CGMMA: CNN-GNN Multiscale Mixed Attention Network for Remote Sensing Image Change Detection

Yan Zhang, Xinyang Song, Zhen Hua, and Jinjiang Li

***Abstract*—Remote sensing change detection (CD) networks have been increasingly powerful with the application of convolutional neural networks (CNNs) and transformers. The CNN-based CD method with a CNN backbone has been widely used and plays an significant role. In complex spatial relationships within remote sensing images, CNNs may face limitations due to the restricted receptive field, making it challenging to handle intricate pixel relationships effectively. Therefore, to address this limitation of CNNs, we introduce vision graph neural network (ViG) to tackle the constrained receptive field issue. In addition, we propose a backbone network named Congraph, which integrates convolution and graph interaction. Congraph simultaneously leverages local information from CNNs and global information from GNNs, enabling more comprehensive feature extraction for more accurate change detection. Furthermore, we introduce a multiscale mixed attention (MMA) module to make the model focus on different scale feature information. MMA replaces small-scale features in the multilayer encoder with self-attention to capture global feature information within small-scale features. Finally, we feed bitemporal features into a transformer module to obtain feature difference information and generate the ultimate feature difference map. Through extensive experiments on the LEVIR-CD, WHU-CD, and GZ-CD datasets, our method demonstrates more significant performance advantages compared to the current state-of-the-art change detection methods.**

***Index Terms*—Attention mechanism, convolutional neural network (CNN), graph convolution, remote sensing change detection (CD), transformer.**

I. Introduction

REMOTE sensing change detection (CD) is a vital research area in the remote sensing community. The detection of changes occurring on the Earth's surface over time has significant implications. This research typically focuses on high-resolution satellite remote sensing images captured over the same area at different points in time [1]. Various image processing and pattern recognition techniques are employed to extract change information from the multitemporal remote sensing data to detect and characterize surface changes in relevant areas.

Fig. 1. Remote sensing image building change detection task diagram. The yellow boxes indicate nonchange regions possessing different features, and the red boxes indicate nonchange regions with nontarget features.

The CD task has various application scenarios based on the variation markers in the given scenario. Currently, CD has been successfully applied in urban management [2], damage assessment [3], deforestation [4], environmental monitoring [5], and cropland change [6]. For example, in Fig. 1, even the same target object shows slight feature differences between the yellow boxes. In addition, the performance requirements are higher because some regions are considered pseudovarying regions, and therefore should not be labeled in the final binary detection map, such as the region in the red box with seemingly varying features.

Early traditional CD images used algebraic operations to calculate the differences in RS data, enabling the processing of lower resolution RS data. Typically, clustering [7], threshold-based approaches [8], or change vector analysis [9] were used to compute binary CD images based on image segmentation techniques.

With the significant advancements in deep learning (DL) for image processing, the efficiency of remote sensing image processing has greatly improved. Numerous convolutional neural network (CNN)-based methods have been introduced into CD tasks [10], [11], and [12]. Building upon the foundation of pure convolutional CD methods, some researchers have attempted to stack additional convolutional layers [13], [14], [15], [16], [17], and [18] or utilize expanded convolutions [15] to extend the receptive field (RF). Attention mechanisms have also emerged as a new direction to expand the RF for better contextual modeling [13], [14], [19], [11], and [20]. Attention-based CD

Manuscript received 30 October 2023; revised 30 December 2023; accepted 22 January 2024. Date of publication 25 January 2024; date of current version 29 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62272281, Grant 62002200, Grant 61972235, Grant 62202268, and Grant 62002200, and in part by the Shandong Natural Science Foundation of China under Grant ZR2023MF026 and Grant ZR2021MF068. *(Corresponding author: Zhen Hua.)*

Yan Zhang, Xinyang Song, and Jinjiang Li are with the School of Information and electronic engineering, Shandong Technology and Business University, Institute of Network Technology (ICT), Yantai 264005, China.

Zhen Hua is with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: huazhen@sdtbu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3358298

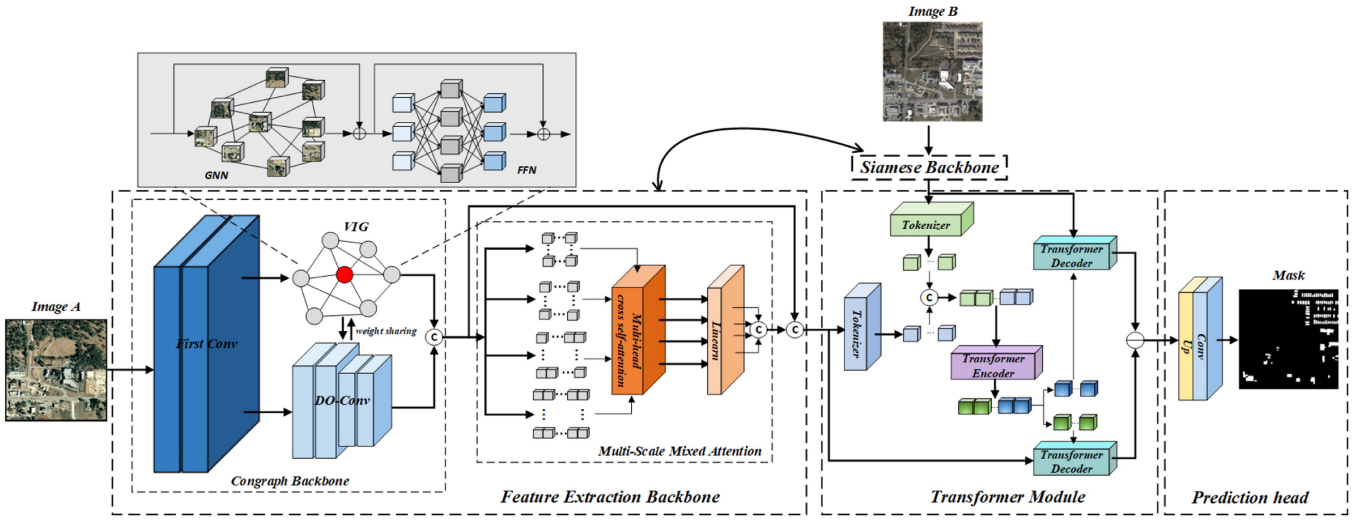

Fig. 2. Overall network architecture of CGMMA, which includes siamese feature extraction backbone and transformer module. The feature extraction backbone consists of Congraph and a multiscale mixed attention module. The transformer module consists of a tokenizer, a transformer encoder and a transformer decoder.

methods have been developed to improve detection accuracy to some extent. However, previous channel or space-based attention mechanisms rely on stacked convolutional layers for feature extraction backbone. Although self-attention has a good effect of modeling interpixel relationships, the computational power consumption increases.

With the successful adoption of transformer in the CD task [21], their powerful contextual modeling capability has been utilized to capture longer range dense relationships, which compensates for the limitation of traditional convolutional and attentional mechanisms. In fact, even using a shallow CNN backbone, transformer-based approaches have achieved better results than deep pure convolutional methods. A large number of transformer-based methods have been proposed [22], [23], and [24], demonstrating their significant prominence in the field of CD.

In recent years, graph convolution-based image processing methods in the remote sensing field have received increasing attention. Remote sensing images have high dimensionality, complex textures, and large scales, making traditional image processing methods often unable to handle this type of data. Therefore, graph convolution-based methods have become a new solution. Graph convolution is a technique for convolutional operations on graph data, different from traditional pixel-based convolution. It treats the image as a graph, with each pixel corresponding to a node in the graph and the relationships between pixels corresponding to edges in the graph. By performing convolution operations on the nodes and edges of the graph, spatial features and local structures in the image can be effectively captured. In remote sensing image processing, graph convolution-based methods can be used for tasks, such as image classification [25], [26], and [27], object detection [28], and image segmentation [29]. Currently proposed graph convolutional networks use the capabilities of both CNNs and GCNs to capture the dependencies between features, enhance feature representation, and achieve good results [30], [31], [32], and [33]. For example, Liu et al. [25] proposed a heterogenous deep network called CNN-enhanced GCN (CEGCN), where complementary feature maps are generated at the pixel and superpixel levels. Similarly, methods such as [26] and [27] also use a combination of CNN and GCN. Although these methods achieve feature complementarity between CNNs and GCNs, this is only a simple feature interaction mechanism, and the intermediate features of CNNs and GCNs are still ignored.

In this article, we introduce a CNN-GNN interactive multiscale mixed attention (CGMMA) remote sensing image CD network, as shown in Fig. 2. In our approach, we propose a high-performance hybrid backbone called Congraph and MMA module that effectively integrates feature information. Congraph allows the CNN branch and GNN branch to complement each other and make full use of their respective information features, which can maximize the perception of local and global features. MMA can better model overall feature information based on information extracted at different scales, enabling better information interaction between features at different scales, which compensates for the global or local features that we ignore without stacking downsampling layers, and has lower computational cost than multihead self-attention (MSA). Finally, we use the bitemporal image features extracted by Congraph and MMA, convert them into tokens, and feed them into a transformer to capture their temporal changes and generate a feature difference map.

Our contribution can be summarized as follows.

1) We designed a hybrid backbone called Congraph. It consists of multiple stages of feature interaction using high-performance DO-Conv, which combines traditional convolution with depthwise convolution, and the vision graph neural network (ViG), which has a wide RF. This design allows for maximum perception of local and global features.
2) We propose an MMA module that balances global and local features, which are used to extract image details (e.g., building edges and shapes) and global structures (e.g.,

target colors and textures). Furthermore, the multiscale cross-attention (MSCA) is used to adaptively focus on features at different scales, thereby improving the performance of the model.

3) We outperform current state-of-the-art CD methods in terms of performance. We conducted extensive experiments on three datasets: LEVIR-CD, WHU-CD, and GZ-CD, and the results demonstrate that our method performs well on all datasets, with even better performance on the most authoritative dataset LEVIR-CD.

