CGFINet: Cross-Scale Guided High-Order Feature Interaction Change Detection Network for Remote Sensing Image

Qichao Han, Xiyang Zhi, Jianming Hu, *Member, IEEE*, Yuanxin Huang, Wenbin Chen, Shikai Jiang, *Member, IEEE*, and Jinnan Gong

***Abstract*—Remote sensing image change detection is a valuable technology for analyzing the earth observation data. It has significant application value in resource monitoring, disaster assessment, and urban planning. However, current change detection methods have not fully explored the interrelationships between bitemporal data, and the extraction process of change information lacks prior guidance and constraints. Therefore, it is easy to produce missed detections and false alarms when facing complex backgrounds and variable objects in remote sensing images. To tackle such issues, we propose a cross-scale guided high-order feature interaction change detection network for dual temporal images. Specifically, a cross-scale guided dual encoder–decoder backbone is proposed to constrain the reconstruction process of change objectives, and guide geometric prior to optimize the representation of target structures. Next, an efficient high-order feature interaction module is designed, employing multilevel receptive fields to enhance the perception ability for multiscale features. Moreover, we construct a bitemporal feature alignment fusion module, which decouples and filters out the interference of background pseudo changes through interactive perception of spatial–temporal differences. Comprehensive experimental validation is undertaken on four representative change detection datasets (LEVIR-CD, WHU-CD, DSIFN-CD, and S2Looking). The findings demonstrate that the network demonstrated state-of-the-art performance.**

***Index Terms*—Change detection, cross-scale guidance, feature bidirectional interaction, remote sensing.**

I. Introduction

CHANGE detection is a process by which the state difference of a geographical area is determined through remote sensing revisiting and observing [1]. Change detection technology is widely used in disaster damage assessment [2], urban expansion [3], agricultural pest monitoring [4], and natural resource management [5]. Capturing change information of interested objects on the earth's surface is a highlight in remote sensing monitoring research. In recent decades, very high resolution (VHR) optical remote sensing satellites (such as Quickbird, WorldView-3, Gaofen series) have the advantages of space-based observation perspective and high revisit frequency. The increasingly rich multitemporal VHR optical remote sensing image data brings greater opportunities for change detection applications [6], [7].

Challenges still persist in change detection of VHR optical remote sensing images. First, the scenes in optical images are more complex, and background changes such as illumination intensity and shadows interfere with target edge extraction, as the content in red box of Fig. 1. Second, the discrepancy in

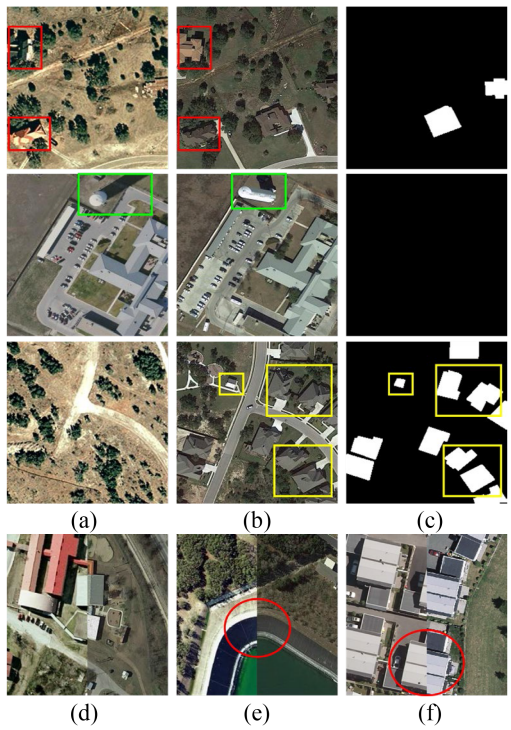

Fig. 1. Bitemporal VHR optical remote sensing images. The red box represents target texture and shadow change, the green box indicates the coupling effect of observation angle and shadow, the yellow box displays targets with diverse scales, and the red circle shows register deviation. (a) T1 images. (b) T2 images. (c) Ground truth. (d) Register accurate samples. (e) and (f) register deviation samples.

Manuscript received 18 June 2024; revised 18 July 2024; accepted 23 July 2024. Date of publication 29 July 2024; date of current version 26 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62305088 and in part by the China Postdoctoral Science Foundation under Grant 2023M740900. *(Corresponding authors: Xiyang Zhi; Jianming Hu.)*

The authors are with the National Key Laboratory of Laser Spatial Information, Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: hanqc_xyer@stu.hit.edu.cn; zhixiyang@hit.edu.cn; hujianming@hit.edu.cn; huangyxopt@163.com; 18b921010@stu.hit.edu.cn; jiangshikai@hit.edu.cn; gongjinnan@hit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3434966

observation angles of bi-temporal images inevitably results in a spatial displacement, and these spatial–temporal differences are prone to false changes, as illustrated in the red circle in Fig. 1. Moreover, the above two problems are often coupled with each other, further increasing the spatial-temporal differences of the background and the instability of changing information, as shown in green boxes of Fig. 1. Finally, characteristics of features in VHR remote sensing images, such as scale, structure, texture, and distribution, are complex and diverse, as illustrated in yellow boxes of Fig. 1. This complexity increases the difficulty of extracting robust features.

Deep learning excels in computer vision, and has an advantage in extracting fine feature details and complex texture information from VHR remote sensing images [8], [9]. This is due to its ability to mine robust deep features, and has attracted increasing attention in change detection [10]. These models fall into two broad categories: single-stream approach and double-stream structure. The former is cascaded or differentially processed on bi-temporal images as inputs into semantic segmentation model to achieve end-to-end change detection [11], [12]. However, images integration operations in the early processing inevitably sacrifices original semantic information, and it is often unable to cope with the interference of space–time displacement.

Currently, the double-stream structure has become the mainstream of bi-temporal change detection, which inputs the bitemporal images as independent data into two feature extraction branches [11], [14], [15]. In this structure, the feature fusion module is needed to mine change information from two feature branches [10], [13], [16]. The double-stream network prevents the mixing of the original images and fuses the bitemporal features of ground objects after the encoder to capture more accurate changing information.

Despite the growing popularity of the double-stream network approaches, there remains a need for effective solutions that address the aforementioned challenges in a comprehensive manner. We contend that its crux lies in the following three aspects.

1) Improving the perception ability for variable scales targets and establishing a robust feature representation model.
2) Enhancing the guidance of geometric prior on the process of changing semantic reconstruction.
3) Constraining the bitemporal feature fusion to suppress the interference of registration error and eliminate the false and retain the truth.

Different from the previous work, we comprehensively considered the above three aspects and proposed cross-scale guided high-order feature interaction change detection network (CGFINet). First, the high-order feature interaction module employs a joint action of translation and interactive operation to generate hierarchical multiscale receptive fields with less computation, thereby enhancing the perception of variable targets. This approach differs from redundant convolution stacking operations or transformer modules with significant computation in previous work. Second, we consider that the plain skip connection is prone to introducing redundant and interfering information. To address this, we propose cross-scale guided enhanced decoders, which employ refined geometric priors to constrain the change semantic reconstruction, thus preventing the loss of change information in complex remote sensing scenes. Third, based on the prior knowledge of the symmetry of binary change detection, we propose a symmetrical interaction strategy of bi-temporal features fusion. This module can symmetrically enhance the characteristics of changing areas and suppress the background features of spatial–temporal dislocation. The main contributions of our article are summarized as follows.

1) A novel change detection network is proposed. It uses cross-scale guided enhancement decoder (CGED) to improve the refined reconstruction of changed semantic features with prior geometric information, capturing finer change regions.
2) A high-order feature interaction module (HFIM) is introduced to efficiently obtain hierarchical receptive fields, which uses a concise group shifting and interaction operation to capture and fuse multiscale features.
3) A bitemporal feature alignment and fusion module (BAFM) is adopted, which suppresses spatial–temporal dislocation based on optical flow. It decouples the background pseudo change interference and enhances changed feature adaptively.

The rest of this article is summarized as follows: Section II provides an overview of current dual stream change detection methods and feature fusion methods. The problem statement and our solutions are also provided in this section. The specific process of our method is introduced in Section III. Section IV presents experiments and discussions. Section V provides a summary of our entire work.

II. Related Works

A. Double-Stream Change Detection Methods

Dual stream networks typically employ a Siamese neural network architecture to process bitemporal images [17], [18]. Then extract changing information by merging the features of two branches. Dual stream networks can be mainly divided into two structures : dual encoder single decoder (DESD) network and dual encoder–decoder (DED) network.

The DESD network fuses bitemporal features before decoding operations. Zhang et al. [16] proposed a Siamese architecture with skip connection for extracting bi-temporal image features. Fang et al. [19] combined the advantages of dual encoder structure and UNet++ semantic segmentation model to obtain change information. Chen et al. [20] and Zhang et al. [21] used implicit neural representation to design the change decoder after the twin encoder, and proposed a new idea to alleviate the spatial resolution difference and alignment deviation of bitemporal features. Recently, Zhao et al. [22] introduced Mamba based encoder in the framework of DESD, which has advantages in capturing context in large-scale remote sensing images. Zhang et al. [23] proposed to fuse the global and local guidance of the dual encoder features, to alleviate Mamba's lack of local cues when dealing with change detection tasks. Chen et al. [24] introduced region similarity before the decoder for bitemporal feature fusion to reduce the impact of spatial offset on the decoding process. The DED network often includes two encoding–decoding branches with shared weights, and feature

fusion operation occurs during the decoding stage. Chen et al. [13] constructed a DED structure for binary change detection, which enhances change information, reduces background interference, and utilizes self-supervised learning strategies to improve change detection performance. Liang et al. [25] used graph convolution to perceive global contextual information and enhance the guidance of low-level features in the decoding and reconstruction process.

