MVAFG: Multiview Fusion and Advanced Feature Guidance Change Detection Network for Remote Sensing Images

Xiaoyang Zhang, Zhuhai Wang, Jinjiang Li, and Zhen Hua

***Abstract*—In recent years, change detection (CD) methods have faced challenges in being applied to various types of remote sensing datasets and related research fields, particularly in the domain of CD in remote sensing images. While convolutional neural networks (CNNs) have significantly advanced CD in remote sensing images, they struggle with modeling long-distance dependencies between image pairs, leading to poor recognition of semantically similar objects with different features. Meanwhile, transformer technology has gained widespread popularity for global applications, but it lacks in extracting local features effectively. Current approaches typically rely on single or dual-branch network structures for mining change-related features in remote sensing images, yet they still lack in extracting both local and global features comprehensively. To address these issues, this article proposes a triple-branch network combining transformer and CNN, comprising CNN, transformer, and channel feature-guided branch. These branches extract and fuse three types of change features from both global and local perspectives. Importantly, the channel feature-guided branch is introduced to capture continuous and detailed change relationship features, thus enhancing the model's change discrimination ability. Experimental results on three datasets (LEVIR-CD, WHU-CD, and GZ-CD) demonstrate the superior performance of the model over state-of-the-art methods.**

***Index Terms*—Change detection (CD), change relation features, channel feature-guided (CGF), convolutional neural networks (CNNs), remote sensing, transformer.**

I. Introduction

CHANGE detection (CD) in dual-temporal remote sensing imagery is a crucial component of terrestrial change monitoring and is extensively applied in fields such as urban planning [1], forest cover mapping [2], and disaster damage assessment [3], [4]. Although the definition of CD varies by application, its core objective is to mark binary changes on the surface from registered images taken at two different times [5], [6], [7], [8]. With the rapid development of optical sensor technology, automated CD techniques have gained increasing attention. Automation can significantly reduce labor costs and time consumption. While high-resolution optical images offer new opportunities for remote sensing applications, conducting CD on a fine scale remains challenging. There are three main issues: first, imaging conditions differences due to varying shooting times—such as lighting, seasonal changes, and appearance variations—can cause the same semantic objects to appear different in color and shape across different image pairs, thereby complicating CD. Second, the scale of the target area varies widely, necessitating deeper consideration of smooth contours to reduce false positives and false negatives. Third, irrelevant changes interfere with the accurate representation of true change features, causing areas with the same semantic concepts to differ in spatial and temporal characteristics, which increases the complexity of CD.

To enhance the accuracy of change detection, numerous excellent CD algorithms have been proposed over the past few decades [9], [10], [11]. In the early stages, due to limitations in computer hardware, such as GPUs, CD tasks relied on traditional methods, including random forests [12], decision trees [13], and support vector machines [14]. These conventional approaches focused on detecting and classifying changed pixels in bitemporal images to produce the final change map. While these methods have achieved notable success in some CD tasks, they often lack advantages in terms of change target recognition and accuracy due to the limitations of algorithmic thresholds and the influence of complex targets and environmental noise present in the images. However, it is important to acknowledge that traditional methods in the CD field have made significant contributions to the subsequent advancements in deep learning technologies [14].

With the advancements in computer hardware performance and artificial intelligence technologies, deep learning has been extensively applied in the CD field, yielding numerous outstanding algorithms [15], [16], [17], [18], [19]. Deep learning, by learning from sample spaces, is capable of identifying change targets. Among various deep learning approaches, those based on convolutional neural networks (CNNs) have gained widespread popularity. CNN-based algorithms excel over traditional methods in handling environmental factors such as lighting and shadows. Early deep learning strategies were primarily derived from semantic segmentation and image segmentation

Manuscript received 16 March 2024; revised 20 April 2024 and 11 May 2024; accepted 28 May 2024. Date of publication 31 May 2024; date of current version 14 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62002200, Grant 61972235, Grant 62272281, and Grant 62202268, in part by Shandong Natural Science Foundation of China under Grant ZR2020QF012 and Grant ZR2021MF068, and in part by Yantai Science and Technology Innovation development plan under Grant2022JCYJ031 and Grant 2023JCYJ040. *(Corresponding author: Zhuhai Wang.)*

The authors are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: 201513255@sdtbu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3407972

algorithms, such as VGG [20], FCN [21], and U-Net [22]. For instance, the FC-EF [23], based on the U-Net architecture concatenates bitemporal images and feeds them into an encoder to extract semantic information at different depths. Through the decoder and skip connections, the changed targets are reconstructed. While this method achieves low resource occupancy, the fused bitemporal image feature extraction previously disrupted the structural information within a single image, leading to suboptimal performance in edge handling by FC-EF [23]. Subsequently, two "late fusion" approaches based on U-Net [23], FC-Siam-conc and FC-Siam-diff, were proposed. Both methods replicate U-Net-s encoder into two identical ones with shared parameters, processing the bitemporal images separately for feature extraction. Their difference lies in the decoding phase: FC-Siam-conc concatenates the skip connections from both encoders, while FC-Siam-diff connects their absolute differences. Zhang et al. [5] followed the same architecture as FC-Siam-conc and employed a differential discrimination operation in the decoder to achieve uniform detection maps. These methods focus on exploring various connection and fusion strategies to fully exploit the spatial neighborhood context, achieving certain successes. However, they treat changed and unchanged information equally in context without discrimination, which limits the performance of CD [24]. A novel context aggregation network (CANet) [25] introduced a comprehensive approach to aggregate intraimage and interimage context to enhance the representativeness of specific category objects, thus enhancing the model's ability to detect changes of various scales. Overall, CNN-based CD methods can achieve satisfactory detection performance, especially on prominently changed or unchanged objects. However, when bitemporal images contain ambiguous areas—where pseudochanges occur or real changes are obscured or damaged—most existing CNN-based CD models encounter difficulties [26]. To effectively address ambiguous areas, Hang et al. [26] proposed a remote sensing image change detection network (AANet) focused on exploring ambiguous areas, effectively resolving pseudochanges caused by seasonal variations, lighting changes, and scenarios where real changes are obscured or damaged.

The self-attention mechanism introduced in the transformer [27] model has been proven effective in capturing long-range dependencies between feature information, leading to its widespread use in the remote sensing field due to its capability to acquire global feature information [28], [29], [30]. Although the transformer [27] model can capture global concepts, its outputs exhibit a uniformly consistent global representation across different stages, resulting in redundancy between shallow and deep layers. To address the challenge of encoding both local features and global concepts simultaneously, an intuitive approach is to integrate CNNs and the transformer into a single network. For example, the BIT [28] model utilizes the transformer encoder to establish global feature relations for fused high-dimensional information, thereby enhancing the intensity of semantic information. However, BIT's use of ResNet18 for different scale feature information has limitations [28], and it lacks finesse in the differential marking process during image restoration and decoding phases. The subsequently developed Changeformer [29] employs a multihead self-attention module as the backbone network for feature information extraction, but it also significantly increases resource utilization. ICIF-Net [30] adopts a parallel approach, using both CNN and transformer [27] backbone networks simultaneously to extract feature information, and creates a dual-branch, multistage cross-temporal network to fuse the features obtained from the two different backbone networks.

Although these deep networks have achieved good results in CD, they also have some shortcomings. Single-branch structures that concatenate two input images and then feed them into the network often lack detailed and locational information about individual images [21]. Two-branch structures typically employ conjoined encoders to extract features from each image separately and fuse these features at the decoder stage. The architectural contradiction between the dual-branch encoders and a single decoder leads to vanishing gradient propagation and affects the learning of low-level features from the two original images [22]. Moreover, these networks do not explicitly mine the features of change relationships, leading to a blurred understanding of the changes between the two images. To address the problem of undercharacterizing mining change relationships, we propose a network structure with three branches, each dedicated to mining three types of change features: the features of the prechange remote sensing image, the postchange remote sensing image, and the change relationship features between the prechange and postchange images. Starting from the perspective of extracting both global and local information from the images, we combine transformers and CNNs to further extract and fuse these three types of features. In the upsampling stage of the decoder, we address the deficiencies in detail features extracted by CNNs and transformers by introducing a channel feature guidance module, which enhances the model's precise detection performance. The triple-branch structure extracts local multiscale and global multiscale features in parallel, achieving the final binary map of change through feature guidance and fusion.

In summary, the contributions of this work can be summarized as follows.

1) We first introduce a triple-branch network MVAFG, which, through its triple-branch structure, can more deeply mine the features of images before and after change, as well as the change relationship features between them. This capability facilitates better performance in CD. The triple-branch network extracts, interacts, and fuses features from bitemporal images by combining transformers and CNNs.
2) We propose a channel feature-guided (CGF) branch to construct the third branch of the network, which is a feature module based on an encoder–decoder architecture. Through the channel feature-guided encoder (CGFE), this branch deeply mines feature information. After operations such as channel swapping, and spatial and channel feature fusion, the channel feature-guided decoder (CGFD) generates target outputs more accurately. This approach enables the learning of continuous and detailed features of changes

in remote sensing images, further enriching the model's information on change relationships.

3) We propose an adaptive feature refinement module (AFRM) with reweighting capabilities to further enrich and enhance the features of bitemporal images extracted by the transformer across different dimensions, significantly improving feature extraction. Additionally, we introduce a global context integration module (GCIM) to substantially enhance the CNNs feature extraction capabilities. We also employ a deep convolutional attention (DCA) module to interact with the bitemporal features extracted by the dual branches of CNN and transformer, capturing feature information under different branches and better integrating bitemporal feature information.
4) We conducted a series of experiments to validate the effectiveness and superiority of the proposed methods. The experimental results demonstrate that the MVAFG network achieved better or more competitive detection performance compared to nine of the most advanced models on three benchmark datasets.

The rest of this article is organized as follows. Section II discusses related work associated with the CD framework. Section III provides a detailed description of the methods we propose. To validate the effectiveness of the approach, we conducted several ablation experiments and comparisons with proposed methods in Section IV. Finally, Section V concludes this article.

II. Related Work

With the continuous development of deep learning technologies, many new methods based on CNNs, attention mechanisms, and transformers have emerged in the field of CD tasks [31]. Currently, significant efforts have been made by these methods to explore how to better execute CD tasks, resulting in numerous innovative approaches based on various technologies [32]. Given the tremendous success of deep learning-based methods in the CD domain, we will next briefly review the three most relevant branches, which form the research foundation of the work. These clearly categorized approaches will help to deepen the understanding of different aspects of bitemporal remote sensing image CD models, thereby further advancing the field of remote sensing CD.

A. CNN-Based Module

Remote sensing CD involves using remote sensing imagery to compare differences and similarities at two different time points, identifying changes on the Earth's surface. In recent years, methods based on CNNs have become one of the mainstream approaches in the field of remote sensing CD [5], [33], [34], [35], [36], [37]. First, CNNs, as powerful feature extractors, are capable of learning high-level semantic features from remote sensing images. Many existing studies focus on using CNNs to extract feature representations from remote sensing images. For example, some studies utilize pretrained CNN models (such as VGGNet, ResNet) as feature extractors, acquiring high-level semantic features of images through transfer learning. These features are able to capture target shapes, textures, and structures in remote sensing images. Second, to address the scale variations and complexities in remote sensing images, many studies have introduced multiscale feature fusion strategies [33], [34]. These strategies enable feature extraction at different scales and fuse multiscale features to obtain more comprehensive and rich feature representations. For instance, some research employs pyramid-structured CNNs to extract features at various scales, and these features are fused through convolutional or pooling operations [38]. This approach retains both low-level detail features and high-level semantic information, enhancing CD performance. Additionally, to incorporate contextual information and enhance the capability of feature representation, some studies have also adopted attention mechanisms. Channel attention mechanisms [34] can recalibrate the channel responses in feature maps, emphasizing features relevant to CD. Spatial attention mechanisms [36] can highlight important regions in specific locations of the image. By integrating attention mechanisms, CNNs can focus more on learning information relevant to CD tasks, thereby improving CD performance.