The rest of this article is organized as follows. Section II describes the nontransformer-based approach with the transformer based CD approach under the DL approach. In Section III, we describe the design ideas in our method in detail. In Section IV we summarize and analyze a large number of experimental reports. Finally, Section V concludes this article.

II. Related Work

As DL technology continues to advance, numerous novel methods based on convolution, attention mechanisms, and transformers have emerged for the CD task. Currently, these methods have made significant efforts to explore how to better implement the CD task, resulting in many innovative approaches based on different techniques.

According to whether a training sample set is established and labeled samples are used for learning, existing CD methods can be divided into unsupervised CD methods [34] and [35] and supervised CD methods [10] and [36]. Currently, researchers are focusing their efforts on supervised CD methods. Moreover, based on how the network processes remote sensing images, CD methods can be classified into single-stream models [37] and [38] and dual-stream models [11], [19], [21], [22], [39], and [40]. In the single-stream model, the bitemporal images are usually fused during the preprocessing stage (based on connection or difference) and treated as a single input to the entire network, which makes the CD method only need to perform the cost of a single network. On the other hand, the dual-stream model is the current mainstream approach, which converts the bitemporal images into feature maps and then uses various DL-based analysis methods to generate CD masks. In the dual-stream model, the feature extraction process of the bitemporal images can be divided into Siam-based [11], [22], and [40] and pseudo-Siam-based [21] and [24] structures. Recently, there has been a growing focus on the global modeling ability of transformer in the CD task. The continuous improvement of transformer's capabilities has made it a promising direction for further research. In addition, GNN which excel in modeling graph structures, have also shown great potential in remote sensing applications [25], [26], [27], and [28] and have been found to outperform transformers in some cases. In the following, we will introduce related works in the field of CD based on three different approaches: CNN-based CD methods, GCN-based CD methods, and transformer-based CD methods.

A. CNN-Based CD Method

The powerful feature representation ability of CNN has been demonstrated in various fields, making it an indispensable part of the CD task. In the early days of CNN methods for CD, Daudt et al. [10] proposed the fully convolutional neural network (FCN) with three processing methods that are still important for current research. This study was the first to attempt to use excellent segmentation models in the CD task using CNN methods and tried different processing methods for single-stream and dual-stream models. However, early traditional CNN methods were limited by RFs and could not provide global information effectively. Therefore, Song et al. [38] introduced dilated convolution to replace traditional convolution, which increased the RF and improved the contextual modeling ability.

In addition, in order to obtain better feature extraction ability, attention mechanisms have become an integral part of CNN networks and have been continuously improved, playing an important role. Based on Daudt et al.'s work [10], Li et al. [41] used the feature attention ability brought by the attention mechanism to further improve the processing effect of feature difference maps. This also proves that attention mechanism does have a good improvement on the basis of CNN. Therefore, with the development of attention mechanism, a large number of innovative new CD methods have emerged. For example, Fang et al. [40] used channel attention to fuse different scale feature information on the basis of the Unet++ model, and achieved certain improvements in performance compared to pure CNN networks. Chen et al. [13] incorporated a self-attention to emulate spatiotemporal relationships, integrating it into the feature extraction process, resulting in the design of a self-attention-based CD network.

Simply relying on the attention mechanisms for improving the network's performance is not enough. Methods, such as dense connections, which integrate multiscale features [40] or deep supervision [11] also play a crucial role in enhancing performance. While CNN-based and attention-based approaches have improved feature representation, FCN-based frameworks with fixed local RFs are limited in their ability to model long-range dependencies. Furthermore, to achieve superior results, CNN backbones have become increasingly complex, with many networks stacking additional attention modules on top of already deep convolutional stages.

Based on our analysis, we propose a CD method that utilizes MMA module. We leverage the long-range modeling capability brought by self-attention and design a mixed method for multiscale self-attention, which considers both local and global features. We have integrated the MMA module into the Congraph hybrid backbone, resulting in more performance improvements in the overall network.

B. GCN-Based CD Method

In high-resolution remote sensing images, the images are characterized by high dimensions, complex textures, and large scales, which make it difficult for traditional image processing methods to handle such data. Due to the loss of positional information and global context information, most existing CNN methods are insufficient to extract more details and global features, resulting in incomplete and discontinuous results. In other remote sensing fields, the combination of CNN and GCN has made great progress, which also proves that GCN has a good performance in processing remote sensing images.

In remote sensing image research, many researchers have combined GCN and CNN into a dual-stream model structure to interact the features of GCN and CNN. For example, Liu et al. [25] proposed a heterogeneous deep network called CEGCN, which utilizes the advantages of both CNN and GCN and generates complementary feature maps at pixel and superpixel levels. Similarly, in [28] a similar dual-stream approach using GCN to enhance CNN's feature extraction capabilities was proposed. Liang et al. [27] added information sharing between CNN and GCN in the dual-stream structure.

Another type of structure is a single-stream model that combines GCN and CNN, such as in [26] and [29], where GCN is stacked on top of CNN in a single-stream model.

The aforementioned methods can achieve feature complementarity between CNN and GNN, but they only provide a simple mechanism for feature interaction, ignoring more intermediate features between CNN and GNN. Therefore, we hope to transmit more information separately to the GNN branch and CNN branch, so that the information from both branches can interact at every step of the local and global feature information. Therefore, we designed the Congraph model to allow full information complementarity between GNN and CNN.

C. Transformer-Based Method

Transformer [42] has demonstrated strong capabilities in various computer vision applications, including remote sensing CD. In comparison to traditional attention mechanisms, transformer offers irreplaceable advantages due to its nonlocal attention mechanism. This mechanism can establish better global feature correlations, which makes it ideal for remote sensing applications. Moreover, position information can be used to enhance the modeling ability of context for long-term correlations between pixel features, which further improves the performance of transformer.

Currently, transformer-based methods for CD tasks can be divided into two structures: the twin transformer method [22] and [23] and the pseudotwin transformer method [21]. Both of these methods have achieved good results. In [21], Chen et al. first introduced transformer into CD tasks and achieved good results using only a shallow ResNet as the feature extraction backbone. This also proves that transformer has more complete modeling capabilities in the spatiotemporal domain, and this approach achieves performance improvement while maintaining low computational cost compared to previous CD methods. In the parameter-sharing [22], transformer is used as the encoder and completely replaces the use of CNN. Multiple layers of transformer are used for parameter sharing, and decoding is performed in a multilayer perceptron (MLP) to achieve CD. Similarly, in [23], Swin Transformer is used instead of transformer for encoding. This is also an important exploration of pure transformer structures in CD tasks. However, this pure transformer-based approach increases the network's computational cost to some extent, although it can achieve better performance by using more long-term connections.

As an efficient module for modeling global context, transformer's application in the field of CD is an inevitable trend.

Algorithm 1: Implementation Process of Our CGMMA Model.

Input: A, B $(bitemporal\ image))$
Output: $M(a\ prediction\ change\ mask)$
// $step1 : Congraph\ feature\ extraction$
$\mathbf{F_1^{HB}} = Congraph(\mathbf{A})$
$\mathbf{F_2^{HB}} = Congraph(\mathbf{B})$
// $step2 : Multi - Scale\ mixed\ Attention\ Module$
for $i\ in\ \{1, 2\}$ **do**
$\mathbf{F_i}^{\mathbf{4}}, \mathbf{F_i}^{\mathbf{3}}, \mathbf{F_i}^{\mathbf{2}}, \mathbf{F_i}^{\mathbf{1}} = AvgPool(\mathbf{F_i^{HB}})$
$\mathbf{F_i^{MS}} = MMA(\mathbf{F_i}^{\mathbf{4}}, \mathbf{F_i}^{\mathbf{3}}, \mathbf{F_i}^{\mathbf{2}}, \mathbf{F_i}^{\mathbf{1}})$
end
// $step3 : Fusion\ of\ multi - stage\ features$
for $i\ in\ \{1, 2\}$ **do**
$\mathbf{F_i} = CAT(\mathbf{F_i^{HB}}, \mathbf{F_i^{MS}})$
end
// $step4 : Feature\ difference\ extraction$
for $i\ in\ \{1, 2\}$ **do**
$\mathbf{T_i} = Semantic\ Tokenizer(\mathbf{F_i})$
$\mathbf{F_i}^{\mathbf{new}} = Transformer(\mathbf{F_i}, \mathbf{T_i})$
end
// $step5 : Obtain\ the\ change\ mask\ by\ prediction\ head$
$\mathbf{M} = PH(|\mathbf{F_1^{new}} - \mathbf{F_2^{new}}|)$

Moreover, it has been demonstrated in BIT [21] that transformer has good performance in extracting image difference information. Therefore, by leveraging the transformer module, we can further improve the effect of feature extraction and calculate the difference image of features.

III. Methodology

In this article, we propose a CGMMA for remote sensing image CD. The overall architecture of the network is presented in Fig. 2. Our proposed method comprises a Siamese feature learning backbone and a transformer module. The feature learning backbone includes the Congraph hybrid backbone and the MMA module. To provide a more intuitive understanding of our method, we present the detailed reasoning process in Algorithm 1.

A. Congraph Hybrid Backbone

1) Overview: CNN's ability in feature learning tasks is undeniable. In recent CD research, CNN still plays a crucial role. Most studies enhance the feature extraction ability and improve network performance by increasing the number of linear and nonlinear layers in CNN. The recently proposed DO-Conv [43] has shown stronger feature extraction ability than traditional convolutional methods, and achieved significant performance improvement with minimal parameter increase. However, this pure CNN approach is still limited by the RF and difficult to obtain better global features.