DESD directly fuses the bitemporal encoder features. However, general encoder feature includes invalid background information and noise interference, causing redundancy and deviation in change detection. Dual-decoder is additionally designed in DED which has potential in reconstructing change characteristics and suppressing background interference. However, due to the change of imaging conditions such as season and illumination in different time phases, pseudo change features are inevitable in bitemporal images. This brings challenges to the current DED works to accurately reconstruct bitemporal features.

Varying from the aforementioned works, we consider geometric edge, context, and semantic information simultaneously to provide enhanced guidance on high-level semantic reconstruction. Concurrently, we employ high-order feature interaction to generate a hierarchical multiscale receptive field, which enhances the perception of multiscale features with reduced computational complexity.

B. Bitemporal Feature Fusion Methods

In double-stream approach, bitemporal fusion module is often used as a direct means to accurately extract the changing feature. In order to fully exploit the potential benefits of different fusion methods, Chen et al. [13] proposed parallel fusion methods such as summation and difference. On this basis, Chen et al. [26] integrated channel attention and spatial attention to better perceive global contextual information. Jiang et al. [27] and Zhao et al. [28] established attention module to achieve bitemporal feature interaction. The current approach focuses more on change information enhancement and background region suppression [10], [16], [29], ignoring images registration error when revisiting the same area. However, the VHR remote sensing images with extremely fine semantic details are more sensitive to registration errors, which disturbs the effectiveness of the fusion method mentioned above. The spatial–temporal dislocation caused by side-looking problems, which is likely to be amplified into pseudo changes. Song et al. [30] analyzed the pixel mismatch problem that still exists under regional coregistration, and reduced the impact of registration bias by downsampling the original high-resolution image and discarding shallow differential feature layers. But it inevitably leads to the loss of feature information. Liang et al. [25] used feature interaction strategy to achieve spatial information exchange between dual temporal features. However, due to registration bias, feature interaction during the encoding stage often produces pseudo changing semantics. Considering the third aspect that affects the performance of change detection network mentioned above, we construct the dual temporal feature alignment and fusion module.

In this article, we creatively integrate the idea of spatial-temporal registration into the fusion strategy and propose a symmetrical interaction strategy of bitemporal features. This module can symmetrically enhance the characteristics of changing regions and suppress the background features of spatial-temporal dislocation.

III. Proposed Method

A. Method Overview

The complete process architecture of our approached CGFINet is displayed in Fig. 2. A dual encoder–decoder backbone is designed for CGFINet, including dual-encoders, multilevel feature fusion module, dual-decoders, and change decoders. Dual-encoders with shared weights extracts geometric texture information and scene semantics from bitemporal images. Dual-decoders with a series CGEDs will fuse geometric features with the deep high-level semantic features, which is detailed in Section III-B and Fig. 3. The HFIM based on hierarchical receptive fields mines and fuses multiscale target features, which is detailed in Section III-C and shown in Fig. 4. In change information decoders, a series of BAFMs is applied to realize the alignment and fusion of bitemporal features, which is described in detail in Section III-D and shown in Fig. 5. The change decoders output the enhanced change features and the change head produce the prediction change map.

B. Cross-Scale Guided DED Backbone

Dual-encoders based on Siamese network structure are commonly operated to model the original images into the same feature space. The encoder features contain rich scene semantic information, and then the feature map is input into decoders to reconstruct changing area. We analyze the advantages and disadvantages of DESD structure and DED structure in Section II-A. Considering this issue, we choose DED network as the backbone. It uses dual decoders with shared weights, thus avoiding the confusion of dual temporal information in the process of semantic reconstruction. This structural design is not only conducive to improving the temporal and spatial accuracy of change features, but also conducive to further integrating multilevel feature information.

Dual-encoder branches with shared weights are composed of four encoder blocks, which is shown in Fig. 2. Encoder blocks down sample the size of the feature map as (H_0/2, W_0/2), (H_0/4, W_0/4), (H_0/8, W_0/8), and (H_0/16, W_0/16), and increase the channel number by 32/64/128/256, where H_0 and W_0 respectively indicate the height and width of our input images.

Each encoder block uses SE-ResNet structure [31], and it can be formulated as follows:

$$F_{ei} = f_{\text{MAP}}\left(\text{BaseConv}\left(F_{e(i-1)}\right)\right) + f_{\text{MAP}}\left(f_{\text{SE}}\left(\text{BaseConv}\left(\text{Baseconv}\left(F_{e(i-1)}\right)\right)\right)\right) \quad (1)$$

where $\text{BaseConv}(\cdot)$ refers to the combination of base convolution layer, batch normalization and rectified linear unit.

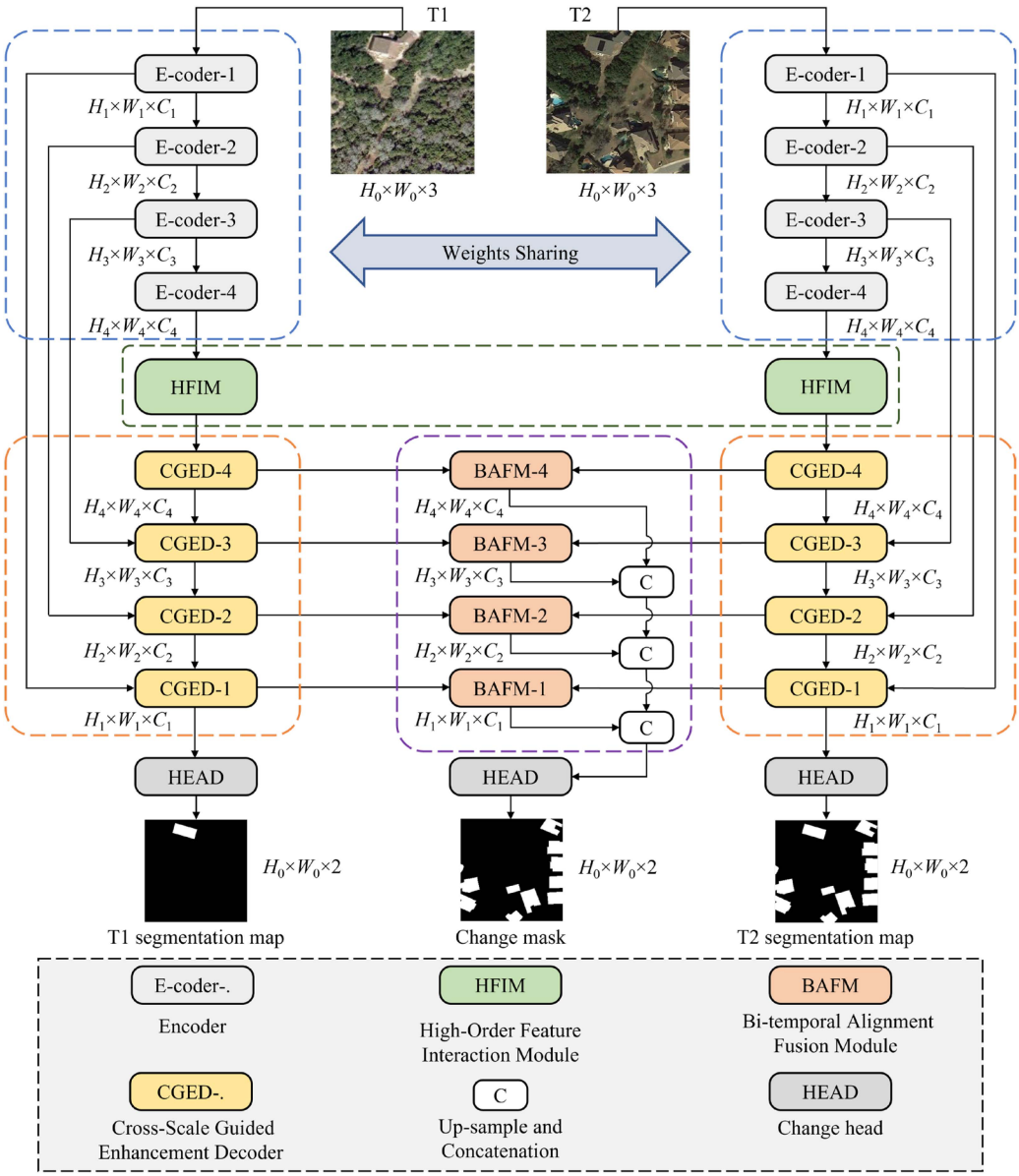

Fig. 2. Overall framework of the proposed CGFINet.

$f_{\mathrm{MAP}}(\cdot)$represents max pooling, and $f_{\mathrm{SE}}(\cdot)$ denotes the SE block, as shown in Fig. 3(b). $F_{ei} \in \mathbb{R}^{H_i\times W_i\times C_i}(i=1,2,3,4)$ refers to the ith encoder feature, and $H_i = H_0/2^i$, $W_i = W_0/2^i$, $C_i = 2^{i+4}$respectively represent its height, width, and channels number.

The skip connection in UNet structure introduces rich structure into the decoder features, which is superior in medical image semantic segmentation field. However, the scenes in VHR remote sensing images are more complex and diverse, which leads to the traditional skip connection mode is easy to cause occlusion and interference between the region of multiscale targets. In light of the aforementioned characteristics of remote sensing images, we introduced the CGED to obtain the edge non aliasing multiscale target spatial domain enhancement feature, as shown in Fig. 3.