Overall, remote sensing CD methods based on CNNs effectively enhance the performance of CD in remote sensing images by leveraging high-level semantic features extracted by CNNs, multiscale feature fusion, and the introduction of attention mechanisms. The development of these methods provides a feasible and effective solution for remote sensing image CD. However, it is important to note that while multiscale feature maps with local features extracted by CNNs are crucial for CD in bitemporal remote sensing image pairs, CNNs lack the capability to model long-range dependencies and global context interactions, which are essential for accurately reflecting change relationship features. The three-branch network structure we propose addresses the shortcomings of CNNs in capturing global contextual information, allowing for a deeper exploration of change relationship features and resolving the lack of long-range dependence modeling in CNNs.

B. Transformer-Based Model

Since their inception, transformers have been widely applied in the field of natural language processing to model long-range dependencies with ease. With their superior modeling capabilities, transformers have demonstrated considerable, and sometimes superior, performance, subsequently causing a significant stir in the field of computer vision. More recently, transformer-based methods have also gained attention in the field of remote sensing CD and have achieved some important research outcomes [39], [40]. First, as a powerful method for sequence modeling, transformers can handle spatiotemporal sequence data in remote sensing images and capture long-term dependencies and global context interactions within images. Compared to traditional CNN-based methods, transformers are better equipped to model nonlocal and long-range dependencies. Second, some studies have transformed remote sensing images into a series of spatiotemporal sequence data and fed them into a transformer model. Through its self-attention mechanism, the transformer can interact with different positions within the image and learn the feature representations of each position.

This mode of sequence modeling effectively captures the spatial and temporal correlations in remote sensing images, thereby enhancing the performance of remote sensing CD. Additionally, to further enhance the CD capabilities of the transformer model, some research works have also incorporated attention mechanisms. Attention mechanisms can highlight important areas within the image and enhance the feature representation capabilities of these areas. By combining attention mechanisms with the transformer model, researchers are able to more specifically learn features relevant to CD tasks.

In summary, transformer-based methods for remote sensing CD leverage the modeling of spatiotemporal sequence data, the incorporation of attention mechanisms, and the integration of techniques such as CNNs to enhance the performance of CD in remote sensing images. The development of these methods provides a novel and effective solution for remote sensing image CD. However, transformers lack the capability to extract local feature information. The triple-branch network combines the local feature extraction capabilities of CNNs with the global feature extraction capabilities of transformers, enabling a more effective extraction of both global and local feature information.

C. Model Combining Transformer With CNN

Recent studies have shown that combining transformers with CNNs has made significant progress in remote sensing CD [28], [30]. This approach effectively utilizes the spatial feature extraction capabilities of CNNs and the global feature modeling capacity of transformers to further enhance the performance of remote sensing CD. First, employing CNNs as a feature extraction branch captures low-level and local features within the image. Typically, CNNs extract feature representations of the image through multiple convolutional and pooling operations, which are then passed to the transformer branch. Second, the transformer, serving as another branch, learns the high-level semantics and global features of the image. It models long-term dependencies within the image through its self-attention mechanism and captures interactions among features through multiple layers of self-attention blocks. This enables the model to understand the image holistically and better model the CD task. In the process of combining CNNs with transformers, a common approach is to use feature fusion. Feature fusion can be achieved through concatenation, stacking, or averaging, merging the low-level features extracted by CNNs with the high-level features extracted by transformers, resulting in a richer, more robust feature representation. Such fusion effectively leverages the respective strengths of CNNs and transformers, enhancing the performance of remote sensing CD. Additionally, some studies have introduced attention mechanisms to enhance the fused feature representation. By incorporating channel attention or spatial attention mechanisms, the model can adaptively adjust the feature weights of different channels or locations, further enhancing the expression of key features. In summary, remote sensing CD methods based on the combination of transformers and CNNs capitalize on the spatial feature extraction advantages of CNNs while utilizing the global feature modeling and long-term dependency modeling capabilities of transformers. This approach enhances the accuracy and robustness of remote sensing CD, holding significant applicative value for complex remote sensing image CD tasks.

Although the methods combining transformers with CNNs have made significant progress in the field of remote sensing CD, there remain issues such as insufficient capture of local features and a lack of advanced semantic features that affect the accuracy of CD. Although transformers can learn high-level semantic features and global context information of images, in some cases, relying solely on local information is insufficient to capture the comprehensive features of changes in remote sensing images. Transformer models may not fully utilize the low-level features extracted by CNNs to gain more advanced semantic information, leading to inadequate feature representation. The proposed triple-branch network model addresses these issues through a channel feature guidance branch that better compensates for insufficient feature representation. By starting from both global and local feature information, it delves deeper into exploring change relationship features, effectively resolving the issues of insufficient local feature capture and missing advanced semantic features found in methods that combine transformers with CNNs.

III. Methodology

A. Network Architecture

In this section, we will provide a detailed introduction to the MVAFG model, as shown in Fig. 1. The CD framework mainly consists of three important branches: CNN, transformer, and the CGF branch. These three branches are processed in parallel to effectively obtain both local and global features of the input images. Specifically, in the CNN branch, we adopt the widely used ResNet18 [41], while in the transformer branch, we employ the state-of-the-art PVTv2-B1 [42] to obtain multiscale feature maps. Additionally, to enhance the features extracted by both CNN and transformer, we introduce an AFRM with reweighting functionality to improve the transformer's ability to extract features of different dimensions. We also incorporate a GCIM to enhance feature extraction in the CNN branch significantly. Furthermore, by employing the DCA, we deeply fuse the features extracted by CNN and transformer, allowing for comprehensive complementarity of features from the dual-temporal images. To provide a more intuitive understanding of the approach, we detail the reasoning process in Algorithm 1.

To highlight the features extracted by both the CNN and transformer branches, a personalized filtering module (Filter) was added, as illustrated in Fig. 2(a). The features obtained from low dimensions are applied to those obtained from high dimensions to obtain f_1, focusing on collecting features such as edges and spatial structures. However, the learned features may be sensitive to background noise. As a supplement, the features obtained from high dimensions introduce rich semantic information into low-dimensional features to obtain f_2, aiding in localization and noise suppression. Therefore, the output combines multiple valuable features in the multiscale feature

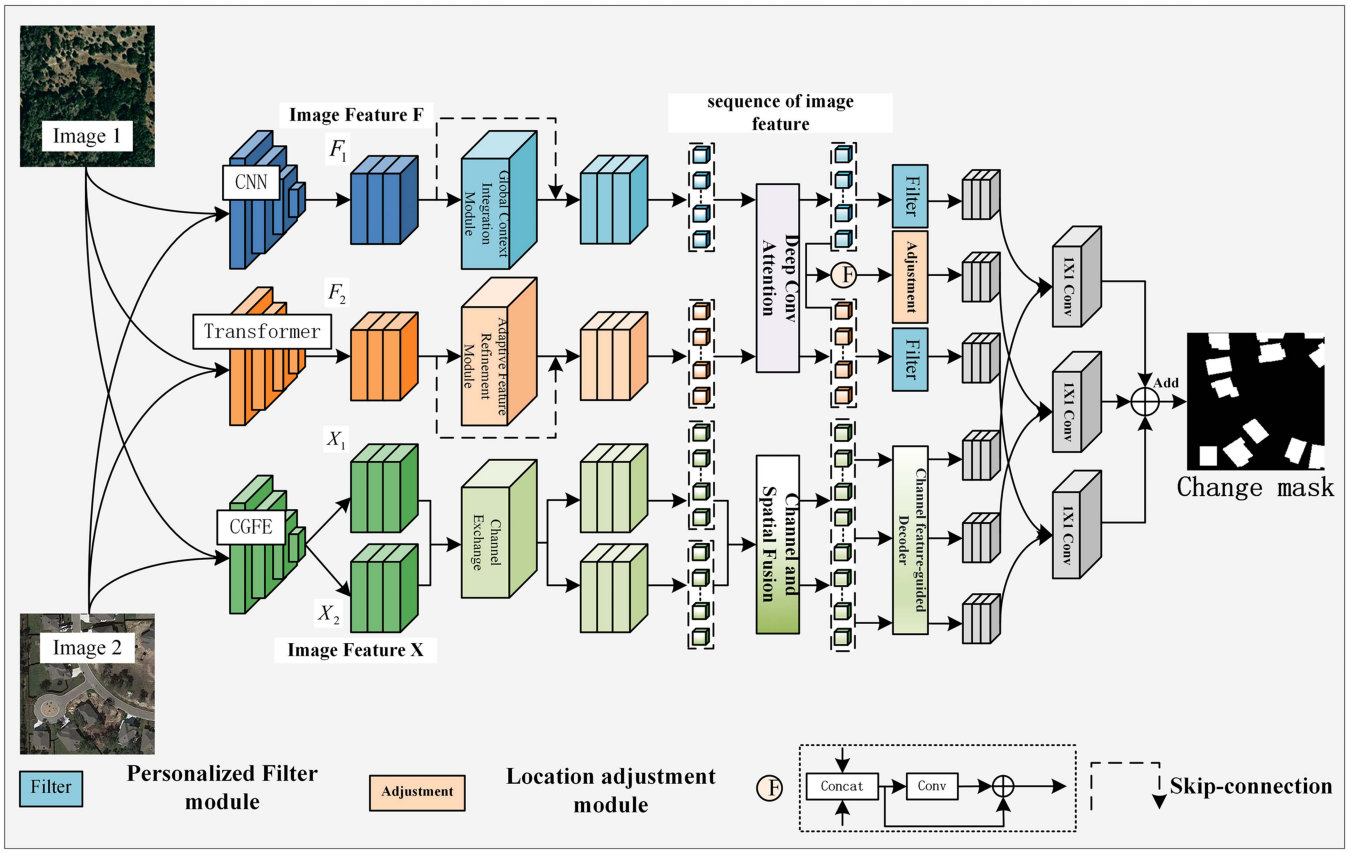

Fig. 1. Proposed MVAFG architecture consists of three branches composed of CNN, transformer, and the CFG branch. These three branches extract richer multidimensional features for fusion to generate the final output image.

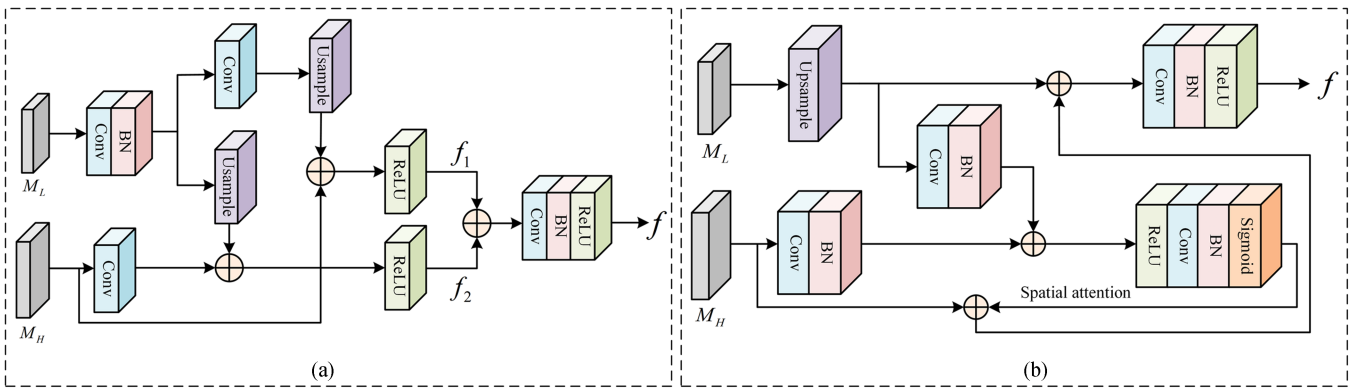

Fig. 2. Personalized filtering module (a) integrates high- and low-dimensional features to combine multiple valuable features in multiscale features. The location adjustment module (b) combines local features with long-range dependent features to share information between local and global features.

f. The relevant formulas are represented as follows:

$$f_1 = \text{Apply}_1(L, H) \tag{1}$$

$$f_2 = \text{Apply}_2(H, L) \tag{2}$$

$$f = \text{Combine}(f_1, f_2) \tag{3}$$

where L represents low-dimensional features, H represents high-dimensional features, Apply_1 denotes the operation of applying low-dimensional features L to high-dimensional features H, and Apply_2 represents the operation of applying high-dimensional features H to low-dimensional features L. Combine is the operation of merging f_1 and f_2 to obtain the final output feature f. Thus, the output feature f combines multiple valuable features, including the boundary and spatial structure features of low-dimensional features as well as the rich semantic information of high-dimensional features, aiding in localization and noise suppression.