To obtain more comprehensive feature information, we consider using ViG, which have a broad and flexible view, to complement CNN features. Graph representation is a general data structure. Compared with grid or sequence methods, graphs

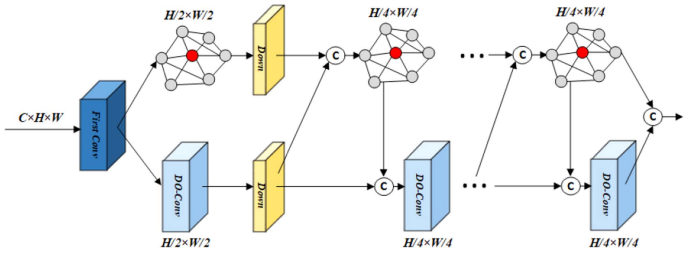

Fig. 3. Construction of Congraph includes two branches, the CNN branch and the GNN branch. Each branch will pass its own feature information to the other branch for interaction to enhance the feature extraction ability and generate more complete feature maps.

can more flexibly handle complex objects with irregular shapes, which is suitable for processing remote sensing images. Compared to vision transformer (ViT), ViG models the relationships between nodes, thus better handling structural information and global relationships in images.

In order to utilize both local and global features for better feature extraction, we designed a concurrent network structure named Congraph, as shown in Fig. 3. Our design is inspired by Conformer [44]. Congraph is a dual-stream model consisting of a CNN branch and a GNN branch. Considering the complementary nature of the two types of features, we integrate the features from the ViG branch and the DO-Conv branch to perform information interaction and execute information fusion of the two branches in parallel. We will explain the different parts of Congraph as follows.

2) Depthwise Overparameterized Convolutional: We employ DO-Conv [43] as the primary unit in the CNN branch. DO-Conv offers advantages over traditional convolution in that it combines depthwise and traditional convolution, delivering superior results under equal computational load. Let us consider an input patch as $P \in \mathbb{R}^{C\times M\times N}$, where the trainable kernels for depthwise convolution are represented by $D \in \mathbb{R}^{(M\times N)\times D_{\text{mul}}\times C_{\text{in}}}$, and for standard convolution by $W \in \mathbb{R}^{C_{\text{out}}\times D_{\text{mul}}\times C_{\text{in}}}$. Here, $D_{\text{mul}} = M \times N$ denotes the depth multiplier for depthwise convolution, where M and N are the spatial dimensions of the patch. C_{in} stands for input channels, and C_{out} represents output channels. We define the DO-Conv operation as $\circledast$, and its workflow can be represented as follows:

$$\begin{aligned} O &= (D, W) \circledast P \\ &= \left(D^T \circ W\right) * P \end{aligned} \tag{1}$$

where $*$ means traditional convolution, $\circ$ means depthwise convolution, and $D^T \in \mathbb{R}^{D_{\text{mul}}\times(M\times N)\times C_{\text{in}}}$ is the transpose of $D \in \mathbb{R}^{(M\times N)\times D_{\text{mul}}\times C_{\text{in}}}$ on the first and second axes. The RF of DO-Conv remains $M \times N$ throughout the operation. Although DO-Conv slightly increases the number of parameters compared to traditional convolution, its depthwise and pointwise convolutions can be independently calculated, making the model more computationally efficient. Moreover, the separation of depthwise and pointwise convolutions in DO-Conv makes it easier to interpret the model's intermediate results and feature maps, leading to a better understanding of how the model works. In summary, DO-Conv offers several advantages, including fewer parameters, high computational efficiency, strong generalization performance, and interpretability. These advantages enable CNNs to achieve good performance without significantly increasing the number of parameters.

3) Vision GNN: The GNN branch operates on the input feature $F \in \mathbb{R}^{\mathbb{C}\times\mathbb{H}\times\mathbb{W}}$. First, the image is divided into N patches as in ViT. By converting these patches into a feature vector $x_i \in \mathbb{R}^D$, we obtain $X = [x_1, x_2, \ldots, x_N]$, where D is the feature dimension. To represent positional information, we add a positional encoding vector to each node feature

$$x_i \longleftarrow x_i + e_i \tag{2}$$

where $e_i \in \mathbb{R}^D$. The features in X are treated as a set of unordered graph nodes, represented as $V = \{v_1, \ldots, v_n\}$. Each node v_i can find K nearest neighboring nodes $N(v_i)$ and add an edge e_{ij} from $v_j \in N(v_i)$ to v_i. In this way, we can obtain a graph $\mathcal{G} = (V, \varepsilon)$, where ε represents all edges.

A graph is constructed based on the features of $\mathcal{G}$,and the ability to aggregate the features of neighboring nodes and exchange information through the graph convolution layer is performed as follows:

$$\begin{aligned} \mathcal{G}' &= F(\mathcal{G}, W) \\ &= \text{Update}\left(\text{Aggregate}\left(\mathcal{G}, W_{\text{agg}}\right), W_{\text{update}}\right) \end{aligned} \tag{3}$$

where W_{agg} and W_{update} are learnable weights for aggregation and update operations, respectively. The aggregation operation computes the current node's feature by aggregating the features of all neighboring nodes surrounding each node

$$x_i' = h(x_i, g(x_i, N(x_i), W_{\text{agg}}), W_{\text{update}}) \tag{4}$$

where $N(x_i)$ is the set of x_i neighbor nodes. For convenience and efficiency, maximum relative graph convolution [45] is used here. The aggregated features are divided into h heads, which are then updated using different weights and finally concatenated into the final value of

$$x_i' = \left[h_1 W^1_{\text{update}}, \ldots, h_n W^n_{\text{update}}\right]. \tag{5}$$

This update operation allows information within subspaces to be processed in parallel while simultaneously updating information in the spatial domain.

In previous GNNs, it was common to aggregate features extracted by iterative graph convolutional layers. The oversmoothing phenomenon in deep GCNs [46] and [47] reduces the degree features of node features, leading to a decline in performance for target area recognition. To alleviate this problem, ViG blocks introduce more feature transformations and nonlinear activations. After the graph convolution, a linear layer is employed to project node features into a common domain, enhancing feature diversity. To prevent layer collapse, a nonlinear activation function is introduced following the graph convolution. This enhanced module is referred to as the Grapher module, denoted as

$$Y = \sigma\left(\text{GraphConv}\left(XW_{\text{in}}\right)\right)W_{\text{out}} + X \tag{6}$$

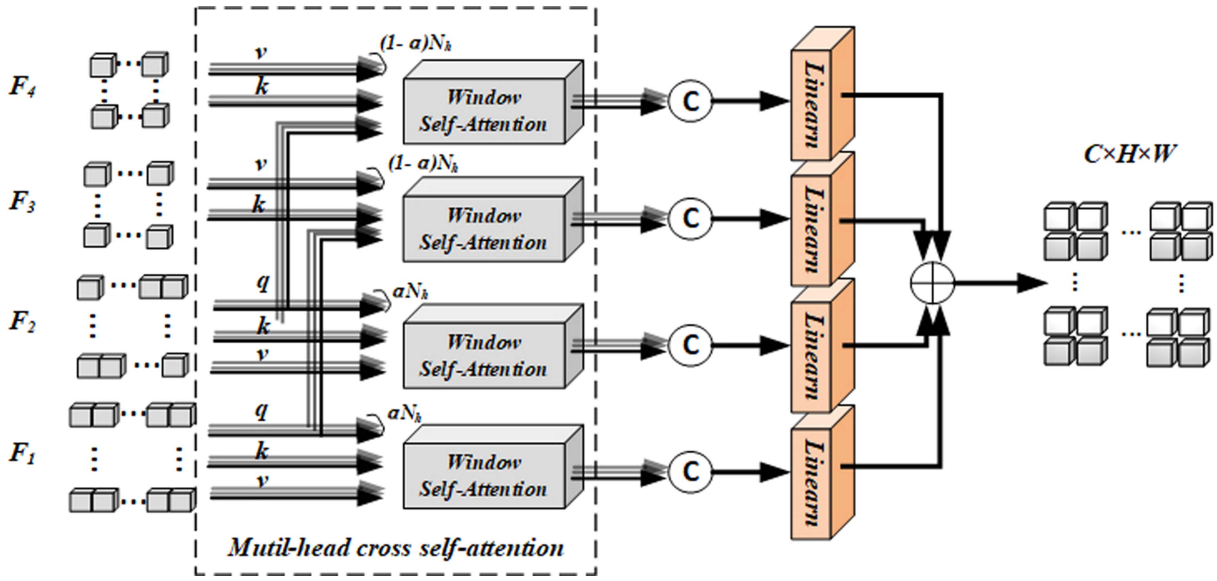

Fig. 4. Construction details of the multiscale mixed attention module are as follows. First, the input features are processed into four different scales of feature maps through a pooling layer, and then input to the multihead cross self-attention (MSCA) module. The larger scale features are responsible for extracting fine-grained features, while the smaller scale features are responsible for extracting global features. The local features and global features are then fused using MSCA, thereby generating multiscale features.

TABLE I
CONGRAPH'S STRUCTURE, INCLUDING THE CNN BRANCH AND THE GNN BRANCH

stage	output	CNN branch	Feature fusion	GNN branch
c1	128 × 128	3×3, C=3→64, stride=2		
		3×3, C=64		
c2	128 × 128	3 × 3, C=64, D=3		K=9, C=64, N=768, D=2
c3	64 × 64	downsample,stride=2 ↓ 3 × 3, C=192, D=16	→ ←	downsample,stride=2 K=9, C=128, N=768, D=6 ↓
c4	64 × 64	1 ×1,192→128 ↓ 3 × 3, C=384→256, D=3	→ ←	↓ K=9, C=256, N=768, D=2 ↓

D represents the number of blocks, C represents the number of channels, K represents the number of neighboring nodes, N represents the number of nodes, and → indicates the direction of feature transmission.

where $Y \in \mathbb{R}^{N\times D}$, W_{in} and W_{out} are the weight of fully connected layers, σ denotes the activation function.