CGED uses spatial features from encoders guiding the change semantic reconstruction process. The introduction of geometric feature enhancement mechanism makes skip connections more efficient and accurate, simultaneously suppress occlusion interference in the scene. The input of CGED is a low-level encoder feature map F_{ei} and a deep decoder feature map $F_{d(i+1)} \in \mathbb{R}^{H_{i+1}\times W_{i+1}\times C_{i+1}}$, and its output is a deep decoding feature map $F_{di} \in \mathbb{R}^{H_i\times W_i\times C_i}$, Specifically, in order to fully capture the geometric features of the changing region, we employ two parallel branches. First, we utilize the underlying features to obtain local fine geometric features. Second, we guide the adaptive enhancement of the global feature channels in the changing scene. We introduce angular differences pixel difference convolution (APDC) [32] to mine the geometric edge information of local regions, and integrate it into decoding features. As shown

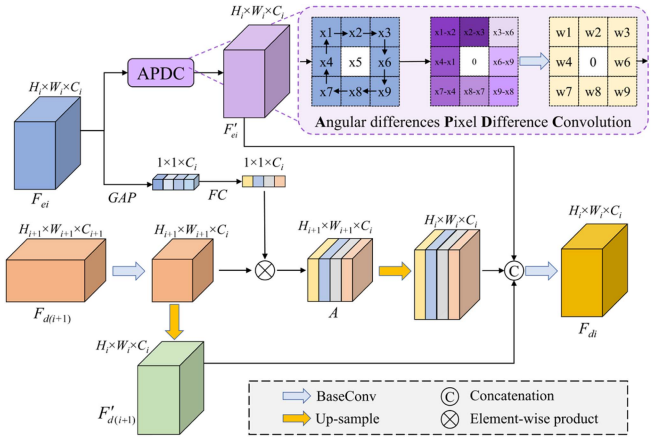

Fig. 3. Structure of cross-scale guided enhancement decoder.

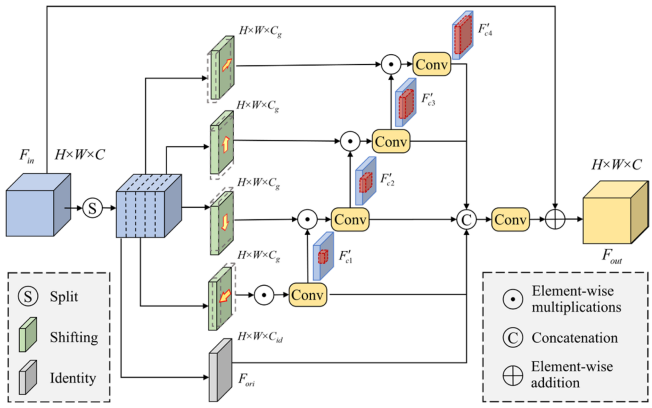

Fig. 4. Structure of hierarchical receptive field module.

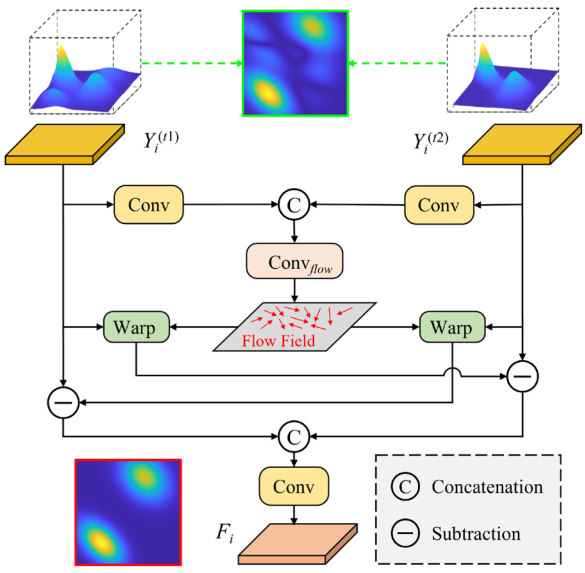

Fig. 5. Structure of bi-temporal alignment fusion module. (Green box indicates that directly fusing misaligned dual temporal features can easily generate false alarms. Red box indicates that the BAFM can filter out the false detection caused by registration deviation.).

in Fig. 3, the mathematical expression of APDC is as follows:

$$\begin{aligned}\mathcal{F}_{\text{APDC}}(x_i) &= w_1 \cdot (x_1 - x_2) + w_2 \cdot (x_2 - x_3) + \cdots \\ &= (w_1 - w_4) \cdot x_1 + (w_2 - w_1) \cdot x_2 + \cdots \\ &= \hat{w}_1 \cdot x_1 + \hat{w}_2 \cdot x_2 + \cdots = \sum \hat{w}_i \cdot x_i\end{aligned} \tag{2}$$

where $x_i$represents pixels to be processed, and $w_i$and $\hat{w}_i$ refers to convolution kernel weight and corresponding weight difference, respectively.

In another branch, we use attention mechanism to obtain feature channel weights. The low-level feature F_{ei} is global average pooled and subsequently compressed to a vector of $1 \times 1 \times C_i$ as guide feature. We use its value as weight to characterize the importance of each feature channel. Then, we compress the number of channels of $F_{d(i+1)}$ into C_i by convolution operation and multiply it with the guide feature vector pixel by pixel. Then, we use residual connected structure can preserve the spatial domain details with semantic information. The deep features $F_{d(i+1)}$ are compressed in channels and upsampled in spatial, it is concatenated with the upsampled feature map A and edge feature F'_{ei} along the channel dimension. To realize the adaptive fusion expression, we compress the channel number of concatenated feature to C_i using a BaseConv with kernel size of 1×1. The aforementioned process is mathematically represented by the following formula:

$$\begin{cases} A = \left(f_{\text{FC}}\left(f_{\text{GAP}}\left(F_{ei}\right)\right)\right) \otimes \text{BaseConv}\left(F_{d(i+1)}\right) \\ F'_{d(i+1)} = \text{Up}\left(\text{BaseConv}\left(F_{d(i+1)}\right)\right) \\ F_{di} = \text{BaseConv}\left(\text{Concat}\left(F'_{ei}, \text{Up}\left(A\right), F'_{d(i+1)}\right)\right) \\ i = 1, 2, 3 \end{cases} \tag{3}$$

where $f_{\text{FC}}(\cdot)$ is full connection layer, and $f_{\text{GAP}}(\cdot)$ refers to global average pooling. $Up(\cdot)$ represents upsampling. In addition, the $\text{Concat}(\cdot)$ denotes to the concatenating operation along channel dimension and $\otimes$ represents the element-wise multiplication.

C. High-Order Feature Interaction Module

The features in remote sensing images have large scale variations. Therefore, it is essential to extract multiscale hierarchical information using large receptive fields effectively and fuse them efficiently. Rao et al. [34] considered that the superior performance of ViT [33] compared with CNN was due to the high-order spatial interaction modeling ability of ViT [33].

Inspired by Rubik's cube [35], we introduced an HFIM, which uses high-order channel interactions to construct hierarchical receptive fields for fusing multiscale features of targets. The module is added after the encoder. Through the operation of shift and channel interaction, the high-order channel interaction of feature layer is realized in a simple and effective way to enhance the multilevel feature mapping.

The structure of HFIM can be seen in Fig. 4. To be specifically, we divide the encoder feature F_{in} into five sections along the channel axis using a splitting operation, one of which remains unchanged, and the other four are shifted in the left, right, up, and down directions, respectively. In order to keep the input feature spatial structure unchanged, zeros are filled to the vacant position and out-of-focus pixels are discarded after shifting

operation. The shifted feature map can be expressed as (4) shown at the bottom of the this page, where C_{id} denotes the number of channels that remain unchanged, and C_g refers to the number of channels that perform shifting operation, and $C_{id} + 4 \times C_g = C$. We set the number of shifted pixels to 1 by default and analyze its advantages in Section IV-E. Through zero FLOP and zero parameter shift operations, we efficiently split the input characteristic graph X into $F_{\text{ori}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C_{id}}}$ and $\{F_{c1}, F_{c2}, F_{c3}, F_{c4}\} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C_g}}$.

After the grouping and panning operations, we alternate with convolution layers and element-wise multiplication to integrate the information in the four shifting operation groups. Specifically, multilevel channel interaction is realized by element level multiplication with the following shifting groups in turn. Equations (5) and (6) represent first-order channel interaction, and (7) and (8) represent high-order channel interaction

$$F'_{c1} = \mathrm{Conv}_{1\times1}\left(F_{c1}\right) \tag{5}$$

$$F'_{c2} = \mathrm{Conv}_{1\times1}\left(F'_{c1} \otimes F_{c2}\right) \tag{6}$$

$$F'_{c3} = \mathrm{Conv}_{1\times1}\left(F'_{c2} \otimes F_{c3}\right) \tag{7}$$

$$F'_{c4} = \mathrm{Conv}_{1\times1}\left(F'_{c3} \otimes F_{c4}\right) \tag{8}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ refers to convolution operation with 1×1 kernel size. The multiorder channel interaction features are concatenated in the channel axis and subsequently fused by convolutional operation. Finally, the input feature is added to obtain the result of Rubik's cube convolution, which is as follows:

$$F_{\text{out}} = \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left[F_{\text{ori}}, F'_{c1}, F'_{c2}, F'_{c3}, F'_{c4}\right]\right) + F_{\text{in}}. \tag{9}$$

HFIM realizes multilevel channel interaction through simple convolution layers with small kernel and element-wise multiplication. The coupling effect of splitting and shifting operations and multilevel channel interaction enables us to obtain hierarchical multiscale receptive fields. It realizes the fusion of multiscale features and makes the features input into the decoder more discriminative against multiscale targets.