Algorithm 1: MVAFG: Multiview Fusion and Advanced Feature Guidance Change Detection Network for Remote Sensing Images.

Input: $Image_1 = x_i, i \in [1, n]$;
$Image_2 = x_i, i \in [1, n]$;
$GT = G_i, i \in [1, n]$;
Output: Final change mask: Y;
for $i = 1, 2, 3$ **do**
$F_i = CNN(I_1, I_2)$
$F_i = transformer(I_1, I_2)$
$F_i = CGFE(I_1, I_2)$
end for
return $Filter(F_1, F_2)$; $Adjustment(F_3)$; $CGFE(F_1, F_2, F_3)$;
$Final_1 = Conv_{(1\times1)}(Filter(F_1), CGFD(F_1))$
$Final_2 = Conv_{(1\times1)}(Filter(F_2), CGFD(F_2))$
$Final_3 = Conv_{(1\times1)}(Adjustment(F_3), CGFD(F_3))$
$Output = Add(Final_1 + Final_2 + Final_3)$
return fusion result change mask: Y

For the part where features extracted from both CNN and transformer are fused, we introduce a position adjustment module based on attention mechanism (Adjustment), as shown in Fig. 2(b). By integrating high-dimensional features and low-dimensional features to model the overall spatial attention distribution associated with both local and global features, we aim to obtain more accurate fusion features

$$f = \text{Concat}(\text{Spaital}(\text{RELU}(L, H)), L). \tag{4}$$

Here, L represents low-dimensional features, H represents high-dimensional features, Spaital stands for spatial attention mechanism, and Concat denotes concatenation operation along the feature dimension to obtain the final fusion operation f.

More importantly, to complement the feature extraction from both the CNN and transformer branches, we propose a third branch called the CGF branch to further enhance the feature extraction capability of our network architecture. This module transfers the features extracted by the channel feature guidance encoder to the deeper layers of the encoder, enabling rough identification of change regions in the encoder by leveraging the semantic features of the dual-temporal information. This provides guidance for the dynamic localization of CD features in subsequent steps. The following illustrates the model training process:

$$\begin{aligned} Y = \text{Add}(&\text{CNN}(I_1, I_2) + \text{transformer}(I_1, I_2) \\ &+ \text{CGFE}(I_1, I_2)) \end{aligned} \tag{5}$$

where CNN, transformer, and CGFE represent the three different branches of the model graph, Y represents the training result image of the model, and I_1, I_2 denote the input dual-temporal images.

B. Channel Feature-Guided Branch

The CGF branch comprises the CGFE, channel exchange (CE), spatial and channel fusion module (SCF), and the CGFD. The CGFE includes multiple semiconvolutional, semipooling modules. Through these modules, input features are split into two halves along the channel dimension; one half is passed through a convolution layer for enhancement, while the other undergoes max pooling to preserve essential feature information. The features from the two branches are then concatenated and shuffled in an alternating pattern along the channel dimension to produce the output features. The channel shuffling operation ensures sufficient cross-channel interaction and the preservation of critical features, which benefits gradient backpropagation and feature reuse, thereby facilitating effective feature extraction. Details of the channel feature guidance encoder implementation are illustrated in Fig. 3. As the CGFD extracts bitemporal image features, performing channel exchange (CE) on these features enriches the bitemporal information. This means that the bitemporal features are joined, ensuring that features, which previously contained only single-time information, now include rich spatial characteristics, which can be utilized to refine change areas and precisely locate change objects at the same temporal phase. Semiexchange of the extracted bitemporal features in the CGFE involves alternatingly swapping half of the input bitemporal features along the channel dimension through the channel exchange (CE) module, thus accentuating the change feature information of the b-temporal images. The (CE can be formalized as follows:

$$T_1, T_2 = M * t_1 + (1 - M) * t_2 \tag{6}$$

where t_1 and t_2 represent the dual-temporal features, T_1 and T_2 represent the exchanged dual-temporal features, and M represents a 1-D exchange mask with a length equal to the channel dimension of the dual-temporal features, alternating between 0 and 1. Thus, each exchanged feature contains half of the dual-temporal features, implying that each exchanged feature encompasses the dual-temporal semantic features of the dual-time remote sensing images.

The spatial and channel fusion module effectively merges the dual-temporal semantic features of the dual-time remote sensing images after channel swapping. The specific implementation details of the spatial and channel fusion module are illustrated in Fig. 4(a). This method utilizes channel and spatial attention to identify important parts of the features. In the channel branch, the input dual-temporal features are aggregated spatially through global pooling across spatial dimensions. In the spatial branch, the input dual-temporal features are aggregated across channel dimensions through global pooling. The deep fusion of both channel and spatial information results in the output aggregated dual-time features containing rich global dual-temporal information features. The spatial and channel fusion module can be formalized as follows:

$$S_c = \text{Concat}\begin{pmatrix} \text{Channel dim}\,(\text{Avg}(T_1), \text{Max}(T_1)) \\ \text{Channel dim}\,(\text{Avg}(T_2), \text{Max}(T_2)) \end{pmatrix} \tag{7}$$

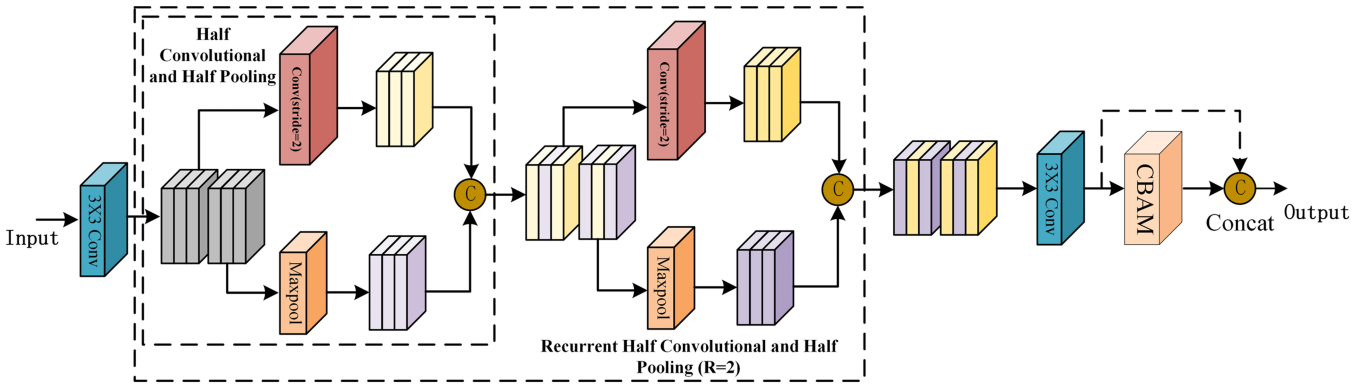

Fig. 3. CGFE comprises four semiconvolutional pooling modules, enabling the extraction of rich dual-temporal information and facilitating precise localization of changing targets.

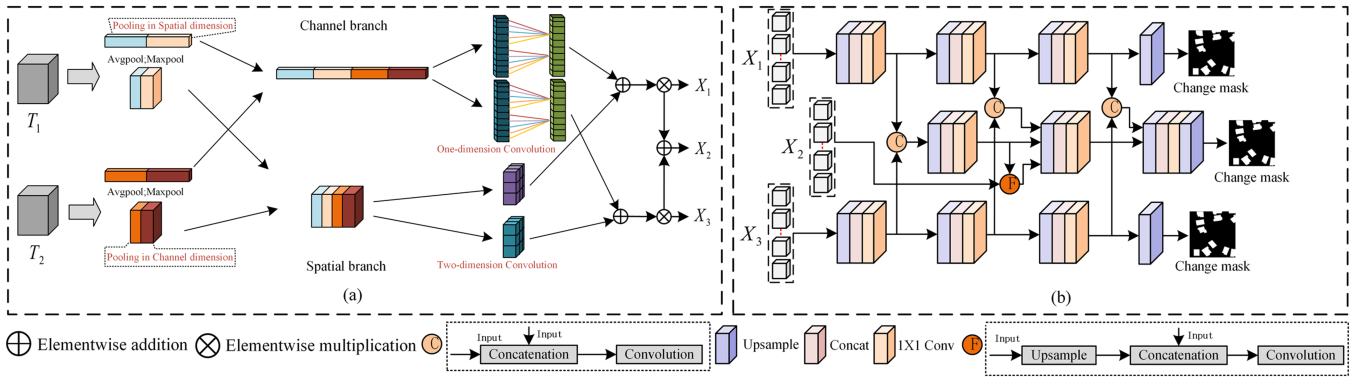

Fig. 4. Spatial and channel fusion module (a) utilizes channel and spatial attention to achieve a more comprehensive fusion of temporal information between diachronic features. The channel feature-guided decoder (b) to accurately localize all change objects and generate change maps as model results.

$$S_s = \text{Concat}\begin{pmatrix}\text{Spatial dim}\,(\text{Avg}(T_1), \text{Max}(T_1)) \\ \text{Spatial dim}\,(\text{Avg}(T_2), \text{Max}(T_2))\end{pmatrix} \tag{8}$$

$$X_1 = X_3 = \text{Conv}(S_c, S_s) \tag{9}$$

$$X_2 = X_1 + X_3 \tag{10}$$

where T_1 and T_2 represent the dual-temporal features, S_c represents the fused channel dual-temporal features, S_s represents the fused spatial dual-temporal features, and X_1, X_2, X_3, respectively, denote different output features.

The CGF branch first accurately extracts layer-by-layer features of the dual-temporal images through the CGFE. It utilizes spatial and channel attention for deep fusion of the dual-temporal features. After passing through the CGFD, it accurately locates all changed objects, ultimately generating a change map as the model's result. The implementation details of the CGFD are illustrated in Fig. 4(b). The training process of the entire module branch can be represented as follows:

$$Y = \text{CGFD}(\text{SCF}(\text{CGFE}(X, T))) \tag{11}$$

where X represents the input feature map, and T represents the temporal information. CGFE stands for the CGFE, which takes input feature map X and temporal information T, and outputs the processed feature map after encoding. SCF represents the spatial and channel fusion module, which takes the channel-swapped dual-temporal features and utilizes spatial and channel attention to determine important feature parts, facilitating a more thorough fusion of feature information between the dual temporalities. CGFD represents the channel feature-guided decoder, which accepts the input feature map after deep fusion processing and generates the output change map. Y represents the final result of CD in this branch.

C. Adaptive Feature Refinement Module

The AFRM is a module designed to address the issues of insufficient contextual guidance, excessive noise, and difficulty in information aggregation in the multilevel feature fusion process of the feature pyramid network, as illustrated in Fig. 5. It recalibrates features to provide improved multilevel feature fusion capability.