To enhance the capability of feature transformation and alleviate the smoothing effect, each node is equipped with a feedforward network (FFN). The FFN module is a simple MLP composed of two fully connected layers, as illustrated in the following:

$$Z = \sigma\left(YW_1\right)W_2 + Y \tag{7}$$

where $Z \in \mathbb{R}^{N\times D}$, W_1 and W_2 are the weight of fully connected layers.

The ViG block, which serves as the fundamental building block of the GNN branch, consists of a stacked Grapher and an FFN module.

4) CNN-GNN Feature Interaction: In the final Congraph, we divide it into four steps, as shown in Table I. The operations in the CNN branch are similar to ResNet, with n 3 × 3 DO-Conv layers in each stage. Similarly, in the GNN branch, each stage includes *n* ViG blocks.

First, both the CNN branch and the GNN branch go through a convolutional layer to transform the number of channels and feature size, and then enter the DO-Conv block and ViG block in the second step, respectively. In the DO-Conv block, we use a 3 × 3 convolution and set the number of layers to 3. For the ViG block, we divide it into 16 × 16 patches, set the number of neighbor nodes to 9, and the number of nodes to 768. The number of layers in ViG is set to 2. In the third step, to save computational cost, downsampling is performed first to reduce the feature size to $\mathbb{R}^{C\times H/4\times W/4}$. Then, the features from both branches are fused. The information from the CNN branch is passed into the GNN branch's features, fused, and then passed into the next ViG block. The fused graph information is still divided into 16 × 16 patches, and the number of nodes is set to 768. The number of layers in this ViG block is set to 6. The output of ViG is passed back into the CNN branch for feature fusion and then goes through a 16-layer deep DO-Conv block. The operations in the fourth step are the same as those in the third step, except that the number of layers in the DO-Conv block is increased to 3, and the number of layers in ViG is set to 2. The outputs of both branches in the fourth layer are fused and then passed to the next step.

The reason why the second layer is set to be relatively shallow is that the computational cost of a large size is too high, but we still want to obtain more detailed information. The third layer is the main layer for information extraction. Deep feature extraction enables the network to better learn complex features. Since too much downsampling is not suitable for segmentation tasks, we use it as the main feature extraction size. In the third step, the model has sufficient perception of global and detail information of the image, so there is no need to further increase

the number of layers or perform downsampling in the fourth layer. This step only performs a simple fusion and extraction of the information in the third step, while ensuring that both branches obtain sufficient information and retaining more original features.

This concurrent structure means that the CNN and GNN branches can, respectively, maximize the perception of local and global features, and complement each other in information, achieving better feature extraction.

B. Multiscale Mixed Attention Module

To better integrate local detailed features and global features, we designed MMA module, as shown in the Fig. 4. Using multiscale average pooling to generate multiple scales of features allows the model to adaptively focus on different scale feature information, thereby improving the performance of the model. Then, using MSCA mechanism to weight different scale features can better focus on the information of different scale features. Unlike MSA, which enforces the same global attention on all image blocks without considering other scale features, in MSCA we divide MSA into four paths and use cross self-attention to obtain complete feature information.

First, the input spatiotemporal images are processed through the Congraph backbone to obtain a finely detailed feature map. After undergoing the MMA operation of multiple scale average pooling, four different scales of features, namely F_1, F_2, F_3, and F_4, the sizes are 64, 32, 16, and 8, respectively. F_4 is the coarsest feature layer, while F_1 is the finest feature layer. The encoding object in F_1 is more focused on local details, so for the extraction of detailed features, we use local window self-attention, which is more efficient than using global attention to obtain the desired results. The multihead operation used in F_1 can be represented as follows:

$$\mathrm{MH}(F_1, F_1, F_1) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_n)W^0 \tag{8}$$

where head_i can be expressed as

$$\begin{aligned}\mathrm{head}_i &= \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \\ &= \sigma\left(\frac{QK^T}{\sqrt{D_h}}\right)V\end{aligned} \tag{9}$$

where W_i^Q, W_i^K, and W_i^V are the linear projection matrices for query, key, and value, respectively. The dimension of each head is denoted as D_h. We set F_1 and F_2 to be feature maps that focus more on local details. Therefore, the same operation as in F_1 is applied to F_2.

The coarse feature map F_4 is used in a similar manner as F_3, with the difference being that it has an even lower spatial resolution and a higher focus on global features. The multihead attention operation on F_3 can be expressed as follows:

$$\mathrm{MH}(F_1, F_3, F_3) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_n)W^0 \tag{10}$$

where W^0 is the linear projection matrix for concatenating the multihead attention. The representation of the head here is the same as in (7), but the query is generated from feature F_1.

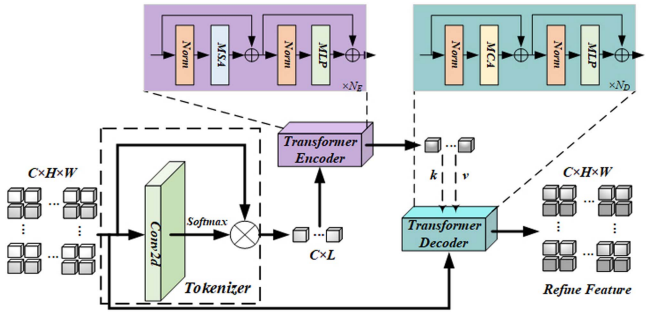

Fig. 5. Construction details of the transformer module. It includes three parts: tokenizer, transformer encoder, and transformer decoder.

In the allocation of heads, to achieve better efficiency, our MSCA assigns double the number of heads in the standard MSA layer to the four scaled features, where F_1 and F_2 are allocated the same number of heads with a split ratio of α. F_3 and F_4 are allocated the remaining heads with a split ratio of $(1-\alpha)$. The attention maps of different scales are then aggregated to generate the final feature output.

C. Transformer Module

After extracting features from the paired images, we pass them through a transformer module for analyzing feature differences. In the transformer module, features from the paired temporal images are first processed by a tokenizer to represent them as several high-level semantic features. These features are then fed into an encoder to extract difference information, which is subsequently mapped back to the original features in the decoder, as shown in the Fig. 5.

For the bitemporal features $X_i \in \mathbb{R}^{N\times C}$, we begin by converting pixel-level information into two sets of tokens $T_1, T_2 \in \mathbb{R}^{L\times C}$ that encompass high-level semantic concepts using spatial attention. Here, L represents the size of concept information within each token. The tokenization process can be expressed as follows:

$$T_i = (A_i)^T X_i = (\sigma(\phi(X_i; W)))^T X_i, i \in \{1, 2\} \tag{11}$$

$\phi(\cdot)$ denotes the pointwise convolution carried out using the learnable kernel $W \in \mathbb{R}^{C\times L}$. In addition, $\sigma(\cdot)$ serves to normalize the attention maps $A^i \in \mathbb{R}^{HW\times L}$ for each semantic group.

T_1, T_2 are passed into the transformer encoder for contextual modeling. The encoder is an N_E-layer iterative structure composed of normalization, MSA, and MLP. The MSA used here is the classic MSA operation in transformer. After the above-mentioned operations, new tokens $T_1^{\mathrm{new}}, T_2^{\mathrm{new}}$ will be obtained.

In the decoding operation, the optimized tokens T_i^{new} will be mapped back to the original features X_i through the decoder structure, thus updating new weights. In the decoder, we have improved the original MSA with a multihead cross-attention (MCA) mechanism. In this approach, query are obtained from X_i, while key and value are acquired from T_i^{new}. The advantage of this approach is that it helps avoid excessive computation resulting from the dense relationships between pixels. This can

be represented as follows:

$$\mathrm{MCA}\left(F_i^C, T_i^{\mathrm{new}}\right) = \mathrm{CAT}\left(h_1, h_2, \ldots, h_n\right) W^0 \tag{12}$$

where $W^0 \in \mathbb{R}^{hd \times C}$ is a linear projection matrix and n is the number of attention heads.

The transformer module concatenates two feature tokens to obtain a vector that contains information from both remote sensing images. This makes it easier to understand and explain the model's differential prediction results. The self-attention mechanism can calculate the correlation between different positions and features of input vectors, thus better understanding the relationship between different positions. This helps to improve the accuracy and robustness of the model. Moreover, this design only requires one transformer model to process the merged vector, which reduces the model's parameter count and improves its training and inference efficiency. This approach has been well demonstrated in the [21].

D. Other Network Details

Network structure: The reason for adopting a three-layer structure in the encoding part is because we considered that downsampling to a smaller size might not provide significant assistance in feature extraction during the early stages. Therefore, we downsized it to 64 × 64. In addition, we aimed to capture feature difference information from a relatively clear size. Thus, we avoided reducing the feature size too much in the final layer to prevent uncertainty in feature difference information extraction due to excessively small sizes, while too large sizes would incur substantial computational overhead. While it might yield better results to perform difference feature extraction at each layer, it comes with significant computational costs, given the exponential growth in the computational expense of self-attention.

Loss function: In the CD method, its essence is similar to binary classification tasks in semantic segmentation. Therefore, we use the minimization of cross-entropy loss to optimize the model parameters. The loss function is defined as follows:

$$L = \frac{1}{H_0 \times W_0} \sum_{h=1,w=1}^{H,W} l\left(P_{hw}, Y_{hw}\right) \tag{13}$$

where $l(P_{hw}, y) = -\log(P_{hwy})$ is the cross-entropy loss, P_{hw} and Y_{hw} is the label for the pixel at location (h, w).