D. Bitemporal Alignment Fusion Module

Bitemporal feature fusion is an important step to obtain change features in the change detection network. Common feature fusion methods include direct fusion (addition, subtraction, or splicing), convolution enhancement fusion and fusion based on attention mechanism. They focus on enhancing the bitemporal characteristics and reducing the interference of background noise. However, they do not address the impact of spatial registration errors on abstracting changing regions. The complexity of remote sensing image scene inevitably leads to dislocation or side view problems, despite the input bi\temporal images have been preprocessed by image registration.

In order to avoid the impact of matching errors or side view problems in the process of bitemporal feature fusion, we introduce a BAFM based on optical flow method, as shown in Fig. 5. The feature maps $F_{di}^{(t_1)} \in \mathbb{R}^{H_i\times W_i\times C_i}$ and $F_{di}^{(t_2)} \in \mathbb{R}^{H_i\times W_i\times C_i}$ output by the dual decoders are spliced along the channel dimension, and then input to a semantic flow field extraction subnetwork. The convolution unit contains two convolution layers, and its kernel size is 5×5. The subnetwork output a semantic flow field information $\Delta f_i \in \mathbb{R}^{H_i\times W_i\times 4}$, which is expressed by the following formula:

$$\Delta f_i = \mathrm{Conv}_{\text{flow}}\left(\mathrm{Concat}\left(F_{di}^{(t_1)}, F_{di}^{(t_2)}\right)\right) \tag{10}$$

where $\mathrm{Conv}_{\text{flow}}(\cdot)$represents the semantic flow field extraction subnetwork. After the flow field information is calculated, Δf_i is divided evenly along the channel dimension to get $\Delta f_i^{(t_1)} \in \mathbb{R}^{H_i\times W_i\times 2}$ and $\Delta f_i^{(t_2)} \in \mathbb{R}^{H_i\times W_i\times 2}$. Then, the semantic flow field information is used to correct the spatial position of the bi-temporal features respectively, and two corrected feature maps are obtained by bilinear interpolation. Feature alignment based on flow field is aimed to avoid the influence of matching error or side view problem, and improve the spatial consistency representation ability of bitemporal features, which helps to extract more accurate change information, and eliminate the pseudo changes caused by spatial offset. Subsequently, we calculate Euclidean distance between each feature map and the other temporal corrected feature map respectively, and splice the two distance features along channel dimension. Finally, a forward propagation convolution layer is designed to extract the changing feature. Mathematically, the output fusion features can be expressed as follows:

$$\begin{aligned} F_i = \mathrm{Conv}_{1\times1}\Big(\mathrm{Concat}\Big(&\left[\mathrm{wrap}\left(F_{di}^{(t_1)}, \Delta f_i^{(t_1)}\right) - F_{di}^{(t_2)}\right], \\ &\times \left[\mathrm{wrap}\left(F_{di}^{(t_2)}, \Delta f_i^{(t_2)}\right) - F_{di}^{(t_1)}\right]\Big)\Big) \end{aligned} \tag{11}$$

where $\mathrm{wrap}(\cdot)$ refers to the calculation operation of correction feature map based on bilinear interpolation. The forward propagation convolution layer halves the feature channel dimension, and keep it consistent with the input single temporal feature to achieve effective feature fusion.

E. Loss Function

The loss function we designed includes two parts: binary change loss L_c and semantic segmentation auxiliary loss L_s. As shown in Fig. 6. Among them, L_c aims to constrain the deviation between our predicted change map and the truth change

$$\begin{aligned} F_{ori} &= F\left[0:H, 0:W, 0:C_{id}\right] \leftarrow F_{in}\left[0:H, 0:W, 0:C_{id}\right] \\ F_{c1} &= F\left[0:H, 1:W, C_{id}:C_{id}+C_g\right] \leftarrow F_{in}\left[0:H, 0:W-1, C_{id}:C_{id}+C_g\right] \\ F_{c2} &= F\left[0:H, 0:W-1, C_{id}+C_g:C_{id}+2C_g\right] \leftarrow F_{in}\left[0:H, 1:W, C_{id}+C_g:C_{id}+2C_g\right] \\ F_{c3} &= F\left[0:H-1, 0:W, C_{id}+2C_g:C_{id}+3C_g\right] \leftarrow F_{in}\left[1:H, 0:W, C_{id}+2C_g:C_{id}+3C_g\right] \\ F_{c4} &= F\left[1:H, 0:W, C_{id}+3C_g:C_{id}+4C_g\right] \leftarrow F_{in}\left[0:H-1, 0:W, C_{id}+3C_g:C_{id}+4C_g\right] \end{aligned} \tag{4}$$

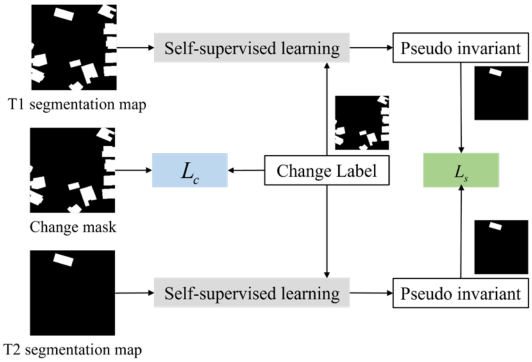

Fig. 6. Schematic diagram of constraint learning in CGFINet.

label. Inspired by the research of Chen et al. [13], the semantic segmentation auxiliary loss L_s is used to constrain the prior difference between the outputs of two semantic segmentation branches. That is, the semantic information of invariant regions should be as similar as possible, while the semantic difference of changing regions should be as large as possible. It can be expressed by the mathematical formula as follows:

$$\begin{cases} L_s = F_{\text{loss}}\left(\left\{S^{(t_1)} \cap Y_u\right\},\ \left\{S^{(t_2)} \cap Y_u\right\}\right) \\ Y_u = \{(n,m)|y_{n,m}=0\} \end{cases} \tag{12}$$

where Y_u represents the invariant pixel set in the ground change truth label, and $S^{(t_1)}$ and $S^{(t_2)}$ refers to the output result graph of two semantic segmentation branches, respectively., and $\cap$ denotes the intersection of logical operations. We use the addition of binary cross entropy loss l_{BCE} and dice number loss l_{dice} to calculate the loss function of each part (L_c and L_s), which is described as follows:

$$\begin{cases} l_{\text{BCE}} = \sum_{h=1,w=1}^{H\times W} \frac{(y_{h,w}\log x_{h,w} + (1-y_{h,w})\log(1-x_{h,w}))}{H\times W} \\ l_{\text{dice}} = 1 - \frac{2\times|\text{CM}\cap I_{\text{truth}}|}{|\text{CM}|+|I_{\text{truth}}|} \end{cases} \tag{13}$$

where H and W respectively indicate the height and width of the truth change label, h and w denote to the pixels position in image, CM and I_{truth} refer to the prediction change map and truth change label, $x_{h,w}$ and $y_{h,w}$ represent the value of pixels in the prediction change image and the truth change image, respectively, and $y_{h,w}=1$ ($y_{h,w}=0$) represents the changing (unchanged) pixels in the ground change map.

Our final loss function is defined as follows:

$$L = L_c + \alpha L_s \tag{14}$$

where α denotes the weight parameter, which is set as 0.2 in subsequent experiments.

IV. Experimental Results and Analysis

A. Datasets and Evaluation Metrics

In this section, four universal datasets are selected for fully evaluating effectiveness of our proposed method. During our experiment, we divided images into patches of 256 × 256 pixels without any overlap, and specific details are as follows.

1) *LEVIR-CD* [36] is produced for building change detection using Google Earth as the data source, with spatial resolution of 0.5 m. The images comprise various types of buildings and include the effects of seasonal change and solar illumination changes. We divide it into three groups according to the volume of 7120, 1024, and 2048, which are respectively used as training set, verification set, and test set.
2) *WHU-CD* [37] is a bitemporal change detection dataset, and records the changes of buildings reconstructed after the earthquake. The dataset includes aerial images of the region from 2012 and 2016, containing over 10 000 buildings. The dataset is divided into three groups according to the volume of 6096, 760, and 760, which are respectively used as training set, verification set and test set.
3) *DSIFN-CD* [16] is a dual temporal remote sensing image dataset collected from Google Earth with spatial resolution of 2 m. It contains diverse ground buildings imaged by different sensors under different seasonal conditions. The dataset is divided into three groups according to the volume of 14 389, 13 590, and 192, which are respectively used as train set, validation set and test set.
4) *S2Looking* [38] is a building change detection dataset for urban and rural scenes, which provides side view remote sensing image pairs. The presence of rural scenes makes the phenomenon of farmland or vegetation changes caused by seasonal changes more common, and improves the challenge of building change detection. The image has a spatial resolution of 0.5–0.8 m. We divide it according to the volume of 56 000, 8000, and 16 000, which are respectively used as train set, verification set, and test set.