In the AFRM module, let d_1 represent the input features. Initially, a 1×1 convolutional layer with an S-shaped activation function is employed to generate the change map c_1. This indicates that within the AFRM module, convolutional operations with activation functions are introduced to generate the change map c_1, representing the recalibration of features based on the

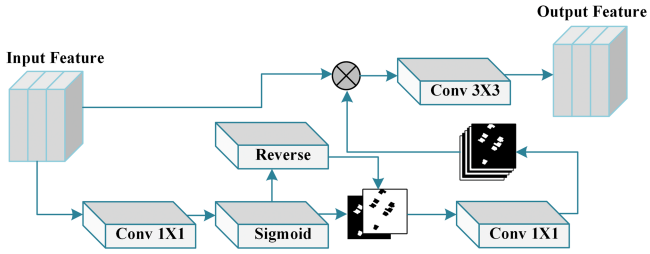

Fig. 5. AFRM enhances the fusion capability of multilevel features by recalibrating them.

input features d_1. The formula is as follows:

$$c_1 = \sigma(\text{Conv}_{(1\times1)}(d_1)) \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid function, used to transform the change mapping into the range [0, 1]. Subsequently, we obtain the reverse change mapping $c_1^r = \text{Rev}(c_1)$ through inverse operations. $Rev(c_1)$ is computed by subtracting c_1 from a matrix E, where all elements are 1. The derived change map c_1 and the reversed change map c_1^r depict contextual information about changed objects and unchanged background. Here, we employ a 1×1 convolutional layer to generate the pixelwise attention mask a_1 from c_1 and c_1^r, which has the same shape as d_1, as shown as follows:

$$a_1 = \text{Conv}_{(1\times1)}(\text{Concat}(c_1, c_1^r)). \tag{13}$$

Subsequently, we recalibrate d_1 using a_1 to obtain the feature guided by the AFER module. This process can be described as follows:

$$d_1^r = \text{Conv}_{(3\times3)}(d_1 \otimes a_1). \tag{14}$$

Here, $\otimes$ denotes elementwise multiplication operation. In such a formula, a_1 can be regarded as a pixelwise gate used to guide the learning process of d_1, allowing only information relevant to the changed objects to pass through. Finally, the supervised attention-guided feature d_1^r generated by AFRM is further fused with higher-level features. This allows AFRM to reweight the features extracted by the transformer across different dimensions, thereby refining the accuracy of the features extracted by the transformer and further enhancing the feature extraction capability of the transformer branch.

D. Global Context Integration Module

Capturing long-range dependencies can harness useful contextual information that benefits visual understanding tasks. Traditional CNN models may be restricted by local features when addressing long-range dependencies, making it difficult to fully capture global information. This limitation often results in suboptimal performance in tasks that require an understanding of global context. In this work, we propose a GCIM that amalgamates contextual information to enhance the low-level, mid-level, and high-level feature information extracted by CNNs, thereby improving the model's semantic understanding capabilities. As illustrated in Fig. 6, "R" denotes the number of iterations for the deep associative attention module, with multiple deep associative attention modules sharing parameters to augment the CNN's ability to identify critical features in CD more effectively and efficiently. Specifically, the GCIM consists of two deep associative attention modules. The inner structure of the deep associative attention is depicted within the dashed lines in Fig. 6. Through two consecutive deep associative attention processes, each pixel can aggregate feature information from all pixels, thereby capturing the context information of its surrounding pixels across intersecting paths. By engaging in further iterative operations, each pixel can ultimately capture the long-range dependencies of all pixels, further enhancing the feature extraction capabilities of CNNs. The training process for this module is elucidated by the equations below. The following describes the training process of this module:

$$\text{GCIM}(X) = \text{DAA}(\text{DAA}(X)) + X \tag{15}$$

where X represents the input features (comprising information from low, mid, and high dimensions), $GCIM$ represents the global context integration module, and DAA represents the deep associative attention module.

The GCIM module proposed in this work is an independent module that can be seamlessly integrated into any CNN architecture at any point in the network to capture rich contextual information. The module is computationally inexpensive and adds only a few adjustable parameters, resulting in minimal GPU memory usage.

E. Deep Convolutional Attention Module

To facilitate the deep fusion of bitemporal features extracted by both CNN and transformer, capturing complementary information of the same scene from two different perspectives, we propose a linearized deep convolutional attention module (DCA) inspired by traditional cross-attention mechanisms for interactive propagation. This approach encourages one branch to retain the original features while glimpsing feature information from the other branch. We achieve feature coupling by alternately using the linearized expansions of Q from both CNN and transformer branches. Specifically, for cross-branch interaction, we apply the Q from one branch to connect with the K and V from the other branch, facilitating feature interaction between different branches and enabling deeper exploration of the feature space. The implementation details are illustrated in Fig. 7. The following equations demonstrate the specific implementation process of this module:

$$\text{DCA} = \text{Concat}[\text{softmax}(K) \times V \times Q, \text{Hadamard}(\text{Conv}(V), Q)] \tag{16}$$

$$M_1 = \text{DCA}(Q_t, K_c, V_c) \tag{17}$$

$$M_2 = \text{DCA}(Q_c, K_t, V_t) \tag{18}$$

where M_1 represents the output features of the CNN branch after feature interaction, M_2 represents the output features of the transformer branch after feature interaction, Q_t, K_t, V_t, represent the key-value pairs of the image features extracted by the transformer Q_c, K_c, V_c, represent the key-value pairs of

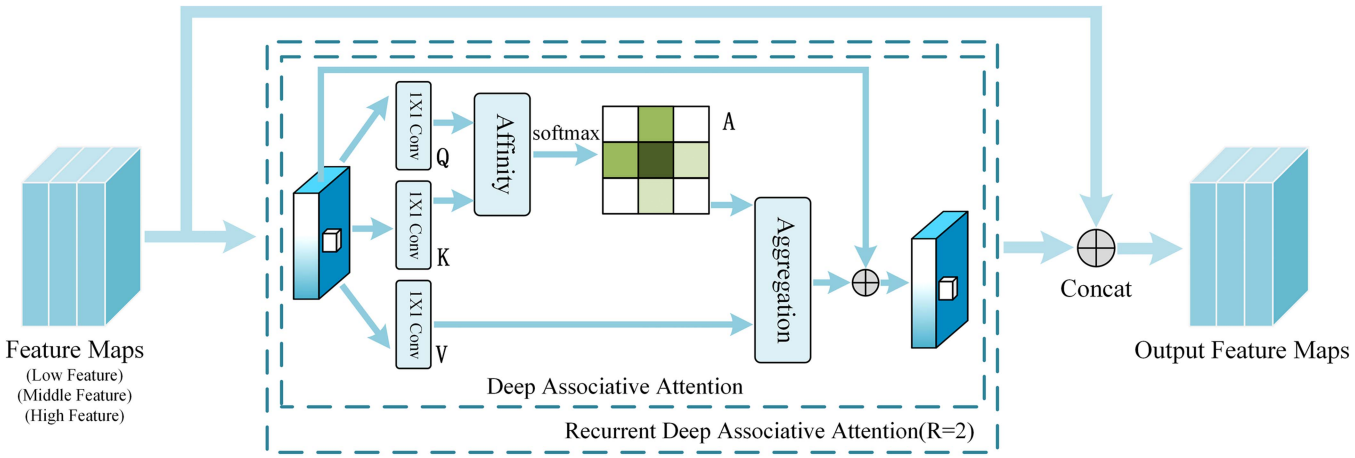

Fig. 6. Global context integration module consists of two deep associative attention modules designed to capture contextual information and enhance feature extraction. The internal structure of the deep associative attention module is shown within the dashed line and is designed to capture the long-range dependencies of all pixels.

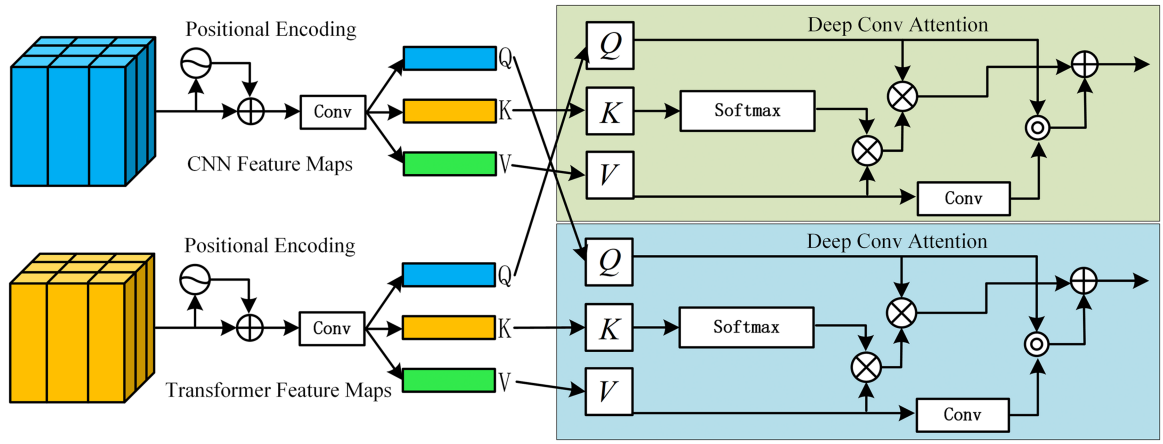

Fig. 7. DCA module: Linearly expanding the feature maps of CNN and transformer into vectors to capture complementary information from the same image under two different perspectives, facilitating a better deep fusion of bitemporal features.

the image features extracted by the CNN, and Hadamard denotes the elementwise product. In Fig. 7, it is represented as $\odot$.

F. Loss Function

In remote sensing CD tasks, we need to train a model to learn to classify remote sensing images to differentiate between changed and unchanged regions. To ensure that the model accurately classifies the images and sensitively learns changes, we have selected the cross-entropy loss function as our optimization objective, as shown in (19), where $p(x, y)$ represents the ground truth (GT) pixel values, and $\widehat{p}(x, y)$ represents the predicted pixel values. The cross-entropy loss function measures the difference between the model's predicted results and the GT labels to drive the model to learn accurate classification boundaries. This loss function can be expressed as the average of the negative log-likelihood function. Specifically, for each sample in the remote sensing CD task, we calculate the cross-entropy loss using the model's predicted probability distribution and the true label information. During training, we optimize the model parameters by minimizing the cross-entropy loss between f_1, f_2, f_3, and the GT individually. The total training loss function is shown in (20)

$$L_{CE(p,\widehat{p})} = \frac{-1}{WH}\sum_{x=0}^{W-1}\sum_{y=0}^{H-1} p(x,y)\log\widehat{p}(x,y) + (1-p(x,y))\log(1-\widehat{p}(x,y)) \tag{19}$$

$$L_{\text{total}} = \lambda_1 L(g,\phi(f_1)) + \lambda_2 L(g,\phi(f_2)) + \lambda_3 L(g,\phi(f_3)) \tag{20}$$

where $\phi(\cdot)$ consists of the Sigmoid function applied pixelwise along the channel dimension and upsampling operation consistent with the size of the input image, λ_1, λ_2, and λ_3 are tradeoff parameters used to adjust the influence of each loss. Following

experimental comparison, here, we select the parameters to be 1, 0.5, and 0.5, respectively. During the inference stage, the predicted mask is computed by performing pixelwise Argmax operation along the channel dimension on the sum of $\phi(f_1)$, $\phi(f_2)$, and $\phi(f_3)$.

IV. Experiment

This section provides a detailed introduction to the experimental part, including the equipment information used in the experiments, the setting of experimental parameters, the configuration of the loss function, as well as the design of comparative experiments and ablation studies. These experiments aim to evaluate the detection performance and effectiveness of the proposed network framework MVAFG.

A. Experiment Details

The network architecture MVAFG was implemented using the PyTorch toolkit and the program was run in the PyTorch1.8 environment. Training and inference stages were conducted using two NVIDIA TITAN RTX GPUs, each with 12 GB of memory. For the backbone of the network, we initialized parameters using pretrained ResNet18 [41] and PVT-b2 [42] models. Consistent with other comparative experiments, we applied common data augmentation operations to the input image blocks, including flipping, resizing, cropping, and Gaussian blurring. Through these operations, we increased the diversity of training data, enhancing the robustness and generalization capability of the network. For model optimization, we employed the AdamW [43] optimizer with a momentum parameter set to 0.9 and weight decay set to 0.001, while parameters $\beta 1$ and $\beta 2$ were set to 0.9 and 0.99, respectively. The initial learning rate was set to 0.0005, and the batch size was set to 16. These configurations were chosen to achieve better convergence speed and performance during training, balancing model complexity and computational resource consumption.

B. Datasets

LEVIR-CD dataset [44] is a large-scale dataset specifically designed for CD tasks, consisting of very high-resolution (0.5 meters/pixel) Google Earth images. These images capture various types of changes in buildings that have occurred over the past 5–14 years. The dataset primarily focuses on changes related to buildings, such as building growth and decline. To annotate the change regions in these images, experts used binary masks, where "1" indicates change and "0" indicates no change. There are a total of 31 333 independent change samples in the entire dataset. For ease of model training and evaluation, we cropped the images into patches of size 256 × 256 pixels. Subsequently, the dataset was divided into training, validation, and testing sets in a ratio of 7:1:2.