IV. Experimental Results and Analysis

A. Experimental Setup

Evaluation metrics: In order to evaluate the effectiveness of our method, we use precision, recall, F1-score, intersection over union (IoU), and overall accuracy (OA) as the five evaluation metrics in our experiments. F1-score and IoU are the main evaluation metrics, and higher values indicate better model performance. The expressions for these metrics are as follows:

$$\begin{aligned}
\text{Precision} &= \frac{\text{TP}}{\text{TP+FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP+FN}} \\
F1-\text{score} &= 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \\
\text{Intersection over Union (IoU)} &= \frac{\text{TP}}{\text{TP+FN+FP}} \\
\text{Overall Accuracy (OA)} &= \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}}.
\end{aligned} \tag{14}$$

Implementation details: We conducted experiments on a Ubuntu system, running our DL model using the PyTorch framework with Torch version 1.11. During training, we utilized two Titan RTX GPUs, setting the batch size to 24. We employed the stochastic gradient descent optimization algorithm with a momentum of 0.99. The initial learning rate across all datasets was set to 0.01 and linearly decayed to 0 over 200 epochs. After each training stage, we performed validation and saved the best model.

B. Datasets

We evaluate all CD methods using three datasets.

LEVIR-CD [13] is a high-resolution (0.5 m/pixel) dataset consisting of 637 pairs of Google Earth images, each with a size of 1024 × 1024 pixels. It focuses on building-related changes, including various types of buildings. During the experiments, we cropped the images into nonoverlapping blocks of size 256 × 256, and set the training/validation/testing datasets to be 7120/1024/2048, respectively.

WHU-CD [48] is a large public CD dataset. It consists of a pair of high-resolution (0.075 m/pixel) images with a size of 32 507 × 15 354, and a spatial resolution of 0.075 m/pixel. We cropped the images into samples of size 256 × 256 for training/validation/testing, with dataset sizes of 6096/762/760, respectively.

GZ-CD [49] is a dataset consisting of 19 pairs of high-resolution (0.55 m/pixel) Google Earth images, including 19 pairs of seasonal change images covering the suburban areas of Guangzhou, China over the past decade. The focus is on changes related to buildings. The image sizes range from 1006 × 1168 to 4936 × 5224. We cropped them into nonoverlapping image blocks of size 256 × 256. Finally, we set the sizes of the training/validation/testing datasets to be 2834/400/325, respectively.

C. Comparative Experiment

In this section, we compare CGMMAanchor with several state-of-the-art methods, including three purely convolutional-based methods (FC-Siam-Conc [10], FC-Siam-Di [10], FC-EF [10]), four attention-based methods (DSIFN [11], DTCDSCN [19], SNUNet [40], AERNet [50]), and four transformer-based methods (BIT [21], ChangeFormer [22], MSCANet [6], AMTNet [51]).

1) FC-EF [10] combining UNet and Early Fusion results in FC-EF, where EF stands for early fusion. The EF structure involves concatenating two input images before

TABLE II
COMPARISON RESULTS ON THREE CD TEST SETS

	LEVIR-CD	**WHU-CD**	**GZ-CD**
Models	Pre. / Rec. / F1 / IoU / OA	Pre. / Rec. / F1 / IoU / OA	Pre. / Rec. / F1 / IoU / OA
FC-EF	86.91 / 80.17 / 83.40 / 71.53 / 98.39	84.60 / 71.74 / 77.64 / 63.45/ 98.01	77.49 / 65.74 / 71.13 / 55.19 / 94.89
FC-Siam-Diff	89.53 / 83.31 / 86.31 / 75.92 / 98.67	79.78 / 80.16 / 79.97 / 66.62 / 98.06	73.18 / 59.21 / 65.26 / 48.65 / 94.01
FC-Siam-Conc	91.99 / 76.77 / 83.69 / 71.96 / 98.49	83.07 / 85.19 / 84.12 / 72.59 / 98.45	80.59 / 72.34 / 76.24 / 61.61 /95.68
DTCDSCN	88.53 / 86.83 / 87.67 / 78.05 / 98.77	91.84 / 89.16 / 90.48/ 82.62/ 99.09	88.19 / 78.38 / 83.00 / 70.93 / 96.92
DSIFN	**94.02** / 82.93 / 88.13 / 78.77 / 98.87	**96.91** / 73.20 / 83.41 / 71.53 / 98.83	55.70 / 67.41 / 67.00 / 43.99 / 91.74
SNUNet	89.18 / 87.17 / 88.16 / 78.83 / 98.82	91.34 / 85.53 / 88.34 / 79.11 / 98.91	87.55 / 80.05 / 83.64 / 71.87 / 97.00
BIT	89.24 / 89.37 / 89.31 / 80.68 / 98.92	88.71 / 86.27 / 87.47 / 77.73 / 98.81	82.40 / 78.18 / 80.23 / 66.99 / 96.31
MSCANet	89.79 / 87.57 / 88.67 / 79.64 / 98.86	93.16 / 84.50 / 88.62 / 79.56 / 98.95	83.04 / 78.42 / 80.66 / 67.59 / 96.40
ChangeFormer	92.05 / 88.80 / 90.40 / 82.48 / 99.04	88.50 / 85.33 / 86.88 / 76.81 / 98.76	84.59 / 65.23 / 73.66 / 58.30 / 95.53
AMTNet	91.82 / 89.71 / 90.76 / 83.08 / 99.07	91.11 / 89.97 / 90.57 / 82.62 / 99.10	87.98 / 77.44 / 82.38 / 70.03 / 96.83
AERNet	89.97 / 91.59 / 90.78 / 83.11 / 99.07	92.47 / 91.89 / 92.18 / 85.49 / 99.25	88.06 / 81.07 / 84.42 / 73.03 / 97.13
CGMMA	92.27 / **91.34 / 91.81 / 84.85 / 99.17**	95.25 / **90.48 / 92.81 / 86.58 / 99.32**	**91.13 / 81.37 / 85.97 / 75.40 / 97.46**

The highest score is marked in bold. All scores are described as percentages (%).

TABLE III
PARAMETERS AND FLOPS RESULTS FOR ALL METHODS ON THE THREE DATASETS, AND SHOWS THE F1-SCORE AND IOU VALUES ON EACH DATASET

			LEVIR-CD		WHU-CD		GZ-CD	
	Params (M)	FLOPs (G)	F1	IoU	F1	IoU	F1	IoU
FC-EF	1.35	3.57	83.40	71.53	72.01	56.26	71.13	55.19
FC-Siam-Diff	1.35	4.72	86.31	75.92	77.59	63.38	65.26	48.65
FC-Siam-Conc	1.55	5.32	83.69	71.96	77.66	63.47	76.24	61.61
DTCDSCN	41.07	13.21	87.67	78.05	85.03	73.96	83.00	70.93
DSIFN	50.71	82.35	88.13	78.77	74.23	59.03	67.00	43.99
SNUNet	12.03	54.88	88.13	78.83	85.26	74.31	84.25	72.79
BIT	3.55	10.59	89.31	80.68	84.90	73.75	80.23	66.99
MSCANet	16.42	14.77	88.67	79.64	79.64	66.17	80.66	67.59
ChangeFormer	41.03	202.87	90.40	82.48	81.82	69.24	73.66	58.30
AMTNet	21.56	24.67	90.76	93.08	90.57	82.62	82.38	70.03
AERNet	25.36	12.82	90.78	83.11	92.18	85.49	84.42	73.03
CGMMA	17.19	41.78	91.81	84.85	92.81	86.58	85.97	75.40

feeding them into the network, i.e., different channels of one image.

2) FC-Siam-Di [10] is another structural form based on the Unet architecture and FC-EF. Di represents the difference image, where the feature extraction from the paired temporal images, followed by interpolation calculations, is employed to achieve differential image detection
3) FC-Siam-Conc [10] is also a structural form based on the Unet architecture and FC-EF. In this approach, paired temporal images are concatenated at each layer, and the concatenated images are passed backward to the decoder for differential image analysis.
4) DTCDSCN [19] is a twin CNN composed of three subnetworks, designed for dual-task constraints. It incorporates dual-attention modules and an improved focus loss to address the issue of sample imbalance. The network's output consists of a CD map and a segmentation result map.
5) IFNet [11] is a method that utilizes a multiscale feature fusion strategy. After utilizing multilayer feature results and undergoing feature extraction by VGG, features from each layer are passed to the decoder. Deep supervision loss is applied to achieve multiscale feature supervision.
6) SNUNet [40] employs a multilevel feature fusion technique, using the NestedUNet with dense connections for CD. In addition, it incorporates deep supervision to enhance the recognition capabilities of intermediate features and improve the effectiveness of the final features.
7) MSCANet [6] is a feature-level method based on the transformer architecture. It leverages both CNN and transformer, resulting in enhanced performance and improved CD capabilities.
8) BIT [21] is a lightweight approach that relies solely on a shallow ResNet, using transformer to extract difference features. The mentioned transformer approach proves to be highly effective.

TABLE IV
INTERNAL BACKBONE ABLATION EXPERIMENT

Module	CNN	CNN-PPM	Congraph	LEVIR-CD		WHU-CD		GZ-CD	
				F1	IoU	F1	IoU	F1	IoU
Backbone	✓	×	×	91.02	83.53	92.06	85.28	83.33	71.42
Backbone	×	✓	×	90.86	83.26	92.27	85.66	85.78	75.11
Backbone	×	×	✓	91.81	84.85	92.81	86.58	85.97	75.40

These include CA module, MMA module, transformer module.