Our evaluation metrics used for experiments are precision (Pre), recall (Rec), intersection over union (IoU), and overall accuracy (OA). These four metrics are commonly used for evaluating the behavior of change detection method. They are calculated as follows:

$$\text{Pre} = \text{TP}/(\text{TP}+\text{FP}) \tag{15}$$

$$\text{Rec} = \text{TP}/(\text{TP}+\text{FN}) \tag{16}$$

$$\text{IoU} = \text{TP}/(\text{TP}+\text{FP}+\text{FN}) \tag{17}$$

$$\text{OA} = (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FN}+\text{FP}) \tag{18}$$

where TP, TN, FP, FN are calculated by confusion matrix, and indicate true positive, true negative, false positive, and false negative, respectively. We use F1-score (F1) as the main evaluation index, and it is calculated using the following formula:

$$\text{F1} = 2\times\text{Pre}\times\text{Rec}/(\text{Pre}+\text{Rec}). \tag{19}$$

B. Comparison Algorithms and Implementation Details

The selected several state-of-the-art change detection methods include three single-stream methods: FC-EF [39], DDCNN [12], and CLNet [40], seven dual-stream methods: DTCDSCN [41], STANet [36], SNUNet [17], SILICD [20], HANet [44], CGNet [45], and C2FNet [46]. To demonstrate the efficiency

TABLE I
DISTINCTIONS BETWEEN CGFINET AND OTHER COMPARISON ALGORITHMS ON LEVIR-CD AND WHU-CD

Method	LEVIR-CD					WHU-CD				
	Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA
FC-EF [39]	79.34	76.71	78.01	63.94	97.80	78.86	78.64	78.75	64.94	93.03
DDCNN [12]	89.01	87.39	88.19	78.88	99.02	87.74	80.48	83.95	72.34	98.70
CLNet [40]	89.88	88.59	89.23	80.56	99.02	91.93	85.91	88.82	79.89	98.94
DTCDSCN [41]	90.54	85.20	87.78	78.29	99.01	63.92	82.30	71.95	56.19	-
STANet [36]	83.81	91.00	87.26	77.40	98.66	70.47	89.39	78.81	65.03	98.12
SNUNet [17]	89.18	87.17	88.16	78.83	99.02	**93.32**	87.89	**90.52**	**82.68**	**99.28**
BiT [42]	89.24	89.37	89.31	80.68	98.92	86.64	81.48	83.98	72.39	98.75
ChangeFormer [43]	92.05	88.80	**90.40**	**82.48**	**99.04**	92.70	82.28	87.18	77.27	99.14
SILICD [20]	90.55	86.30	88.38	79.18	98.86	89.20	91.87	90.51	82.67	99.25
HANet [44]	91.21	89.36	90.28	82.27	99.02	88.30	88.01	88.16	78.82	99.16
VcT [27]	92.57	87.65	90.04	81.89	99.01	89.39	89.77	89.58	81.12	99.18
Ours	**92.06**	**89.76**	90.90	83.31	99.08	93.49	**90.12**	91.77	84.79	99.37

Note: The best results are in red, and the second-best are in bold. Scores in this table are written in percentage (%).

TABLE II
DISTINCTIONS BETWEEN CGFINET AND OTHER COMPARISON ALGORITHMS ON DSIFN-CD AND S2LOOKING

Method	DSIFN-CD					S2Looking				
	Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA
FC-EF [39]	54.33	69.63	61.03	43.92	84.89	54.73	33.51	37.62	29.33	95.13
DDCNN [12]	62.88	63.90	63.39	46.40	74.02	63.49	47.69	54.47	37.43	98.92
CLNet [40]	66.19	66.81	66.50	49.81	76.37	64.96	48.24	55.36	38.28	98.94
DTCDSCN [41]	53.87	77.99	63.72	46.76	84.91	62.79	52.89	57.42	40.27	98.99
STANet [36]	67.71	61.68	64.56	47.66	88.49	54.46	40.96	46.75	30.51	98.73
SNUNet [17]	73.49	62.94	67.80	51.29	89.84	68.37	53.24	55.62	44.91	96.17
BiT [42]	68.36	70.18	69.26	52.97	89.41	72.29	53.99	61.81	44.73	99.09
ChangeFormer [43]	76.81	69.36	72.90	57.35	91.24	75.91	44.99	56.50	39.37	99.16
SILICD [20]	77.08	69.74	73.23	57.76	82.18	67.16	55.17	60.58	43.45	99.13
HANet [44]	56.52	70.33	62.67	45.64	85.73	61.38	55.94	58.54	41.38	99.04
VcT [27]	83.91	66.47	**74.18**	**58.95**	**92.14**	72.36	53.28	61.37	44.27	99.19
CGNet [45]	47.75	**81.38**	60.19	43.05	81.71	70.18	59.38	64.33	47.41	99.20
C2FNet [46]	56.41	72.15	63.32	46.32	85.79	72.55	**56.37**	**63.38**	**46.40**	**99.21**
Ours	**78.45**	81.47	79.93	66.57	93.05	**73.90**	53.71	62.21	45.15	99.21

Note: The best results are in red, and the second-best are in bold. Scores in this table are written in percentage (%).

of our method, we also select three transformer-based methods: BiT [42], ChangeFormer [43], and VcT [27].

The experiments are trained using a single NVIDIA GeForce RTX 3090 GPU and implemented using the PyTorch toolkit. We apply several data augmentation, including flip, rescale, crop, and Gaussian blur. We use the AdamW optimizer and a linear rate policy. We train these models in 200 epochs on four datasets with batch size as 8. We install the initialization learning rate as 0.001, and beta value as (0.9, 0.999) and the weight attenuation as 0.0001. The model validating on the validation set occurs after each epoch training, and is tested on the test set using the best model of all epoch training.

C. Comparison Experiments

The official code and default parameter settings of the comparison methods are employed in comparison experiments. The comparison results are presented in Tables I and II. More specific results and analyses of experiments on LEVIR-CD, WHU-CD, DSIFN-CD, and S2Looking datasets are given below, respectively.

1) Comparisons on LEVIR-CD Dataset: As demonstrated in Table I, our approach significantly behave better than the comparison method on LEVIR-CD. Specifically, our CGFINet achieves 92.06%, 89.76%, 90.90%, 83.31%, 99.08% on the Pre, Rec, F1, IoU, and OA, and yields optimal F1, IoU, and OA. In these methods, STANet [36] employs a pyramid spatial–temporal attention mechanism to discover the multiscale information and to establish intricate global spatial–temporal associations. It achieves the optimal Rec in our experiments. ChangeFormer [43] makes full use of the layered transformer encoders to obtain a larger receptive field and a stronger long-range semantic mining ability, achieving suboptimal F1, IoU, and OA. VcT [27] considers the invariance of the background region, and effectively reduces false alarms by mining the global dependencies between reliable tokens. It achieves the optimal Pre in our experiments. However, the above methods do not optimally balance Pre and Rec, which may be due to the lack of refinement constraints on the semantic feature decoding process. Different from these methods, CGFINet improves the refined representation of change features by mining the guidance and enhancement ability of low-level geometric feature in the process of high-level semantic reconstruction. It uses BAFM to suppress the pseudo detection caused by spatial-temporal deviation, to obtain the optimal F1 value.

In order to intuitively and qualitatively understand the behavior of our proposed method, we select eight methods with the best F1 values, and show their visualization results in Fig. 7. The selected scenes include densely distributed buildings,

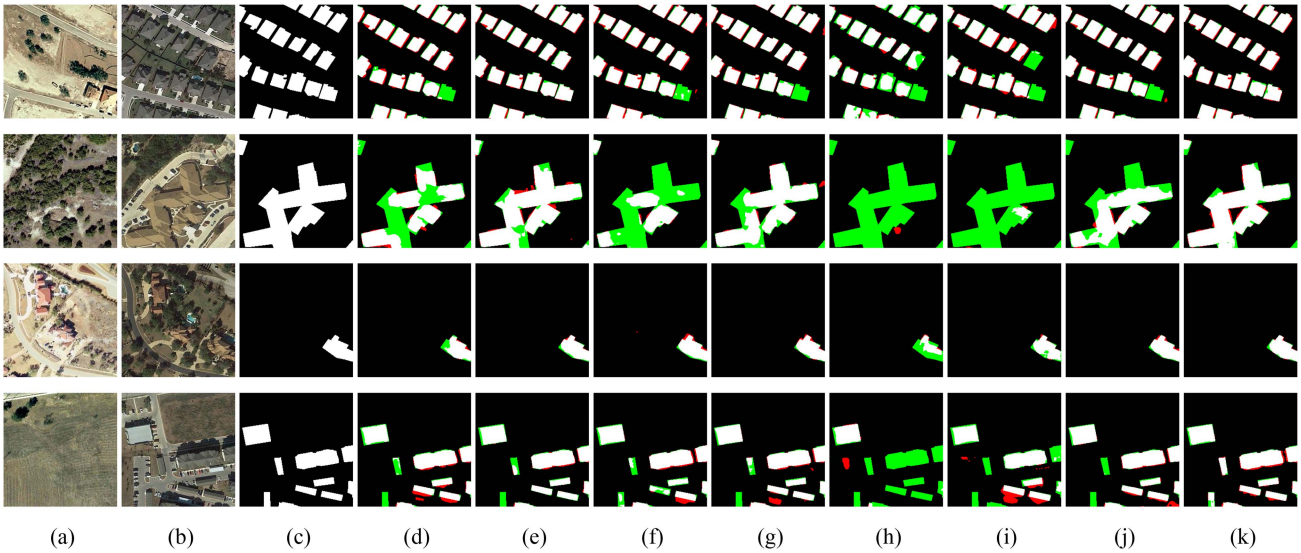

Fig. 7. Several inference results of comparison algorithms on the LEVIR-CD test sets. (a) T1 images. (b) T2 images. (c) Ground truth. (d) DDCNN [12]. (e) CLNet [40]. (f) BiT [42]. (g) ChangeFormer [43]. (h) SILICD [20]. (i) HANet [44]. (j) VcT [27]. (k) Ours. Different colors in change maps indicate true positives (white), false positives (red), true negatives (black), and false negatives (green), and the same annotation method is used in other visualization results of the article.

multiscale buildings, and the presence of occlusion and shadow interference. The results of the comparison demonstrate that the experimental results of the proposed method exhibit minimal red or green components, offering a distinct advantage over other comparison methods. Specifically, ChangeFormer [43] and VcT [27] perform well overall, but when the scenes are complex and the ground objects are irregular, the error rate increases significantly, shown as the third row in Fig. 7. However, CGFINet effectively integrates the spatial information in the semantic reconstruction process and introduces HFIM to improve the perceptual ability of multiscale targets, so the results of CGFINet in intricate scenarios are demonstrably superior to other methods. As illustrated in the fourth and fifth rows of Fig. 7, comparison methods have several degrees of false alarm or missed detection in the detection of the edge of the changed object. This is because the illumination angle and observation angle of the before and after temporal are inconsistent, which makes the shadow around the target change and there is partial occlusion. However, CGFINet reduces the interference of the above factors by aligning and fusing of dual temporal features, to enhance the precision of the detection process.