WHU-CD dataset [45] (The Wuhan University Change Detection) is focused on detecting changes in buildings. This dataset comprises a pair of aerial images with spatial dimensions of 32507 × 15354 and a resolution of 0.2 meters/pixel. These images document the changes in buildings in the Christchurch area of New Zealand following a 6.3 magnitude earthquake in 2011. Due to limited information in reference regarding the sample partitioning scheme, we cropped these large-scale image pairs and generated patches of size 256 × 256 pixels. Subsequently, we randomly split the total patch pairs into 6096 for training, 762 for validation, and 762 for testing, according to certain proportions.

The Guangzhou Dataset (GZ-CD) [36] is a dataset focused on high-resolution satellite imagery CD with a spatial resolution of 0.55 meters per pixel. This dataset covers suburban areas of Guangzhou, China, spanning from 2006 to 2019. The dataset comprises 19 pairs of seasonally varying images, with dimensions ranging from 1006 × 1168 pixels to 4936 × 5224 pixels. For uniform processing, images were cropped into patches of 256 × 256 pixels. The dataset is divided into training, validation, and testing sets, with a total of 2504 training samples, 313 validation samples, and 313 testing samples. Compared to the LEVIR-CD and WHU-CD datasets, the Guangzhou Dataset (GZ-CD) is relatively small in scale. Nevertheless, it still provides an important resource for the research and evaluation of building CD algorithms.

C. Comparison Method

In order to comprehensively compare the performance of this network, we selected mainstream and state-of-the-art methods for comparison. We conducted extensive experiments on three dual-temporal remote sensing image CD datasets to evaluate the effectiveness of the MVAFG model. We compared MVAFG with the following nine state-of-the-art methods, which are FC-EF [23], FC-Siam-Di [23], FC-Siam-Conc [23], SNUNet [29], IFNet [5], BIT [28], ChangeFormer [29], FTN [46], and ICIF -Net [30].

D. Evaluation Metrics

For the performance evaluation of the model, we employ five metrics to measure the similarity between the predicted change probability maps and the GT, including Precision (Pre.), Recall (Rec), F1-score (F1), intersection over union (IoU), and overall accuracy (OA). In binary classification tasks, OA is used to compute the ratio of correctly classified samples to the total number of samples. Its maximum value is 1, indicating better performance as the model's result approaches 1. The F1-score combines Precision and Recall with equal weights. Its maximum value is 1, effectively reflecting the model's accuracy and its ability to correctly identify positive samples. IoU represents the overlap between the model's predicted results and the GT, with a maximum value of 1. Precision indicates the proportion of true positive samples among samples predicted as positive, while Recall indicates the proportion of correctly identified positive samples among the actual positive samples. The maximum values for Precision and Recall are both 1. Each metric is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

TABLE I
MODULE ABLATION MODEL DEMONSTRATION

Method	DCA	GCIM	AFRM
MVAFG-1	×	✓	✓
MVAFG-2	✓	×	✓
MVAFG-3	✓	✓	×
MVAFG	✓	✓	✓

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

$$F1 = \frac{2}{Recall^{-1} + Precision^{-1}} \tag{23}$$

$$IOU = \frac{TP}{TP + FP + FN} \tag{24}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{25}$$

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. It is worth noting that $F1$ is a composite metric that integrates $Precision$ and $Recall$.

E. Ablation Experiment

To further verify the impact of various modules and different branches on the performance of remote sensing CD, this section conducted comprehensive ablation experiments to assess the effects of different modules and branches, and presented the experimental results across three datasets. All experimental results were visualized using GT overlays, where red indicates misidentified areas, and green represents areas that were not detected.

1) Module Verification: The specific models used for the ablation study are shown in Table I.

MVAFG-1 represents the network structure with the DCA module disabled. To verify the importance of feature interaction between CNN and transformer, we disabled the DCA module. Experimental results demonstrate that the DCA module significantly enhances feature interaction between different branches.

MVAFG-2 represents the network structure with the GCIM module disabled. To validate the importance of the GCIM module in enhancing CNN feature extraction, we disabled the GCIM module. Experimental results demonstrate that the GCIM module indeed enhances the ability of CNN feature extraction.

Similarly, in MVAFG-3, we disabled the AFRM module to validate its capability in enhancing features of different dimensions extracted by the transformer. Experimental results indicate that the AFRM module indeed possesses the ability to improve the features extracted by the transformer across different dimensions.

LEVIR-CD dataset: On the LEVIR-CD dataset, we conducted experiments on several dilution models, and some experimental results are shown in Fig. 8. Although MVAFG-3 yielded similar results to MVAFG in handling the second group of samples,

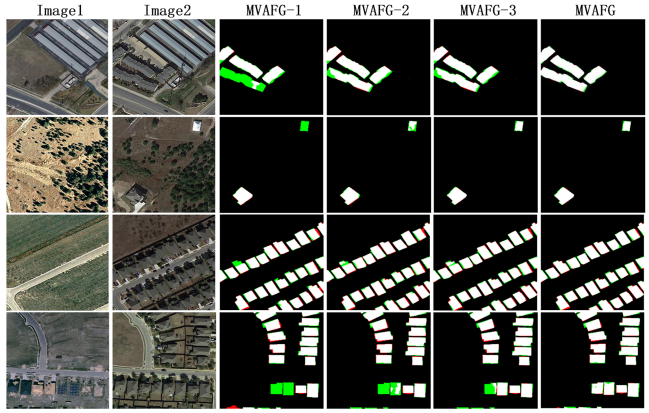

Fig. 8. Results of module ablation experiments on LEVIR-CD.

TABLE II
INDICATORS FOR ABLATION MODELS ON THE LEVIR-CD DATASET

Method	OA	F1	IoU	Precision	Recall
MVAFG-1	99.12	91.49	84.32	92.35	89.97
MVAFG-2	99.15	91.56	84.43	92.68	90.46
MVAFG-3	99.16	91.62	84.55	92.41	89.93
MVAFG	99.19	91.82	84.95	93.12	90.75

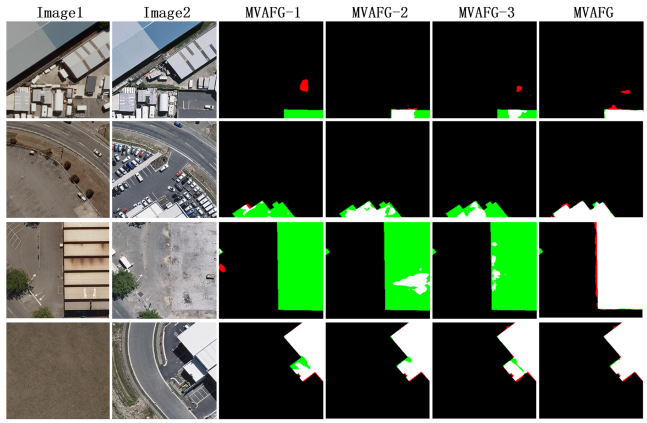

Fig. 9. Results of module ablation experiments on WHU-CD.

varying degrees of unrecognized areas were observed in processing other groups of samples. This indicates that without enhancing the transformer for long-range dependency, it cannot effectively capture change features. Table II lists the performance metrics of several dilution models and MVAFG on the LEVIR-CD dataset. Comparatively, the performance metrics of several dilution models are lower than MVAFG. Overall, through subjective analysis, it is supported that the modules and processing mechanisms corresponding to the five dilution models positively affect the performance of the network. This indicates that the MVAFG network architecture we finally selected is reasonable and effective.

WHU-CD dataset: Fig. 9 illustrates the processing results of various dilution models on the WHU-CD dataset. This dataset

TABLE III
INDICATORS FOR ABLATION MODELS ON THE WHU-CD DATASET

Method	OA	F1	IoU	Precision	Recall
MVAFG-1	99.19	91.52	84.36	92.49	90.37
MVAFG-2	99.23	91.85	85.12	92.98	90.64
MVAFG-3	99.21	91.76	84.92	92.95	90.56
MVAFG	99.39	92.35	85.49	93.12	91.31

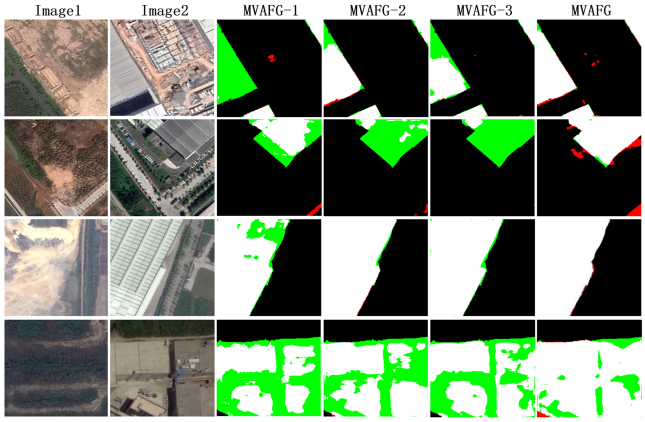

Fig. 10. Results of module ablation experiments on GZ-CD.

TABLE IV
INDICATORS FOR ABLATION MODELS ON THE GZ-CD DATASET

Method	OA	F1	IoU	Precision	Recall
MVAFG-1	97.52	86.88	76.18	89.53	83.21
MVAFG-2	97.74	87.72	76.68	90.12	84.06
MVAFG-3	97.62	87.36	76.58	90.49	82.72
MVAFG	97.89	88.12	77.08	90.13	84.88

TABLE V
BRANCH ABLATION MODEL DEMONSTRATION

Method	CNN	Transformer	CGF
Branch 1	✓	×	×
Branch 2	✓	✓	×
Branch 3	✓	✓	✓

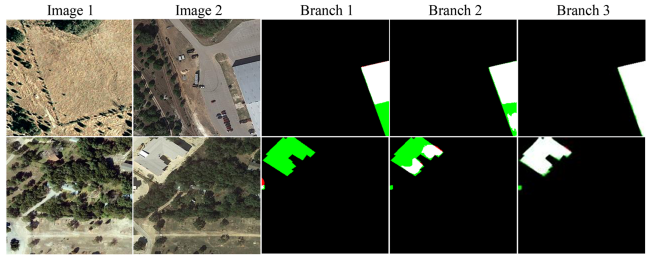

Fig. 11. Results of branch ablation experiments on LEVIR-CD.

has a high resolution, enabling better display of detailed information, but it also introduces more prominent environmental interference factors. In the first and second groups of samples, environmental factors affected MVAFG-1, MVAFG-2, and MVAFG-3, with the results showing that only MVAFG displayed the changed areas more comprehensively, albeit with some slight issues in edge detection. However, overall, the experimental results of MVAFG are the best when compared comprehensively. Table III, lists the performance metrics of several dilution models and MVAFG on the WHU-CD dataset. Through comprehensive comparative analysis, it is supported that the MVAFG network architecture we finally selected is reasonable and effective.

GZ-CD dataset: As shown in Fig. 10, the results of the final model and five dilution models on the GZ-CD dataset are presented. In the second group of samples, MVAFG-1, MVAFG-2, and MVAFG-3 all failed to identify changed areas, with only MVAFG recognizing the changed areas relatively comprehensively, albeit with many false detections. However, overall, the experimental results of MVAFG are the best when compared comprehensively. This indicates that the three-branch structure and feature enhancement modules contribute to improving the model's performance and also validates the rationality of the MVAFG network architecture. Table IV lists the performance metrics of several dilution models and MVAFG on the GZ-CD dataset. The experimental results show that, compared to several dilution experiments, the MVAFG architecture performs well. However, there is still some gap between the experimental results of MVAFG and the GT, indicating that there is room for improvement in the performance of MVAFG. In the future, we will continue to optimize MVAFG to achieve high-quality CD.