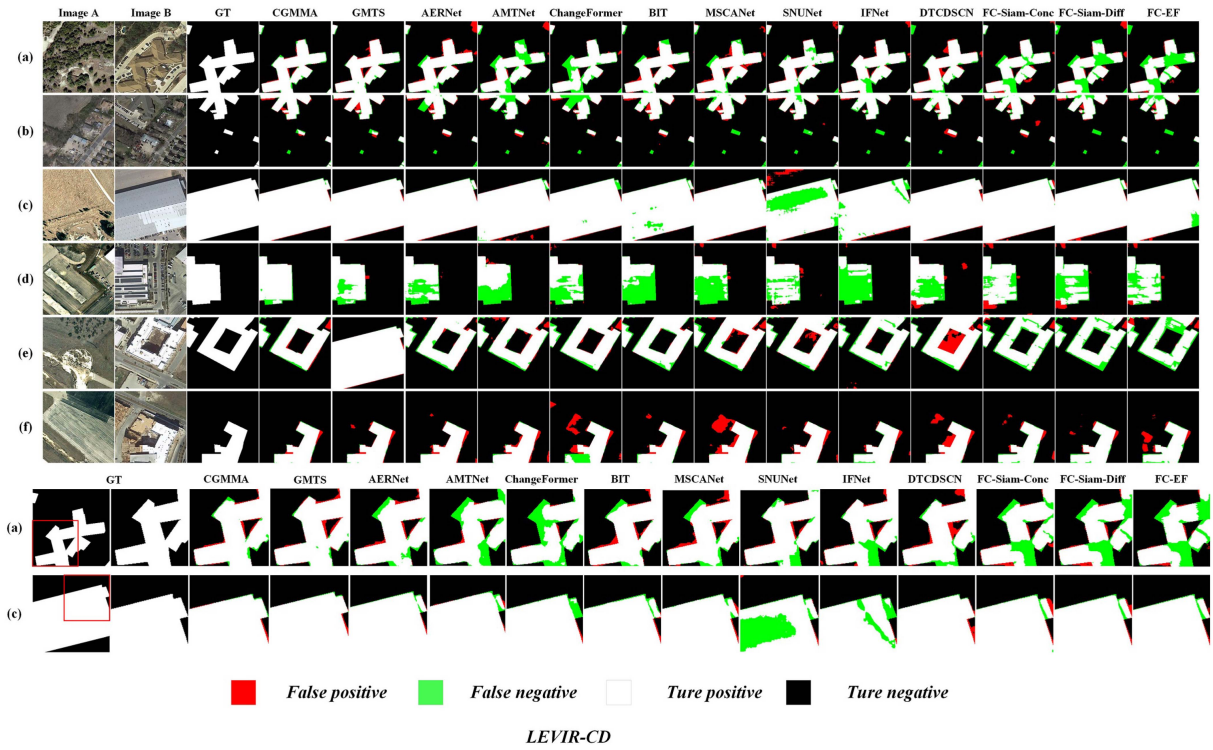

Fig. 6. Visualization results on the LEVIR-CD dataset. Different colors represent different results, with white representing true positive, black representing true negative, red representing false positive, and green representing false negative. The local detail maps of (a) and (c) are shown in the following.

9) ChangeFormer [22] is a pure transformer model that completely forgoes CNN. It incorporates multilevel information through concatenation into the difference information analysis module.
10) AMTNet [51] is a CNN-transformer-based architecture that utilizes ConvNets as the backbone to extract multiscale information. It effectively incorporates context information from paired temporal images using attention and transformer modules.
11) AERNet [50] introduces an attention-guided edge refinement network (AERNet) that utilizes a global context feature aggregation module to aggregate information from extracted multilevel context features. It enhances the network's perception and refinement of change regions by combining attention decoding blocks and edge refinement modules.
12) GMTS [52] introduces a feature extraction method that involves the interaction between GNN and CNN. This method enhances feature extraction capabilities by leveraging the interaction between global features from CNN and local features from GNN. In addition, it utilizes high and low-frequency attention along with pyramid transformers to focus on information at different scales.

We use the common code with default parameters to implement various advanced CD methods mentioned previously for comparative experiments, and use the same number of epochs during training.

Table II shows the evaluation results of all methods on the three CD test sets. The highest score is highlighted in bold. By comparing the results in the table, it is easy to see that our CGMMA method has significant advantages. Currently, methods based on transformer and attention mechanisms have better performance than traditional convolutional methods. However, our method achieves the highest results in most indicators. The DSIFN method performs well in the pre indicator, but not very well in the rec indicator. Usually, precision and recall have a mutually exclusive relationship, so we usually consider the more comprehensive F1-score measure. On the F1-score metric, we

TABLE V
BACKBONE STRUCTURE ABLATION EXPERIMENT

Module	Conformer	Congraph (TC)	Congraph	Params (M)	FLOPs (G)	LEVIR-CD		WHU-CD		GZ-CD	
						F1	IoU	F1	IoU	F1	IoU
Backbone	✓	×	×	16.17	48.39	91.02	83.52	92.58	86.19	84.39	72.99
Backbone	×	✓	×	16.42	122.75	91.62	84.53	92.52	86.06	85.84	75.19
Backbone	×	×	✓	17.19	41.78	91.81	84.85	92.81	86.58	85.97	75.40

TABLE VI
MMA MODULE AND TRANSFORMER MODULE ABLATION EXPERIMENTS

Model	MMA	TR	LEVIR-CD		WHU-CD		GZ-CD	
			F1	IoU	F1	IoU	F1	IoU
MMAT	×	✓	91.61	84.51	92.56	86.12	85.28	74.34
MMAT	✓	×	90.86	83.25	91.27	83.95	83.47	71.64
MMAT	✓	✓	91.81	84.85	92.81	86.58	85.97	75.40

have demonstrated higher values compared to current state-of-the-art methods, outperforming GMTS by 0.24% on the LEVIR-CD dataset and by 0.17% on the WHU-CD dataset. However, we also acknowledge limitations, particularly due to the low spatial resolution of the GZ-CD dataset, where buildings appear more blurred. Deeper networks may cause these fuzzy features to gradually disappear during the sampling process. Therefore, the generalization capability may require further improvement.

Compared to the current pure transformer backbone and widely used ResNet and Unet backbones, our adopted Congraph achieved better results. We increased the RF while maintaining the efficiency of CNN. We believe that the significant improvement in our performance is attributed to the proposed Congraph backbone and MMA module. Obtaining superior feature information is crucial for addressing the majority of segmentation tasks.

To demonstrate the differences in parameter and computation complexity among the various methods, we have conducted tests on all the compared methods and the results are shown in Table III. Our parameter and FLOPs counts are not the best among the methods, but compared to methods of similar levels, our improvement is significant. For instance, when comparing our approach with the advanced methods AMTNet and AERNet in 2023, we have a smaller parameter count but achieve better results.

To address the issue of pseudochanges, we believe that only essential bitemporal feature mixing can accurately determine the correct feature difference component. The feature extraction stage aims to obtain superior feature information, while the acquisition of feature difference information is the key to resolving pseudochange problems, requiring the extraction of change components from two distinct features. Comparable methods, such as BIT, MSCAnet, and GMTS, have employed similar approaches and achieved decent results. CGMMA builds upon these methods, obtaining superior results due to the backbone's design producing better feature extraction, making the extraction of differential features more effective.

To better compare the differences between our CGMMA and other methods, we visualized the results of each method on three datasets and used different colors to represent TP (white), TN (black), FP (red), and FN (green). From the figures, we can intuitively see that our method has a higher overlap with the ground truth, and the proportion of red and green areas is smaller. Moreover, our method has fewer noise points on all three datasets. For example, in Fig. 6(a) and 6(c), for the recognition of complex buildings, our method has better completeness, while other comparative methods lack completeness in object recognition and have unclear edges. Most of the other methods have the problem of incorrect recognition outside the target area, resulting in many red areas. In Fig. 7(b) and 7(d), our method can accurately identify all targets, while many other methods ignore the target area, which further demonstrates that our method performs better in extracting feature details. In Figs. 6(c) and 8(a), our method performs much better in the integrity of large-area targets than other methods, which also proves that our method has better performance in focusing on global features, and the use of ViG and MMA for global feature processing is more effective. We have included detailed images of the visualization results of each dataset below each figure.

From the visualizations, we should be able to roughly observe the changed areas to study surface changes. Therefore, excessive green and red areas indicate challenges in correctly identifying changed regions. The ultimate focus of the CD task is a high-resolution remote sensing image, potentially composed of tens of thousands of pixels. Consequently, low accuracy may be more pronounced in ultralarge-scale images, making the recognizability of the final results crucial. We can use these results to study the scale of urban expansion, analyze land use areas, and explore the development of cities.

D. Ablation Experiments

To demonstrate the effectiveness of the modulesenlrg used in our method, we conducted ablation experiments on three modules, including Congraph, MMA, and transformer.

1) Congraph Backbone: In the ablation experiments of the Congraph backbone, we conducted multiple experiments. Among them, we conducted two ablation experiments on the GNN branch of Congraph. First, we removed the GNN branch in Congraph and only used the DO-Conv block as the backbone. Second, we used a pyramid pooling module (PPM) to expand the CNN RF to verify that ViG can provide more