2) Comparisons on WHU-CD Dataset: As demonstrated in Table I, the indexes of our CGFINet on WHU-CD are 93.49%, 90.12%, 91.77%, 84.79%, 99.37% on the Pre, Rec, F1, IoU, and OA, respectively. This result indicates that our method demonstrates a significant improvement over the comparison methods. SNUNet [17] adopts the idea of dense connection to aggregate and refine the multilevel semantic features, achieving the second-best Pre, F1, IoU, and OA. However, the operation of dense connection is inevitably prone to generate redundant information, and the interference of dual temporal feature space registration error is not fully considered, which affects its detection performance. Our proposed CGFINet uses CGED to extract the spatial information of the change target to guide the semantic reconstruction process, suppressing the interference of irrelevant background noise, and adopts the alignment operation of dual temporal features, to achieve the optimal Pre, F1, IoU, and OA. Compared with suboptimal method, the F1 and IoU of CGFINet increase by 1.38% and 2.55%, respectively, which means that CGFINet can more accurately detect changes in dual temporal images.

To show the intuitive behavior of different approaches on WHU-CD dataset intuitively, we select eight methods with better F1 value performance, and display them in Fig. 8. In general, CGFINet achieves the best change detection performance with less pseudo changes and less missed detection. Moreover, WHU-CD dataset reflects the state of ground building reconstruction after the earthquake, and contains more abundant types of changes. In particular, the first row of Fig. 8 contains surfaces coated with cement or asphalt, which can impede the detection process and lead to the generation of false alarms in the majority of comparison methods. However, CGFINet extracts more discriminative features through the cross-scale guided DED structure, outputting more accurate detection results. The scenes in the second and third rows are more complex, and there are shadows and occlusions, resulting in different degrees of missed detection problems in SILICD [20] and HANet [44]. However, CGFINet effectively reduces the loss of change information by finegrained decoding process and change feature alignment strategy. The experimental results presented in the fourth row demonstrate that the proposed CGFINet performs better compared to other approaches when there is regional overlap between newly built (added) buildings and demolished (reduced) buildings in the scene, which may benefit from its refined extraction of edge information of changing objects.

3) Comparisons on DSIFN-CD Dataset: As shown in Table II, our proposed CGFINet outperforms comparison methods significantly on DSIFN-CD. Specifically, our method

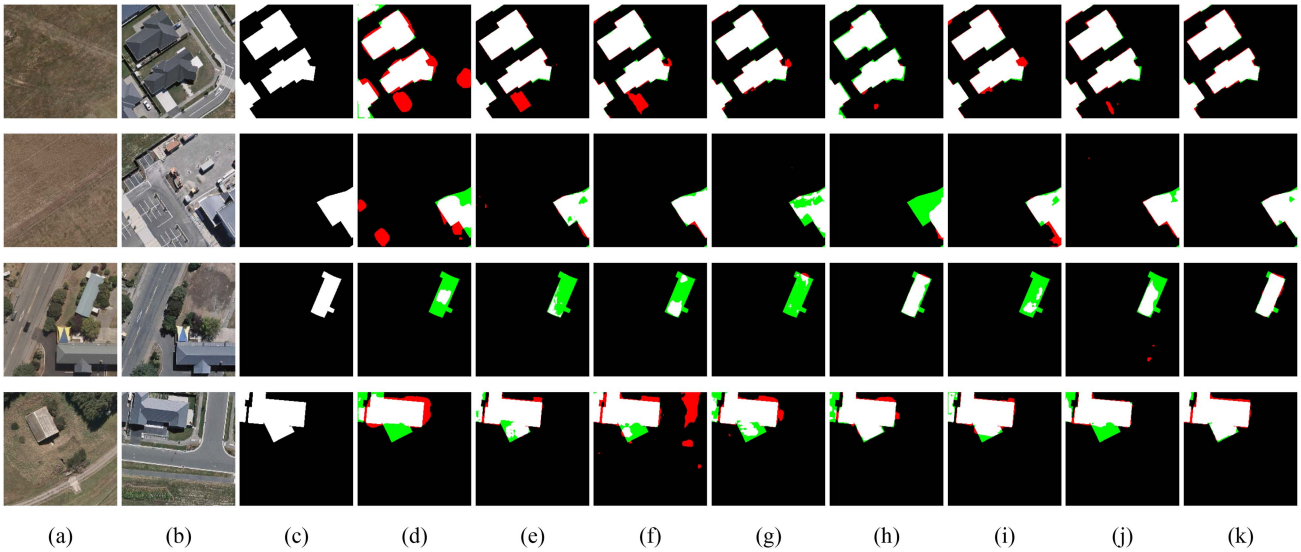

Fig. 8. Several inference results of comparison algorithms on the WHU-CD test sets. (a) T1 images. (b) T2 images. (c) Ground truth. (d) CLNet [40]. (e) SNUNet [17]. (f) BiT [42]. (g) ChangeFormer [46]. (h) SILICD [20]. (i) HANet [44]. (j) VcT [27]. (k) Ours.

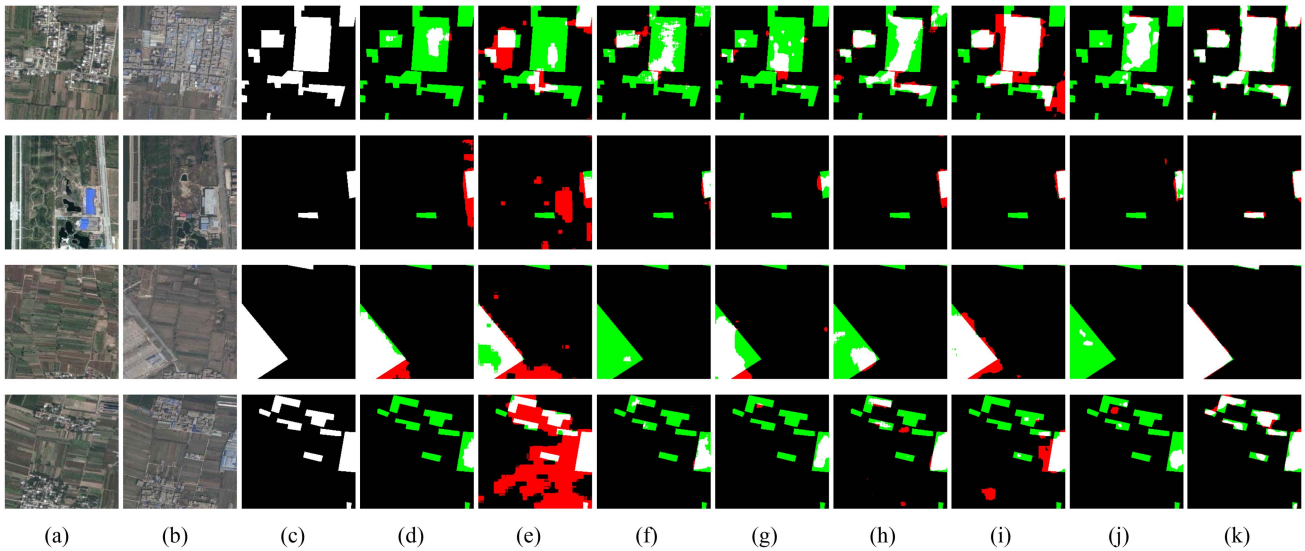

Fig. 9. Several inference results of comparison algorithms on the DSIFN-CD test sets. (a) T1 images. (b) T2 images. (c) Ground truth. (d) CLNet [40]. (e) STANet [36]. (f) SNUNet [17]. (g) BiT [42]. (h) ChangeFormer [43]. (i) SILICD [20]. (j) VcT [27]. (k) Ours.

achieves 78.45%, 81.47%, 79.93%, 66.57%, and 93.05% on the Pre, Rec, F1, IoU, and OA, respectively, achieving the optimal Rec, F1, IoU, and OA. Compared to the suboptimal VcT [27], our proposed method improved F1 by 7.76% and IoU by 12.92%, respectively. The DSIFN dataset contains a more diverse range of urban building types and fuzzy semantic information, which increases the challenge of change detection tasks. VcT [27] combines the global perception advantage of the transformer module to achieve the optimal Precision and suboptimal F1, but the balance between Pre and Rec is not optimal. Our proposed CGFINet enhances the perception of multiscale changes through HFIM, and utilizes CGED to transmit the edge information of multiscale changes to the enhanced semantic information, achieving optimal performance on experiments dataset.