2) Branch Verification: It is important to note that the proposed module is a multibranch structure, thus it is necessary to study the contribution of each branch. We simplified the complete MVAFG into three baseline branches: CNN, transformer, and CGFM. We validate the contributions of multibranch approach by adding branches incrementally. The specific models used for the ablation study are shown in Table V.

Ablation experiments on LEVIR-CD: On the LEVIR-CD dataset, we conducted experiments with several ablation models, with partial results shown in Fig. 11. In the first and second sample groups, thanks to the advantage of transformer's global context modeling, the dual-branch Branch 2 model consistently outperformed the single-branch Branch 1 model in feature handling. Furthermore, by comparison, only the Branch 3 model comprehensively displayed the change areas, directly validating the triple-branch model structure's ability to compensate for the deficiencies in feature extraction by transformer and CNN, thereby proving its dominant position in feature extraction. Table VI lists the performance metrics of the three ablation branches on the LEVIR-CD dataset. By comparison, the performance of the proposed triple-branch network consistently surpasses that of the single-branch and dual-branch networks, demonstrating that the triple-branch network structure is both reasonable and effective.

Ablation experiments on WHU-CD: On the WHU-CD dataset, we conducted experiments with several ablation models, with partial results shown in Fig. 12. In the first and second

TABLE VI
INDICATORS FOR BRANCH ABLATION MODELS ON THE LEVIR-CD DATASET

Method	OA	F1	IoU	Precision	Recall
Branch 1	99.02	90.38	82.47	98.06	89.21
Branch 2	99.14	91.35	84.25	92.94	90.01
Branch 3	99.19	91.82	84.95	93.12	90.75

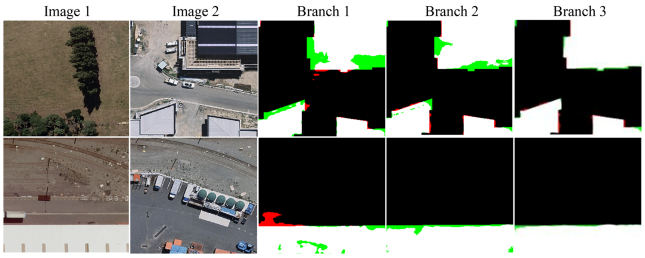

Fig. 12. Results of branch ablation experiments on WHU-CD.

TABLE VII
INDICATORS FOR BRANCH ABLATION MODELS ON THE WHU-CD DATASET

Method	OA	F1	IoU	Precision	Recall
Branch 1	99.04	89.57	82.68	90.89	85.56
Branch 2	99.18	91.49	84.26	92.78	90.34
Branch 3	99.39	92.35	85.49	93.12	91.31

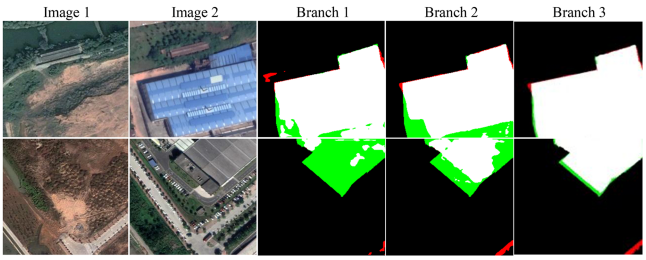

Fig. 13. Results of branch ablation experiments on GZ-CD.

sample groups, thanks to the advantage of transformer's global context modeling, the dual-branch Branch 2 model consistently outperformed the single-branch Branch 1 model in handling features, although there remained some gaps in processing local features. By comparison, only the Branch 3 model comprehensively displayed the change areas, directly validating the triple-branch model structure's ability to compensate for the deficiencies in feature extraction by transformer and CNN, thereby proving the dominant role of the CGF branch in feature extraction. Table VII lists the performance metrics of the three ablation branches on the WHU-CD dataset. By comparison, the performance of our proposed triple-branch network consistently surpasses that of the single-branch and dual-branch networks, demonstrating that the triple-branch network structure is both reasonable and effective.

Ablation experiments on GZ-CD: On the GZ-CD dataset, we conducted experiments with several ablation models, with partial results shown in Fig. 13. In the first and second sample groups, only the Branch 3 model comprehensively displayed the change areas, while Branch 1 and Branch 2 exhibited clear deficiencies in feature extraction. This directly validates the triple-branch model structure's ability to compensate for the deficiencies in feature extraction by transformer and CNN, thereby confirming the dominant role of the CGF branch in feature extraction. Table VIII lists the performance metrics of the three ablation branches on the GZ-CD dataset. By comparison, the performance of the proposed triple-branch network consistently surpasses that of the single-branch and dual-branch networks, demonstrating that the triple-branch network structure is both reasonable and effective.

TABLE VIII
INDICATORS FOR BRANCH ABLATION MODELS ON THE GZ-CD DATASET

Method	OA	F1	IoU	Precision	Recall
Branch 1	96.31	82.19	71.02	86.72	78.68
Branch 2	97.36	86.13	75.79	89.12	83.43
Branch 3	97.89	88.12	77.08	90.13	84.88

This section of this article presents only the most representative ablation models and experimental data. Through ablation studies, we have finalized the architecture of the network, which is MVAFG.

F. Comparison Experiment

After finalizing the network architecture, we conducted comparisons between MVAFG and current mainstream as well as state-of-the-art methods across three datasets. All compared methods underwent a maximum of 200 training epochs, with the best model updated at the end of each epoch. We uniformly set the learning rate to 0.0005. Other parameters for each compared method were assigned as detailed in their respective papers. Experiments were grouped according to the different datasets. All experimental results were visualized using GT visual overlays, where red denotes misidentified regions and green represents unrecognized regions.

1) LEVIR-CD: Fig. 14 illustrates the results obtained by various methods on the LEVIR-CD dataset. We selected the most representative eight sets of samples. In the first set of images, due to differences in the color of the buildings, all previous methods failed to identify the variations caused by external factors affecting the building colors, whereas only the method detected results most similar to the GT. Other methods exhibited significant deficiencies. The second and third sets both involve detection of large buildings. SNUNet, IFNet, and Changeformer exhibited varying degrees of detection omission. Although BIT and ICIF-Net performed well in CD, they both identified erroneous regions, mistaking environmental interference factors as detection targets. Overall, the method yielded results closest to the GT. The fourth and fifth sets contain densely populated residential areas with small and densely packed buildings. Nearly all methods showed varying degrees of detection omission. The method roughly detected the target areas of changes. From the processing results, the method not only outperformed all compared methods in edge handling but also performed well

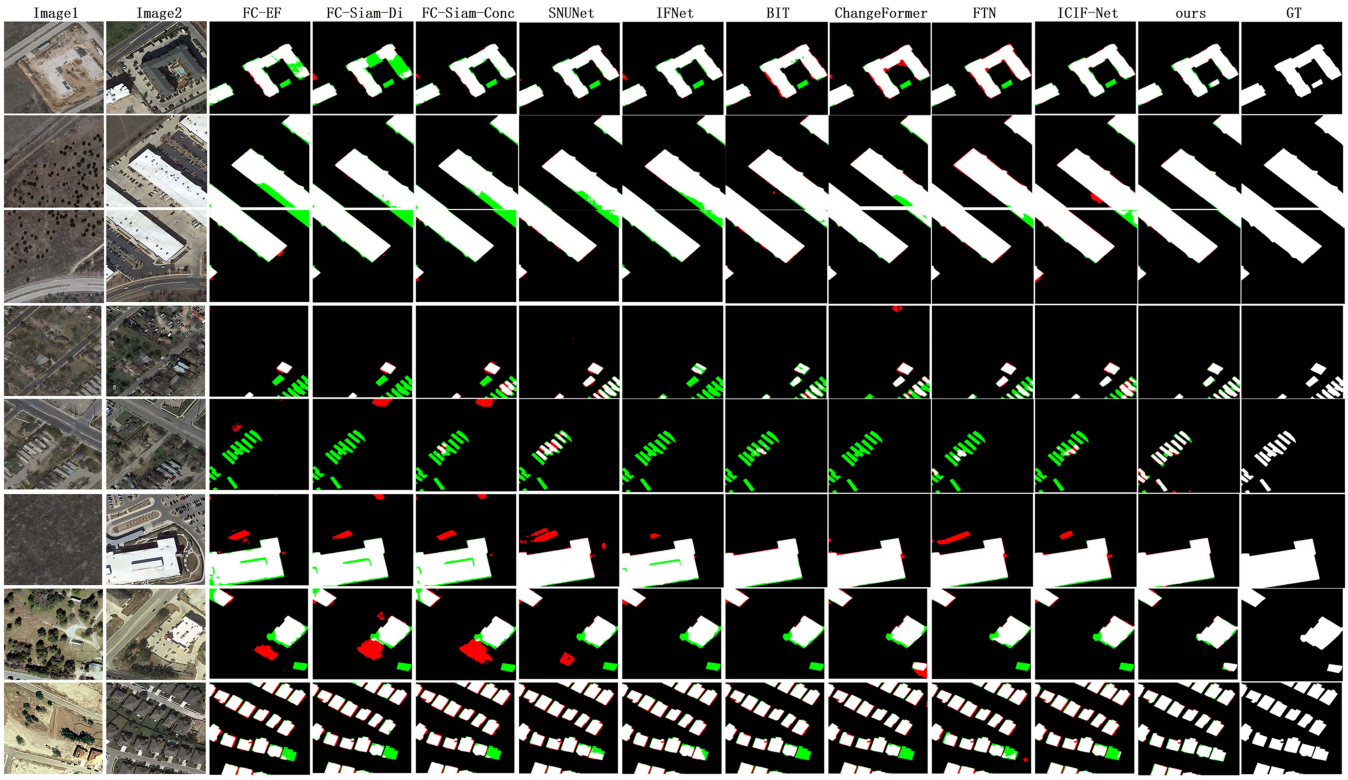

Fig. 14. Comparative experiments on the LEVIR-CD dataset. Different colors are used for better visualization: White indicates true positives, black indicates true negatives, red indicates false positives, and green indicates false negatives.

in detecting small targets. The sixth set involves detection of large buildings with significant interference factors. Although BIT, Changeformer, FTN, and ICIF-Net performed well in target detection, they output factors unrelated to the changing areas, resulting in erroneous detection. In comparison, only our detection results were most similar to the GT. The seventh set comprises buildings with varying degrees of interference from roads and forests. Most methods performed poorly in detecting changes in edge regions, with undetected edge areas. Although there is still a gap between our method and the results obtained by GT in edge handling, compared to all previous methods, we still outperformed all compared methods. In the eighth set of samples, including densely populated residential areas, due to the different acquisition times of the two-temporal images, the target buildings in the samples have different colors. From the processing results, networks such as IFNet, BIT, and ICIF-Net lacked detection of small changes in targets. Our method not only outperformed all compared methods in edge handling but also performed well in detecting small targets.

Table IX presents the performance metrics evaluated for all methods on the LEVIR-CD dataset. MVAFG achieved the highest scores across all five metrics. Notably, in terms of the F1 score, MVAFG outperformed the second-ranked ICIF-Net by 0.64 points. Similarly, MVAFG exhibited superior performance to ICIF-Net by 1.1 point in terms of IoU. Based on a comprehensive evaluation utilizing subjective and objective metrics, MVAFG demonstrated outstanding performance on the LEVIR-CD dataset. MVAFG surpassed mainstream and state-of-the-art methods in target recognition, edge handling, and small target detection. However, MVAFG exhibited suboptimal performance in handling some complex samples, indicating room for improvement in its robustness.

TABLE IX
INDICATOR RESULTS FOR EACH COMPARISON METHOD ON THE LEVIR-CD DATASET

Method	OA	F1	IoU	Precision	Recall
FC-EF	98.38	84.02	72.45	84.39	83.66
FC-Siam-Di	98.71	87.17	77.26	88.63	85.76
FC-Siam-Conc	98.69	87.25	77.39	86.37	88.15
IFNet	98.95	89.32	80.7	92.57	86.29
SNUNet	98.67	86.32	75.93	90.33	82.65
BIT	99.03	90.34	82.39	91.53	89.19
ChangeFormer	99.04	90.4	82.48	92.05	88.81
FTN	99.06	91.01	83.51	92.71	89.37
ICIF-Net	99.12	91.18	83.85	91.13	90.57
MVAFG	99.19	91.82	84.95	93.12	90.75

Red represents the best results, while blue represents the second best results.