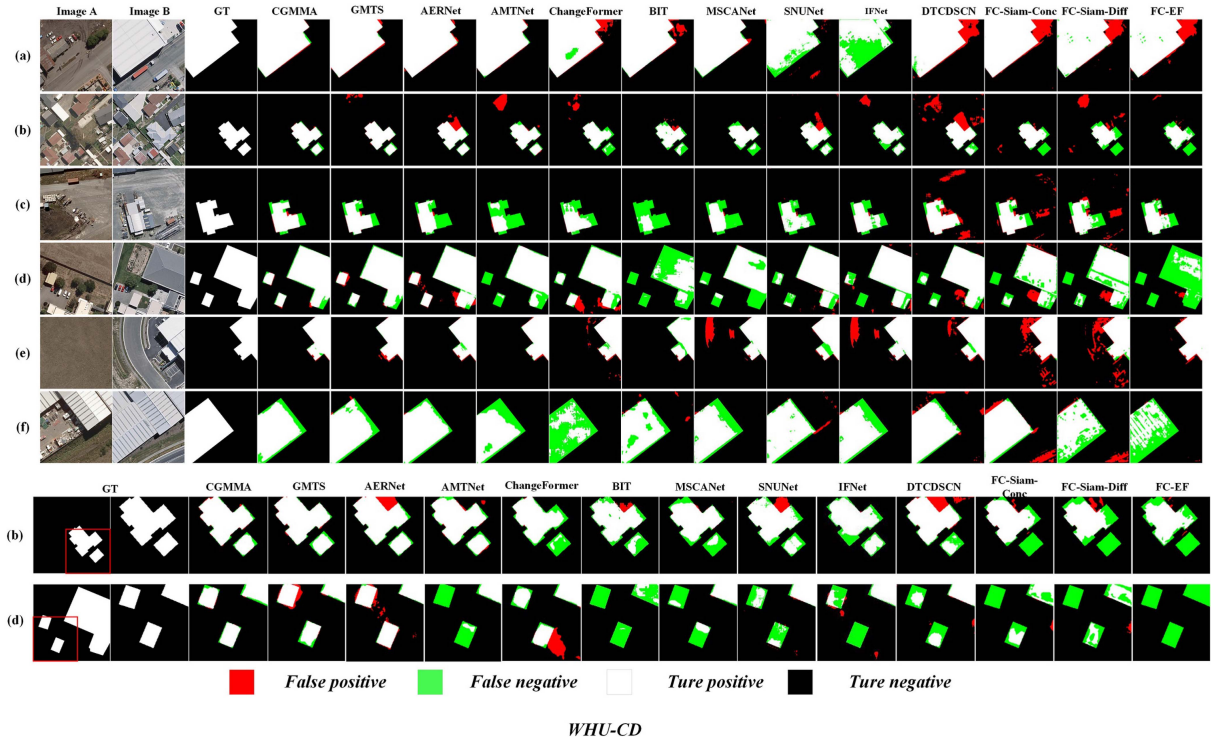

Fig. 7. Visualization results on the WHU-CD dataset. Different colors represent different results, where white represents true positives, black represents true negatives, red represents false positives, and green represents false negatives. The local detail images of (b) and (d) are shown below.

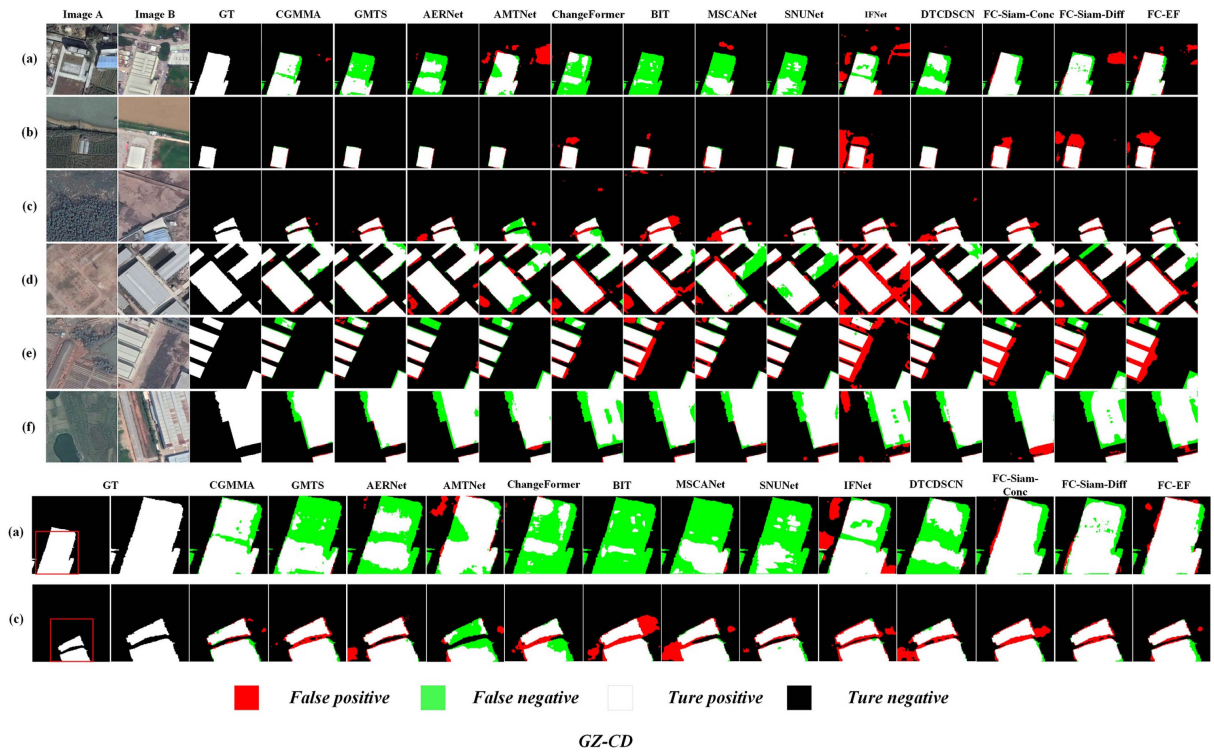

Fig. 8. Visualization results on the GZ-CD dataset. Different colors represent different results, where white represents true positives, black represents true negatives, red represents false positives, and green represents false negatives. The local detail images of (a) and (c) are shown in the following.

TABLE VII
α ABLATION EXPERIMENT IN MMA MODULE

Module	α	LEVIR-CD		WHU-CD		GZ-CD	
		F1	IoU	F1	IoU	F1	IoU
MMA	0.5	91.81	84.85	92.81	86.58	85.97	75.40
MMA	0.4	91.24	83.89	92.62	86.26	83.59	71.81
MMA	0.3	91.13	83.71	92.48	86.03	85.45	74.60
MMA	0.2	91.19	83.81	92.30	85.70	85.05	73.99

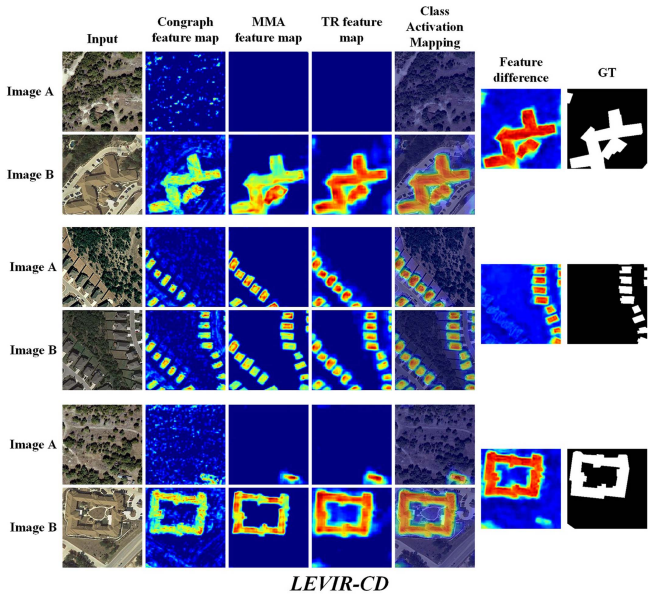

Fig. 9. Visualization results of CGMMA at various stages on LEVIR-CD dataset. These include Congraph feature maps, feature maps of the MMA module, feature maps after transformer, and feature difference maps.

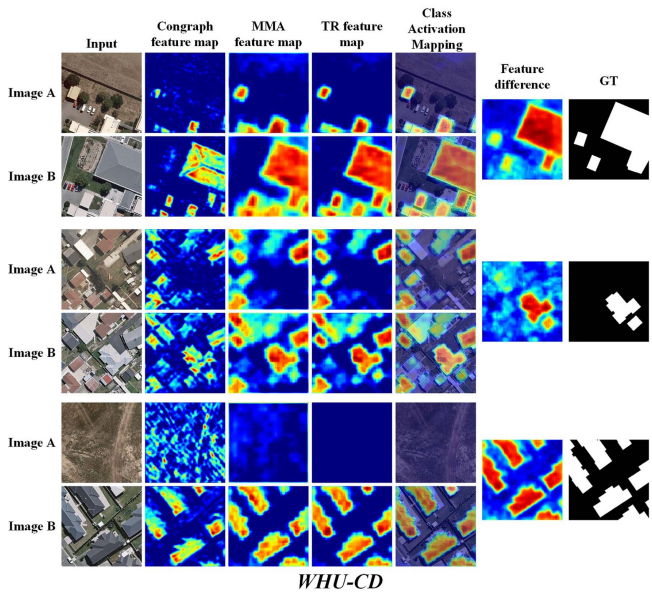

Fig. 10. Visualization results of CGMMA at various stages on WHU-CD dataset. These include Congraph feature maps, feature maps of the MMA module, feature maps after transformer, and feature difference maps.

comprehensive feature information for CNN. The results of the above-mentioned two experiments are shown in Table IV. The experimental results show that the CNN backbone has a good performance improvement after combining with ViG.

We also conducted two ablation experiments on the overall structure of Congraph. First, we compared the CNN-transformer interaction backbone Conformer to demonstrate that Congraph has better performance than Conformer. We used a Conformer with the same computation complexity as Congraph to maintain a fair comparison. Second, we conducted an ablation experiment using traditional convolution instead of DO-Conv. The results of the two experiments are shown in Table V. The experimental results show that we achieved better results than using the Conformer method, and DO-Conv has a significant decrease in computation complexity compared to traditional convolution.