In order to visually and qualitatively understand the performance on DSIFN-CD, we show prediction change maps in Fig. 9. Through visualization results, it can be seen that our method has almost no red or green parts in predicted change map, which is more advantageous compared to other methods. First, our CGFINet can adapt to different urban scenes or the diversity of ground building types, such as in the first and second row in Fig. 9. Some comparison methods, due to limitations in receptive fields or edge extraction, cannot fully detect large building areas or have missed detections for small building variations. However, through multiscale object perception and edge information enhancement, our CGFINet can reduce such missed detections and false detections. Second, our CGFINet is able to handle irrelevant changes caused by seasonal changes

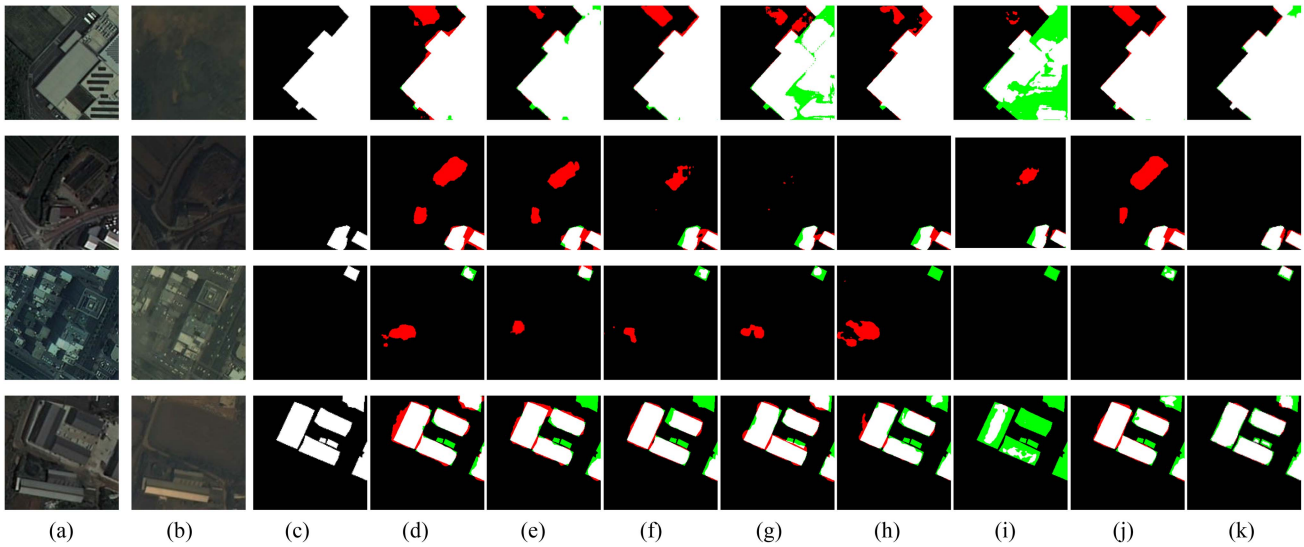

Fig. 10. Several inference results of comparison algorithms on the S2Looking test sets. (a) T1 images. (b) T2 images. (c) Ground truth. (d) DTCDSCN [41]. (e) SNUNet [17]. (f) BiT [42]. (g) ChangeFormer [43]. (h) SILICD [20]. (i) HANet [44]. (j) VcT [27]. (k) Ours.

or semantic ambiguity, such as in the third and fourth rows of Fig. 9. Due to the similarity in color between the top of the building and the farmland, most comparison methods miss building changes in the farmland or mistakenly detect farmland as changes of interest. However, our CGFINet enhances the recognizability of semantic information by utilizing cross scale structural information, which can better avoid the interference of such pseudo changes.

4) Comparisons on S2Looking Dataset: The experimental results are displayed in Table II. The evaluation indicators on this dataset are generally lower than those on other datasets, which further illustrates the challenge of the S2Looking dataset. Nevertheless, our methods achieved 73.90%, 53.71%, 62.21%, 45.15%, 99.21% on the Pre, Rec, F1, IoU, and OA, respectively, achieving the optimal OA and the suboptimal Pre. Compared to the suboptimal BiT [42], our proposed method improved F1 by 0.64% and Pre by 2.23%, respectively. CGNet and C2FNet obtained the optimal and suboptimal F1 and IOU, respectively. CGNet proposes a change prior guided fusion module, which is conducive to the spatial alignment of bitemporal features. This also proves the correctness of designing BAFM to suppress registration errors. It is worth noting that our CGFINet has significant advantages in parameters and running efficiency. See Section IV-E for details. This is due to the fact that CGFINet uses simple but efficient shifting and channel interaction to obtain multiscale receptive fields.

In order to intuitively and qualitatively understand the performance of our CGFINet on the S2Looking dataset, we show the visual results of the comparative experiment in Fig. 10. The visualization results show that there are fewer red and green parts in the experimental results of our method. First, our CGFINet is capable of adapting to the interference of more diverse ground buildings and seasonal changes in rural areas. For example, the first and second rows in Fig. 10 contain significant changes in farmland and vegetation. Some comparison methods mistakenly classify bare land or cement ground areas as changes. Our CGFINet can better avoid the interference of such irrelevant changes. Second, our CGFINet can well deal with the registration error and shadow problems in side view remote sensing images. For example, in the third and fourth rows of Fig. 10, some comparison methods lack robustness to the change of observation angle, resulting in many pseudo changes due to the registration error. By highlighting the edge structure of the changing region and aligning the features in time and space, our CGFINet can reduce such false detections.

D. Ablation Analysis

In this section, CGFINet ablation experiments are organized on the LEVIR-CD, WHU-CD, and DSIFN-CD datasets to verify contribution of the proposed key modules. We designed a base backbone with DED structure. The encoder in base backbone is the same as that in CGFINet. The encoded deep features are directly input into the decoder without HFIM. The decoder in the base backbone adopts the classic UNet structure. The decoded features are concatenated and upsampled layer by layer, and finally the same change detection head is used to get the change map. We introduce the proposed CGED, HFIM, and BAFM modules into the base backbone for ablation research. The performance of different combinations of backbone and the three modules shows that the corresponding modules can improve the performance of our proposed CGFINet. The quantitative analysis results are display in Table III. The formulas of these evaluation indicators are given in Section IV-A.

1) Ablation Study on CGED: The decoder used in the base network concatenates features directly in channel dimensions. We use CGED to replace the decoder in the Base network to verify the contribution of CGED. Take the LEVIR-CD dataset as an example for specific analysis, comparison of the results in the first and second rows of Table III reveals that CGED can improve

TABLE III
ABLATION EXPERIMENTS OF DIFFERENT MODULES

backbone	CGED	HFIM	TAFM	LEVIR-CD			WHU-CD			DSIFN-CD		
				F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
Base				89.78	81.46	98.96	87.57	77.88	99.06	75.24	60.31	82.41
	√			90.55	82.73	99.04	88.92	80.06	99.15	77.34	63.05	84.07
		√		90.51	82.67	99.04	88.04	99.08	99.11	76.77	62.30	81.74
			√	89.98	81.78	98.98	87.68	78.07	99.06	77.15	62.80	83.95
	√	√		90.83	83.21	99.08	89.92	81.69	99.25	78.59	64.73	85.84
	√		√	90.83	83.20	99.07	90.67	82.95	99.22	77.50	63.26	84.10
		√	√	90.70	82.99	99.06	90.24	82.21	99.26	79.22	65.59	84.43
	√	√	√	**90.90**	**83.31**	**99.08**	**91.77**	**84.79**	**99.37**	**79.93**	**66.57**	**93.05**

Note: Best results are in bold. √ means this module is introduced into backbone. Scores in this table are written in percentage (%).

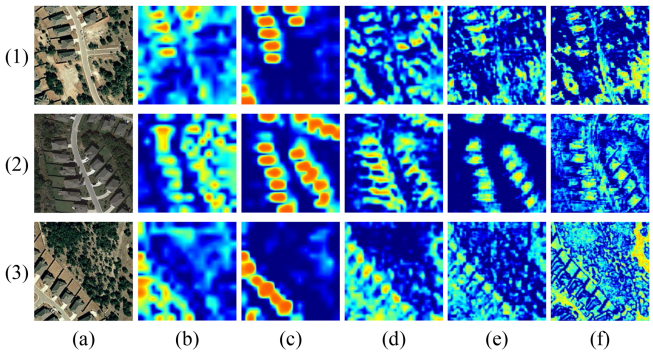

Fig. 11. Visualization of feature maps from HFIM and CGED. (1)-(3) Three images of scenes in LEVIR-CD. (a) Input images. (b) Feature maps before HFIM. (c) Feature maps after HFIM. (d) Feature maps after CGED-3. (e) Feature maps after CGED-2. (f) Feature maps after CGED-1.

indicators of our method. The F1, IoU, and OA result from Base + CGED are 90.55%, 82.73%, and 99.04%, respectively, which are better than those of base network. Comparing the results in the eighth row and the seventh row of Table III, the F1, IoU, and OA of CGFINet removing CGED decrease to 90.70%, 82.99%, and 99.06%, respectively. These results fully demonstrate the contribution of CGED to our CGFINet.

As shown in Fig. 11, the effectiveness of CGED is demonstrated through the presentation of three scenes images in LEVIR-CD. Fig. 11(d)–(f) illustrate the capacity of CGED to reconstruct semantic information accurately across diverse scene types. Furthermore, the spatial information is enhanced through the application of CGED, with the texture and edge features in the reconstructed feature map becoming increasingly discernible as the CGED iteration progresses from CGED-3 to CGED-1.

