ.*, vol. 256, 2021, Art. no. 112308.

[12] A. Vaswani et al., “Attention is all you need,” in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[13] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations*, 2020, *arXiv:2010.11929*.

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

[15] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[17] X. Mao et al., “Enhance the visual representation via discrete adversarial training,” in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 7520–7533.

[18] 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.

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

[20] 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.

[21] X. Su, J. Li, and Z. Hua, “Transformer-based regression network for pansharpening remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5407423.

[22] Z. Fu, J. Li, L. Ren, and Z. Chen, “SLDDNet: Stage-wise short and long distance dependency network for remote sensing change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3000319.

[23] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[24] J. Liu, M. Gong, K. Qin, and P. Zhang, “A deep convolutional coupling network for change detection based on heterogeneous optical and radar images,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Mar. 2018.

[25] M. Wang, K. Tan, X. Jia, X. Wang, and Y. Chen, “A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images,” *Remote Sens.*, vol. 12, no. 2, 2020, Art. no. 205.

[26] W. Zhao, L. Mou, J. Chen, Y. Bo, and W. J. Emery, “Incorporating metric learning and adversarial network for seasonal invariant change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2720–2731, Apr. 2020.

[27] X. Tang et al., “An unsupervised remote sensing change detection method based on multiscale graph convolutional network and metric learning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609715.

[28] 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.

[29] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, “Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15173–15182.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] M. Hao, W. Shi, H. Zhang, and C. Li, “Unsupervised change detection with expectation-maximization-based level set,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 210–214, Jan. 2014.

[35] H. Li, M. Li, P. Zhang, W. Song, L. An, and Y. Wu, “SAR image change detection based on hybrid conditional random field,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 910–914, Apr. 2015.

[36] T. Celik, “Unsupervised change detection in satellite images using principal component analysis and k-means clustering,” *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[37] E. P. Crist, “A TM tasseled cap equivalent transformation for reflectance factor data,” *Remote Sens. Environ.*, vol. 17, no. 3, pp. 301–306, 1985.

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

[39] P. R. Coppin and M. E. Bauer, “Digital change detection in forest ecosystems with remote sensing imagery,” *Remote Sens. Rev.*, vol. 13, no. 3/4, pp. 207–234, 1996.

[40] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, “Deep learning-based change detection in remote sensing images: A review,” *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 871.

[41] 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.

[42] 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.

[43] Y. Sun, L. Lei, X. Tan, D. Guan, J. Wu, and G. Kuang, "Structured graph based image regression for unsupervised multimodal change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 185, pp. 16–31, 2022.

[44] Y. Sun, L. Lei, X. Li, X. Tan, and G. Kuang, "Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4700221.

[45] 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.

[46] 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.

[47] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, vol. 1804, pp. 1–6.

[48] 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.

[49] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-SiamNet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 484.

[50] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603216.

[51] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.

[52] Z. Li et al., "Lightweight remote sensing change detection with progressive feature aggregation and supervised attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602812.

[53] B. Hou, Q. Liu, H. Wang, and Y. Wang, "From W-Net to CDGAN: Bitemporal change detection via deep learning techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1790–1802, Mar. 2020.

[54] Y. Huang, X. Li, Z. Du, and H. Shen, "Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5609414.

[55] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[56] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460.

[57] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossVit: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 347–356.

[58] Y. Zhou, C. Huo, J. Zhu, L. Huo, and C. Pan, "DCAT: Dual cross-attention-based transformer for change detection," *Remote Sens.*, vol. 15, no. 9, 2023, Art. no. 2395.

[59] 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.

[60] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8026–8037.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[62] 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.

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

Mingyao Feng received the B.S. degree in software engineering from Shandong University, Jinan, China, in 2021, where he is currently working toward the M.S. degree in software engineering.

His research interests include change detection, remote sensing image analysis, and machine learning.

Ruifan Zhang received the B.S. degree in software engineering from Qingdao University, Qingdao, China, in 2021. He is currently working toward the M.S. degree in artificial intelligence with Shandong University, Jinan, China.

His research interests include anomaly detection and machine learning.

Hao Wang received the B.S. degree in software engineering from Nanchang University, Nanchang, China, in 2021. He is currently working toward the M.S. degree in software engineering with Shandong University, Jinan, China.

His research interests include machine learning and defect detection.

Yikun Liu received the B.S. degree in mechanical and electronic engineering from Huazhong Agriculture University, Wuhan, China, in 2019, and the M.S. degree in software engineering from Shandong University, Jinan, China, in 2022, where he is currently working toward the Ph.D. degree in software engineering.

His research interests include remote sensing image analysis and machine learning.

Gongping Yang received the bachelor's degree in computer and application, and master's and Ph.D. degrees in computer software and theory from Shandong University, Jinan, China, in 1992, 2001, and 2007, respectively.

Since 2013, he has been a Professor with the School of Software, Shandong University. His research interests include pattern recognition, computer vision, and remote sensing image analysis.

Dr. Yang is a Senior Member of CCF and CAAI, also serve as the Machine Learning Technical Committee of CAAI and Artificial Intelligence and Pattern Recognition Technical Committee of CCF. He was also a Program Committee Member/Organization/Program Chair of several conferences.