2) Multiscale Mixed Attention: In the MMA module, we also conducted several ablation experiments. First, we removed the MMA module, as shown in Table VI. After removing the MMA module, the performance on the test set slightly decreased. After removing the MMA module, there was a slight decrease in performance on the test set. Second, we adjusted the parameter

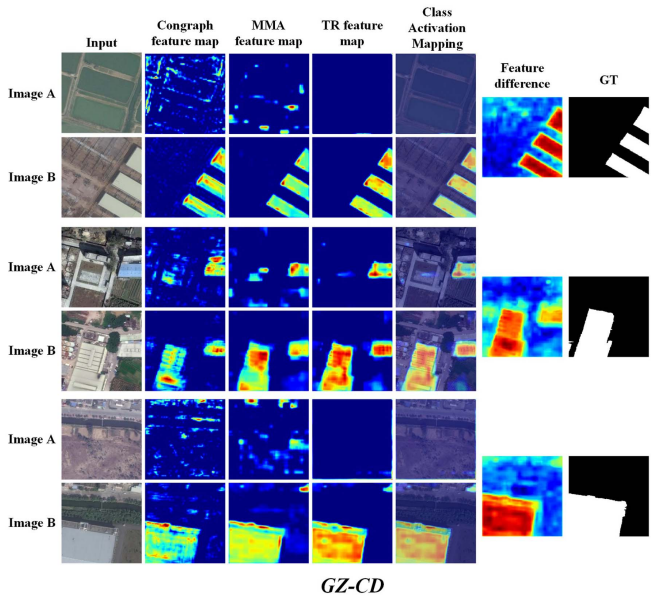

Fig. 11. Visualization results of CGMMA at various stages on GZ-CD dataset. These include Congraph feature maps, feature maps of the MMA module, feature maps after transformer, and feature difference maps.

TABLE VIII
SCALE PARAMETER ABLATION EXPERIMENT IN MMA MODULE

Module	scale size	LEVIR-CD		WHU-CD		GZ-CD	
		F1	IoU	F1	IoU	F1	IoU
MMA	64,32,16,8	91.81	84.85	92.81	86.58	85.97	75.40
MMA	64,16	91.60	84.50	92.37	85.82	85.26	74.30

α inside the MMA module to find its optimal value, as shown in Table VII. After changing the parameter, there was a significant decrease in performance on all datasets. In addition, we also compared the performance of four scales and dual scales, as shown in Table VIII.

3) Transformer Module: For the ablation experiments of the transformer module, we used a deletion approach, as shown in Table VI. After removing this module, the metrics on all three datasets showed a significant decrease.

E. Network Visualization

In order to offer a more comprehensive visual representation of the crucial phases within our CGMMA methodology, we have generated visualizations of the intermediate layers in our network. These visualizations, depicted in Fig. 9 (LEVIR-CD), Fig. 10 (WHU-CD), and Fig. 11 (GZ-CD), encompass a range of visual elements. This includes representations of Congraph features, MMA features, posttransformer features, alongside the incorporation of class activation mapping and feature difference maps.

V. CONCLUSION

In this article, we aim to address the limitations of CNN in handling complex spatial relationships within remote sensing images through the application of GNN. The proposed backbone network, Congraph, integrates convolution and graph interaction, allowing for the simultaneous utilization of local information from CNN and global information from GNN. This integration results in more comprehensive feature extraction, enhancing the accuracy of CD. In addition, the introduction of the MMA enables the model to focus on different scale feature information. MMA replaces small-scale features in the multilayer encoder with self-attention, capturing global feature information within small-scale features. In the final stage, we input bitemporal features into the transformer module to obtain feature difference information, ultimately generating the feature difference map. Extensive experiments conducted on the LEVIR-CD, WHU-CD, and GZ-CD datasets demonstrate that our method outperforms current state-of-the-art CD methods. These findings underscore the efficacy of our proposed Congraph backbone, MMA module, and the overall architecture in addressing the limitations of traditional CNN-based methods. The demonstrated improvements in accuracy and performance pave the way for more effective remote sensing CD, contributing to advancements in urban expansion studies, land use analysis, and the exploration of urban development dynamics.

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] S. Iino, R. Ito, K. Doi, T. Imaizumi, and S. Hikosaka, "Generating high-accuracy urban distribution map for short-term change monitoring based on convolutional neural network by utilizing SR imagery," in *Earth Resources Environmental Remote Sensing/GIS Applications VIII*. Bellingham, WA, USA: SPIE, 2017, pp. 11–21.

[3] B. Peng, Z. Meng, Q. Huang, and C. Wang, "Patch similarity convolutional neural network for urban flood extent mapping using bi-temporal satellite multispectral imagery," *Remote Sens.*, vol. 11, no. 21, 2019, Art. no. 2492.

[4] P. P. De Bem, O. A. de Carvalho Junior, 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.

[5] C. Mucher, K. Steinnocher, F. Kressler, and C. Heunks, "Land cover characterization and change detection for environmental monitoring of pan- Europe," *Int. J. Remote Sens.*, vol. 21, no. 6–7, pp. 1159–1181, 2000.

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

[7] 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.

[8] S. Patra, S. Ghosh, and A. Ghosh, "Histogram thresholding for unsupervised change detection of remote sensing images," *Int. J. Remote Sens.*, vol. 32, no. 21, pp. 6071–6089, 2011.

[9] F. Wang and Y. J. Xu, "Comparison of remote sensing change detection techniques for assessing hurricane damage to forests," *Environ. Monit. Assessment*, vol. 162, pp. 311–326, 2010.

[10] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE 5th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[11] 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.

[12] X. Song, Z. Hua, and J. Li, "LHDACT: Lightweight hybrid dual attention CNN and transformer network for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[13] 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.

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

[15] 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.

[16] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.

[17] 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.

[18] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[19] 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.

[20] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.
[21] 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.
[22] 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.
[23] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.
[24] X. Song, Z. Hua, and J. Li, "Remote sensing image change detection transformer network based on dual-feature mixed attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5920416.
[25] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "CNN-enhanced graph convolutional network with pixel-and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021.
[26] S. Liu, L. Duan, Z. Zhang, X. Cao, and T. S. Durrani, "Multimodal ground-based remote sensing cloud classification via learning heterogeneous deep features," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7790–7800, Nov. 2020.
[27] J. Liang, Y. Deng, and D. Zeng, "A deep neural network combined CNN and GCN for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4325–4338, 2020.
[28] F. Cui, Y. Shi, R. Feng, L. Wang, and T. Zeng, "A graph-based dual convolutional network for automatic road extraction from high resolution remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 3015–3018.
[29] C. Liang, B. Xiao, and B. Cheng, "GCN-based semantic segmentation method for mine information extraction in GAOFEN-1 imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, 2021, pp. 3432–3435.
[30] S. Liang, Z. Hua, and J. Li, "GCN-based multi-scale dual fusion for remote sensing building change detection," *Int. J. Remote Sens.*, vol. 44, no. 3, pp. 953–980, 2023.
[31] 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.
[32] H. Chen, J. Song, C. Wu, B. Du, and N. Yokoya, "Exchange means change: An unsupervised single-temporal change detection framework based on intra-and inter-image patch exchange," *ISPRS J. Photogrammetry Remote Sens.*, vol. 206, pp. 87–105, 2023.
[33] H. Chen, C. Lan, J. Song, C. Broni-Bediako, J. Xia, and N. Yokoya, "Land-cover change detection using paired openstreetmap data and optical high-resolution imagery via object-guided transformer," 2023, *arXiv:2310.02674*.
[34] M. Gong, Y. Yang, T. Zhan, X. Niu, and S. Li, "A generative discriminatory classified network for change detection in multispectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 321–333, Jan. 2019.
[35] Y. Zhou and X. Li, "Unsupervised self-training algorithm based on deep learning for optical aerial images change detection," 2020, *arXiv:2010.07469*.
[36] H. Cheng, H. Wu, J. Zheng, K. Qi, and W. Liu, "A hierarchical self-attention augmented Laplacian pyramid expanding network for change detection in high-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 182, pp. 52–66, 2021.
[37] 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.
[38] K. Song, F. Cui, and J. Jiang, "An efficient lightweight neural network for remote sensing image change detection," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 5152.
[39] X. Song, Z. Hua, and J. Li, "PSTNet: Progressive sampling transformer network for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8442–8455, 2022.
[40] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8007805.
[41] S. Li and L. Huo, "Remote sensing image change detection based on fully convolutional network with pyramid attention," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, 2021, pp. 4352–4355.
[42] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[43] J. Cao et al., "DO-Conv: Depthwise over-parameterized convolutional layer," *IEEE Trans. Image Process.*, vol. 31, pp. 3726–3736, 2022.
[44] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.
[45] G. Li, M. Muller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs?," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9266–9275.
[46] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3538–3545.
[47] K. Oono and T. Suzuki, "Graph neural networks exponentially lose expressive power for node classification," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–37.
[48] 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.
[49] 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.
[50] J. Zhang et al., "AERNet: An attention-guided edge refinement network and a dataset for remote sensing building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617116.
[51] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 202, pp. 599–609, 2023.
[52] X. Song, Z. Hua, and J. Li, "GMTS: GNN-based multi-scale transformer siamese network for remote sensing building change detection," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 1685–1706, 2023.

Yan Zhang received the bachelor's degree in communication engineering in 2022 from Shandong Technology and Business University, Yantai, China, where she is currently working toward the master's degree in electronic information.

Her research interests include computer graphics, computer vision, and image processing.

Xinyang Song received the bachelor's degree in communication engineering from Qingdao Agricultural University, Qingdao, China, in 2020. He is currently working toward the master's degree in electronic information with Shandong Technology and Business University, Yantai, China.

His research interests include computer graphics, computer vision, and image progressing.

Zhen Hua received the B.S. and M.S. degrees in electrical automation from the Taiyuan University of Technology, Taiyuan, China, in 1989 and 1992, respectively, and the Ph.D. degree in electronic information engineering from the China University of Mining and Technology, Beijing, China, in 2008.

She is currently a Professor with Shandong Technology and Business University, Yantai, China. Her research interests include computer aided geometric design, information visualization, virtual reality, and image processing.

Jinjiang Li received the B.S. and M.S. degrees from the Taiyuan University of Technology, Taiyuan, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shandong University, Jinan, China, in 2010, all in computer science.

He is currently a Professor with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China. From 2004 to 2006, he was an Assistant Research Fellow with the Institute of Computer Science and Technology, Peking University, Beijing, China. From 2012 to 2014, he was a Postdoctoral Fellow with Tsinghua University, Beijing, China. His research interests include image processing, computer graphics, computer vision, and machine learning.