2) WHU-CD: Fig. 15 illustrates partial experimental results on the WHU-CD dataset. We have selected the most representative six sets of samples. The first set of samples involves detection of large buildings, but due to shadow effects, SNUet, IFNet, and Changeformer lack detection coherence and fail to detect prominent small target buildings. Although BIT and FTN accurately identify prominent areas, they erroneously output some environmental interference factors as targets. In contrast, our

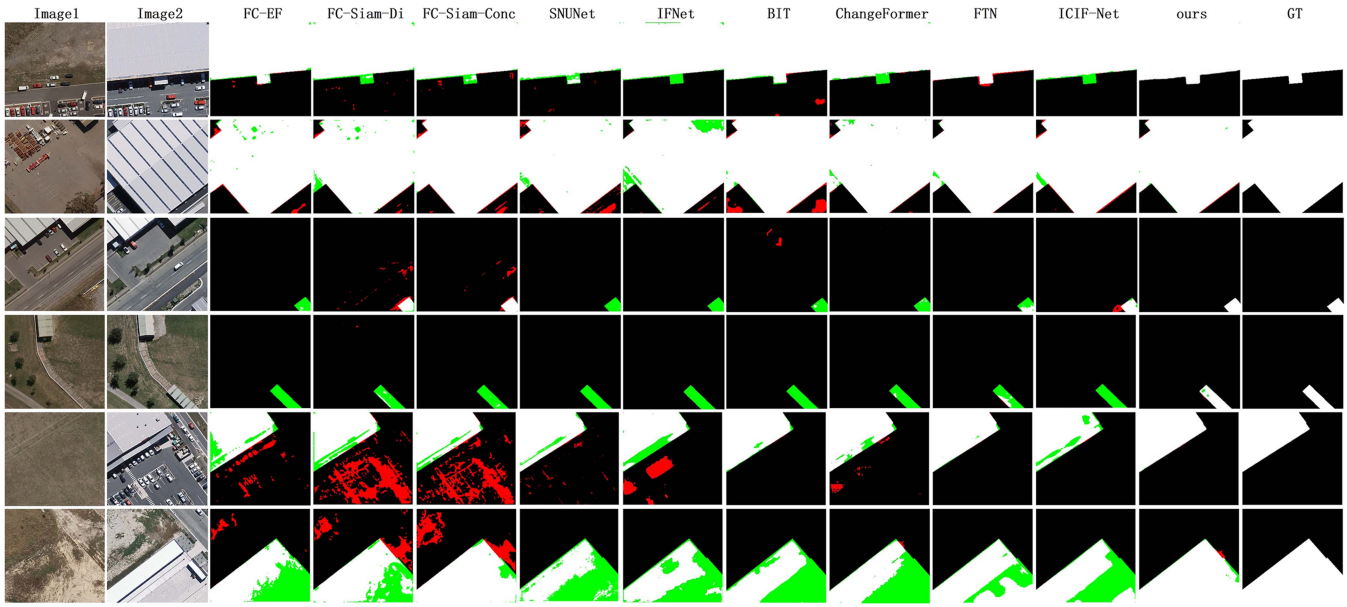

Fig. 15. Illustrates the comparative experiments on the WHU-CD dataset. Different colors are utilized for enhanced visualization: white represents true positives, black indicates true negatives, red signifies false positives, and green denotes false negatives.

detection results are closest to the GT. The second set of samples also pertains to the detection of large buildings, where changes in shadows caused by lighting result in partial occlusion of the edges. Almost all methods exhibit incomplete detection in these edge occlusion areas, whereas only the method fully detects the changed areas, most closely resembling the GT. The third and fourth sets both involve small target change areas, where the influence of surrounding environments and backgrounds makes it challenging for all previous methods to accurately detect the target area changes. In comparison, our method achieves better detection results. The fifth and sixth sets also pertain to the detection of large buildings, but due to excessive background interference, almost all methods exhibit varying degrees of target omission areas. Only our method identifies the changed areas more comprehensively. In the sixth set of samples, our method still has small undetected areas in the edge detection part. However, compared to all other methods, our method clearly has fewer undetected areas and is closer to the GT.

According to Table X, MVAFG achieves the highest scores in all four metrics. Specifically, MVAFG outperforms ICIF-Net by 1.58 points in F1 and by 2.4 points in IoU. However, MVAFG did not achieve the optimal results in some metrics, indicating room for improvement in the network. Nonetheless, it is noteworthy that when faced with challenging environmental factors, MVAFG utilizes a three-branch structure combined with CNN and transformer to extract comprehensive information from multiple aspects, enhancing the feature information at each stage. Additionally, the GCIM module enhances the positional information of global context features, thereby improving the localization of detected targets. However, there is still room for improvement in MVAFG. It still encounters challenges in handling edge features in samples in complex environments, as shown in the sixth set of samples in Fig. 15, where there are still some issues with false detections and omissions. Addressing this limitation will be a focus of our future efforts.

TABLE X
INDICATOR RESULTS FOR EACH COMPARISON METHOD ON THE WHU-CD DATASET

Method	OA	F1	IoU	Precision	Recall
FC-EF	98.02	77.64	63.45	84.61	71.74
FC-Siam-Di	98.06	79.97	66.63	79.78	80.17
FC-Siam-Conc	98.45	84.13	72.6	83.07	85.19
IFNet	98.83	83.41	71.53	96.91	73.2
SNUNet	98.91	88.34	79.11	91.34	85.53
BIT	98.81	87.47	77.73	88.71	86.27
ChangeFormer	98.76	86.88	76.81	88.5	85.33
FTN	99.37	92.16	85.45	93.09	91.24
ICIF-Net	99.13	90.77	83.09	92.93	88.7
MVAFG	99.39	92.35	85.79	93.12	91.31

Red represents the best results, while blue represents the second best results.

3) GZ-CD: The experimental results on the GZ-CD dataset are depicted in Fig. 16. In the processing of the first set of samples, both BIT and Changeformer exhibit inferior capabilities in identifying changes in large buildings compared to our method. This is primarily because both of them utilize single CNN or transformer architectures for feature extraction. Although FTN also solely employs the Swin Transformer as the backbone network for feature extraction, the sliding window and multiscale processing of the Swin Transformer better attend to feature information. FTN shows results similar to ours in the first set of samples. Our method, employing a three-branch structure that combines the strengths of CNN and transformer, can better focus on richer feature information and effectively integrate semantic information. This allows features of large-scale changing targets to be assigned higher weights, ultimately resulting in our method achieving results closest to the GT. The second and third sets both involve the detection of large, dense buildings with complex and dense changes, posing high detection difficulty. Almost

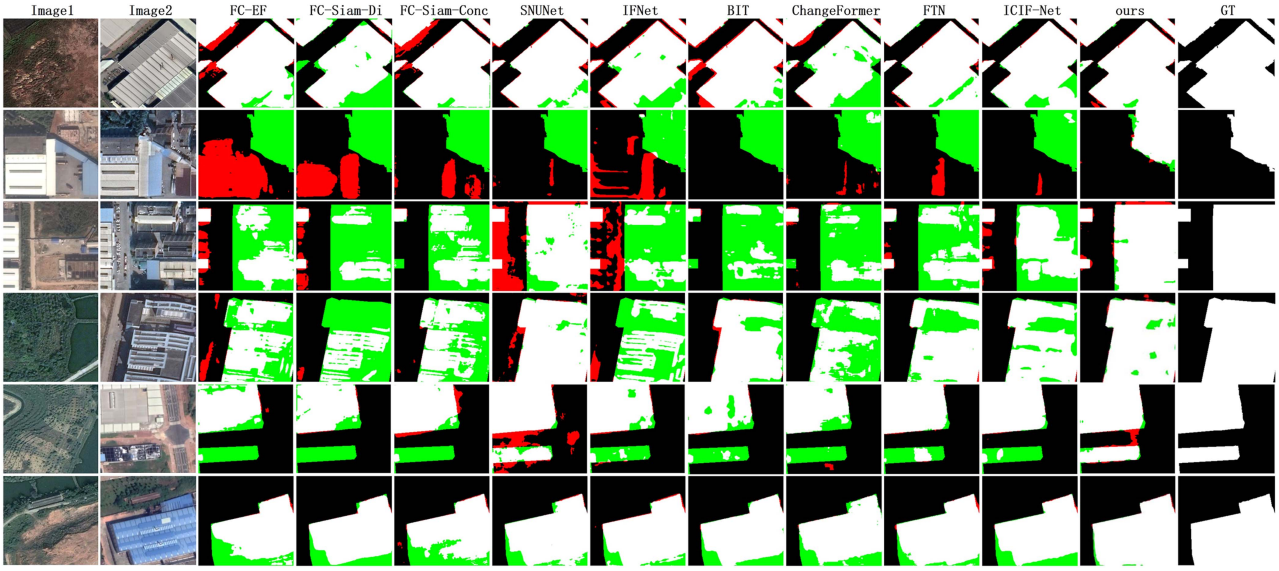

Fig. 16. Illustrates the comparative experiments on the GZ-CD dataset. Different colors are used for better visualization: white represents true positives, black represents true negatives, red signifies false positives, and green denotes false negatives.

all methods fail to detect completely changed targets. In the third set of samples, although the SNUNet method detects the general changed targets, it also exhibits numerous false detection errors. In contrast, only our method achieves results closer to the GT. In the fourth set of samples, due to differences in building colors, almost all methods fail to correctly identify the changed areas of the overall buildings. Although our method also fails to completely identify the changed areas, in terms of detection integrity, our detection results are closest to the GT. In the fifth and sixth sets of samples, almost all methods suffer from detection omission issues. However, comparatively, our method exhibits a high degree of completeness in detecting regions of change and is able to capture most of them relatively completely. Despite some problems with incomplete edge detection, which makes it difficult to accurately localize the edges of the target, the closeness between our results and the GT results is still the most significant. This not only proves the importance and rationality of our proposed architecture in the current research field, but further shows that our proposed architecture is the current mainstream and state-of-the-art method.

In Table XI, MVAFG achieves the best results in all aspects, such as F1 and IoU. Combined subjective and objective evaluations indicate that MVAFG outperforms mainstream and state-of-the-art methods on the GZ-CD dataset. However, it also reveals current issues with MVAFG, such as detecting blurry edges (see Fig. 16, second set) and incomplete handling of complex objects (see Fig. 16, fourth set). In the future, we will continue to optimize the network's feature extraction and complex object localization capabilities to address these issues. Overall, we present a comparative experiment from both subjective and objective perspectives. Through experiments on three datasets, we demonstrate the excellent performance of MVAFG in handling adversarial environmental interference and processing edge information, surpassing current mainstream and state-of-the-art methods. However, we also identified shortcomings of MVAFG, such as incomplete detection when handling blurry edges and complex small targets. In the future, we will focus on enhancing the ability to address these issues.

TABLE XI
INDICATOR RESULTS FOR EACH COMPARISON METHOD ON THE GZ-CD DATASET

Method	OA	F1	IoU	Precision	Recall
FC-EF	94.89	71.13	55.19	77.49	65.74
FC-Siam-Di	94.01	65.46	48.65	73.18	59.21
FC-Siam-Conc	95.68	76.24	61.61	80.59	72.34
IFNet	96.92	82.15	69.71	92.19	74.08
SNUNet	97.07	84.25	72.79	86.82	81.82
BIT	96.31	80.23	66.99	82.4	78.18
ChangeFormer	95.53	73.66	58.3	84.59	65.23
FTN	97.92	85.58	74.79	86.99	84.21
ICIF-Net	97.29	85.09	74.05	89.9	80.76
MVAFG	97.89	88.12	77.08	90.13	84.88

Red represents the best results, while blue represents the second best results.

G. Parameter Analysis

Complexity and parameter and training inference time analysis: In Table XII, MVAFG outperforms all existing methods. Although MVAFG has more parameters and floating-point operations than the CNN-based FC-Siam-Conc, it outperforms FC-Siam-Conc by a wide margin, compared with the hybrid CNN-transformer structure, MVAFG has a slightly larger model complexity and number of parameters due to its three-branch structure, but its performance index is still significantly higher than that of ICIF-Net, which suggests that there is still room for optimization of MVAFG, and our next step will be to further reduce the number of parameters in the model. Our next step will be to further reduce the number of parameters in the model. Although MVAFG's performance metrics on the WHU-CD dataset may slightly lag behind those of FTN, we have the advantage of having a much smaller number of parameters and computational

TABLE XII
COMPLEXITY AND PARAMETER NUMBER

Method	Complexity			LEVIR-CD		WHU-CD		GZ-CD	
	Params(M)	FLOPs(G)	Tt(s)	F1	IoU	F1	IoU	F1	IoU
FC-EF	1.35	3.57	34.80	84.02	72.45	77.64	63.45	71.13	55.19
FC-Siam-Di	1.35	4.72	39.21	87.17	77.26	79.97	66.63	65.46	48.65
FC-Siam-Conc	1.55	5.32	39.73	87.25	77.39	84.13	72.6	76.24	61.61
IFNet	50.71	82.35	121.24	89.32	80.7	83.41	71.53	82.15	69.71
SNUNet	12.03	54.88	405.37	86.32	75.93	88.34	79.11	84.25	72.79
BIT	3.55	10.59	147.54	90.34	82.39	87.47	77.73	80.23	66.99
ChangeFormer	41.03	202.87	373.43	90.4	82.48	86.88	76.81	73.66	58.3
FTN	168.53	45	421.67	91.01	83.51	92.61	85.45	85.58	74.79
ICIF-Net	25.83	25.27	270.5	91.18	83.85	90.77	83.09	85.09	74.05
MVAFG	29.87	36.94	285.5	91.82	84.95	92.35	85.49	88.12	77.08

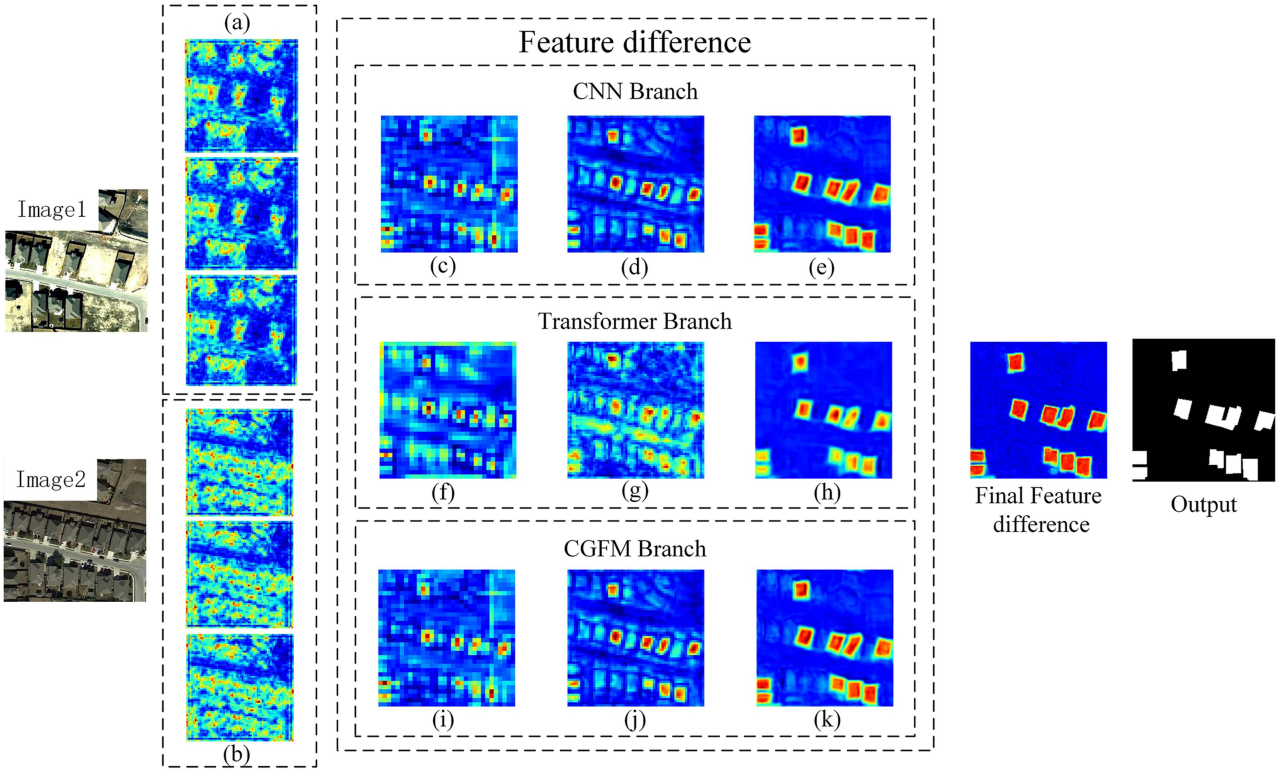

Fig. 17. Using images from the LEVIR-CD dataset as examples, this visualization demonstrates the generation and enhancement of feature maps through various network modules. (a) Feature maps generated by the CNN, transformer, and CGFM branches for Image1. (b) Feature maps generated by the CNN, transformer, and CGFM branches for Image2. (c) Feature maps enhanced by the GCIM module.(d) and (g) Enhanced feature maps obtained through the dual-branch fusion via the DCA module. (f) Feature maps enhanced by the AFRM module. (i) Fusion feature maps produced by the SCF module. (j) Feature maps decoded by the CGFD module. (e), (h), and (k) Feature maps obtained from pairwise fusion of features from the three branches.

load in MVAFG compared to FTN. Not only does MVAFG have much fewer parameters and computational load than the pure transformer-based Changeformer, but it also has much higher performance metrics. However, there is still room for further improvement of our method in terms of the number of parameters compared to methods such as BIT and ICIF-Net. We also compare the inference time on the training set of one epoch on the WHU-CD dataset, which is denoted by Tt(s) in Table XII. In general, the MVAFG model performs well in inference time, which ensures the accuracy of the model and the efficiency of model training. In the future, we will continue to optimize the network structure, reduce the number of parameters, and lower the time complexity while maintaining the performance of MVAFG to meet more specific needs and practical application scenarios.

H. Network Visualization

To better understand our model, we present visualized results of several key stages of the MVAFG model on the LEVIR-CD dataset, as shown in Fig. 17. Given a pair of bitemporal images, multilevel feature maps are initially captured by the CNN, transformer, and CGFE. These maps are then enhanced by the GCIM and AFRM modules for an enriched feature information representation. The CNN and transformer branches,

operating in tandem, facilitate feature interaction via ensemble cross-attention, guiding attention toward the regions of change and mitigating the influence of irrelevant factors. The CGF branch, through channel exchange and information interaction, garners a richer bitemporal information set, endowing features, which preexchange possess only unitemporal information, with comprehensive spatial characteristics essential for refining the change areas and precisely locating change objects existing in the same temporal phase. The visualization results demonstrate that the CGF branch significantly enhances the feature information extracted by the CNN and transformer branches, refining the areas of change. This underscores the rationale and effectiveness of the proposed tribranch network structure.

V. Conclusion

This article introduces a three-branch network architecture that combines CNN, transformer, and CGF branching to improve the model's capability in feature extraction and representation learning. Among them, the CGF branch aims to direct the feature extraction process to focus on task-relevant feature information, thus improving the model performance. The three-branch network structure combines the advantages of CNN in local feature extraction, transformer in global context modeling, and a channel feature guidance branch in focusing on task-related feature changes, which effectively improves the model's performance on complex tasks and demonstrates a high potential for practical use.

However, the model also has some potential shortcomings. In particular, the CNN and transformer branches employ pretrained backbone networks, which increase the number of parameters and computational burden of the model, leading to a relatively bulky model, especially in resource-constrained scenarios. To address this challenge, future research will focus on developing lightweight model architectures. Possible directions include exploring new network compression techniques, more efficient parameter sharing mechanisms, advanced model pruning and quantization strategies, and so on. For example, replacing traditional convolutional operations with grouped convolution to reduce the number of parameters and computational complexity, or using lightweight network architectures such as MobileNet or ShuffleNet as the feature extraction backbone network to alleviate computational and storage requirements. In addition to improvements in the model structure, future research should also explore the usability of this CD model in heterogeneous dual-temporal images, such as CD on dual-temporal synthetic aperture radar and optical images.

References

[1] 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.

[2] S. Ye, J. Rogan, Z. Zhu, and J. R. Eastman, "A near-real-time approach for monitoring forest disturbance using landsat time series: Stochastic continuous change detection," *Remote Sens. Environ.*, vol. 252, 2021, Art. no. 112167.

[3] L. Gueguen and R. Hamid, "Toward a generalizable image representation for large-scale change detection: Application to generic damage analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3378–3387, Jun. 2016.

[4] W. Shi, M. Zhang, H. Ke, X. Fang, Z. Zhan, and S. Chen, "Landslide recognition by deep convolutional neural network and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4654–4672, Jun. 2021.

[5] 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 .

[6] 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.

[7] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 228–239, 2022.

[8] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 101–119, 2022.

[9] 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. Top. Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3439–3451, Aug. 2016.

[10] 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.

[11] 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.

[12] D. K. Seo, Y. H. Kim, Y. D. Eo, M. H. Lee, and W. Y. Park, "Fusion of SAR and multispectral images using random forest regression for change detection," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 10, 2018, Art. no. 401.

[13] Z. Xie, M. Wang, Y. Han, and D. Yang, "Hierarchical decision tree for change detection using high resolution remote sensing images," in *Proc. 6th Int. Conf. Geo-Inform. Sustain. Ecosystem Soc.*, Handan, China, 2019, pp. 176–184.

[14] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel approach to unsupervised change detection based on a semisupervised svm and a similarity measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2070–2082, Jul. 2008.

[15] 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.

[16] X. Xu, Z. Yang, and J. Li, "AMCA: Attention-guided multi-scale context aggregation network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5908619.

[17] Z. Lv, X. Yang, X. Zhang, and J. A. Benediktsson, "Object-based sorted-histogram similarity measurement for detecting land cover change with VHR remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2504405.

[18] Z. Lv, P. Zhong, W. Wang, Z. You, J. A. Benediktsson, and C. Shi, "Novel piecewise distance based on adaptive region key-points extraction for LCCD with VHR remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607709.

[19] W. Zhang, L. Jiao, F. Liu, S. Yang, W. Song, and J. Liu, "Sparse feature clustering network for unsupervised SAR image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5226713.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[22] 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.*, Munich, Germany, 2015, pp. 234–241.

[23] 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.

[24] T. Lei et al., "Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 4507013.

[25] F. Zhou, C. Xu, R. Hang, R. Zhang, and Q. Liu, "Mining joint intra-and inter-image context for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4403712.
[26] R. Hang, S. Xu, P. Yuan, and Q. Liu, "AANet: An ambiguity-aware network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5612911.
[27] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 6000–6010, 2017.
[28] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5607514.
[29] 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.
[30] 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.
[31] H. Lin, R. Hang, S. Wang, and Q. Liu, "DiFormer: A difference transformer network for remote sensing change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 6003905.
[32] R. Guan et al., "Contrastive multi-view subspace clustering of hyperspectral images based on graph convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5510514.
[33] 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.
[34] 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.
[35] 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, pp. 1–16, 2021.
[36] 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, 2021, Art. no. 5604816.
[37] 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.
[38] 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. 6881–6890.
[39] F. Song, S. Zhang, T. Lei, Y. Song, and Z. Peng, "MSTDSNet-CD: Multiscale swin transformer and deeply supervised network for change detection of the fast-growing urban regions," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6508505.
[40] 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.
[41] 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.
[42] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. IEEE Conf. Learn. Representations*, 2015, pp. 1–15.
[44] 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.
[45] 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.
[46] T. Yan, Z. Wan, and P. Zhang, "Fully transformer network for change detection of remote sensing images," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1691–1708.