2) Ablation Study on HFIM: We verify the contribution of HFIM by adding HFIM between the encoder and decoder of base network. Take the LEVIR-CD dataset as an example for specific analysis, comparison of the results in the first and third rows of Table III displays the improvement effect of HFIM. The F1, IoU, and OA result from Base + HFIM are 90.51%, 82.67%, and 99.04%, respectively, which are better than those of base network. By comparing the results in the eighth and sixth rows of Table III, the F1 and IoU of CGFINet without HFIM decrease to 90.83% and 83.20%, respectively. These results prove the contribution of HFIM to our CGFINet.

We present the feature maps before and after HFIM processing to substantiate the efficacy of HFIM in Fig. 11(a)–(c). The feature map resulting from HFIM processing exhibits a notable enhancement of the target of interest, accompanied by the accurate extraction of buildings in disparate scenes, thereby rendering them more discernible from the background.

3) Ablation Study on BAFM: The change decoder in the base network directly subtracts the dual temporal features. We verify the contribution of BAFM by using BAFM as the change decoder of base network. Take the LEVIR-CD dataset as an example for specific analysis, comparison of data in the first and fourth rows of Table III indicates the advancement effect of BAFM. The IoU of Base + BAFM is 81.78%, the F1 value is 89.98%, and the OA value is 98.98%, which are better than the corresponding indicators of base network. The comparison of the results in the eighth and fifth rows of Table III reveals that the F1 and IoU of CGFINet without BAFM decrease to 90.83% and 83.21%, respectively. These results demonstrate the contribution of BAFM to our CGFINet.

As illustrated in Fig. 12, three pairs of scenario diagrams are presented to demonstrate the efficacy of BAFM. Fig. 12(d)–(g) represent the features output from BAFM-4 to BAFM-1 in turn. A comparison of the red boxed areas reveals that BAFM is capable of gradually and effectively suppressing the error detection caused by registration errors and light shadows. The BAFM algorithm, as presented in this article, demonstrates the capacity to accurately and robustly extract change features in a variety of scenes.

In summary, the above experimental results demonstrate that proposed modules enhance the efficacy of our method, and the combination of the three modules can further improve the accuracy. The combination of Base + CGED + HFIM + BAFM yields the optimal performance, with IoU of 83.31%, F1 of 90.90%, and OA of 99.08%.

E. Discussion

1) Effects of the Number of Shifted Pixel: The RubikConv algorithm employs a shift operation within the HFIM framework. In order to ascertain the impact of the number of shifted pixels on the robustness of HFIM, a series of experiments were disposed on four datasets. Different numbers of shifted pixels were designed in RubikConv, and the same optimization strategy was employed in all comparative experiments. The

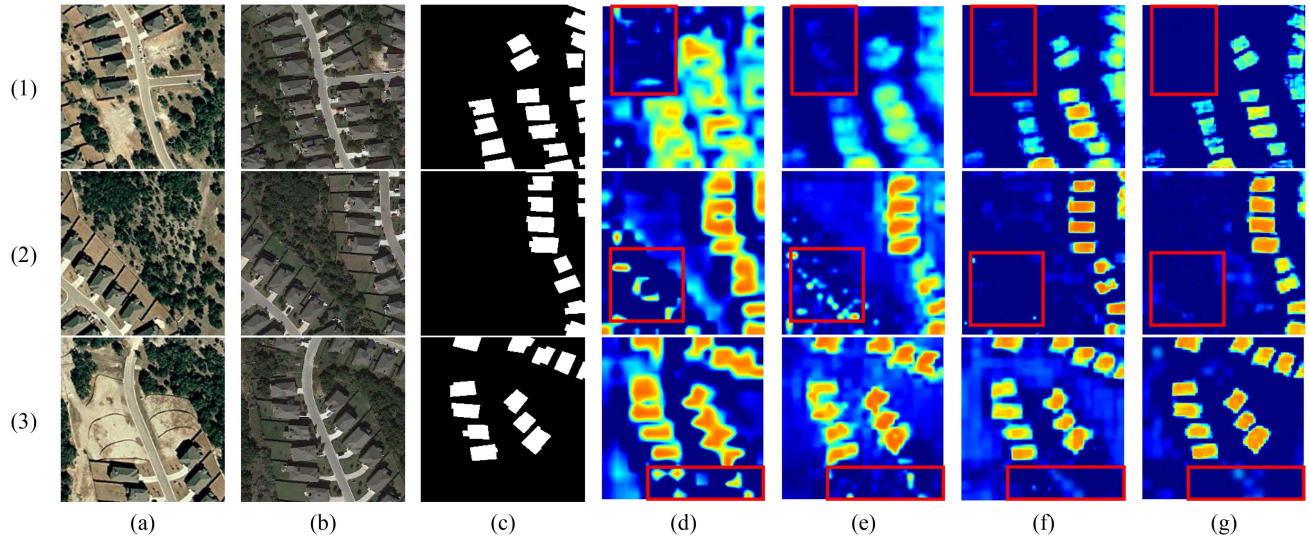

Fig. 12. Features visualization of BAFM. (1)–(3) Three pairs of images in LEVIR-CD. (a) Image1. (b) Image2 (c) Ground Truth. (d)–(g) Features from BAFM.

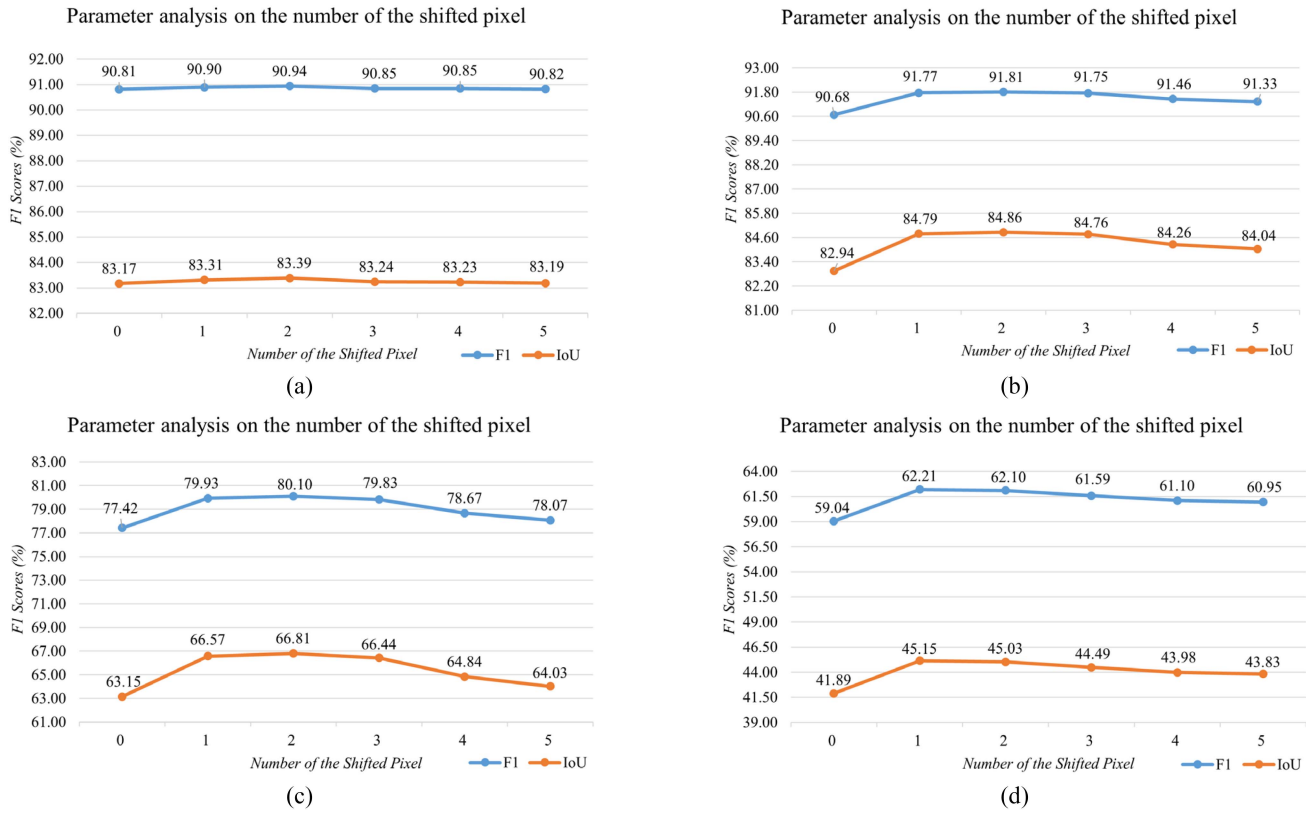

Fig. 13. F1 scores of CGFINet based on different number of the shifted pixel. (a) LEVIR-CD. (b) WHU-CD. (c) DSIFN-CD. (d) S2Looking.

primary evaluation metrics were the F1 and the performance of the IoU analysis algorithm. As illustrated in Fig. 13, the F1 and IoU of CGFINet exhibit a notable enhancement following the implementation of the shift operation. With the augmentation of the number of shifted pixels, the algorithmic performance initially improves and subsequently declines. This phenomenon may be attributed to the incremental complexity of the multichannel interactive reconstruction structure, which is associated with the increase in the number of shifted pixels. Therefore, in this article, we set the default value of the number of shifted pixels to 1. In addition, the impact of shifted pixel number on the robustness of CGFINet is more obvious on WHU-CD and S2Looking. For LEVIR-CD, the performance of CGFINet is more stable, which is mainly due to the spatial distribution

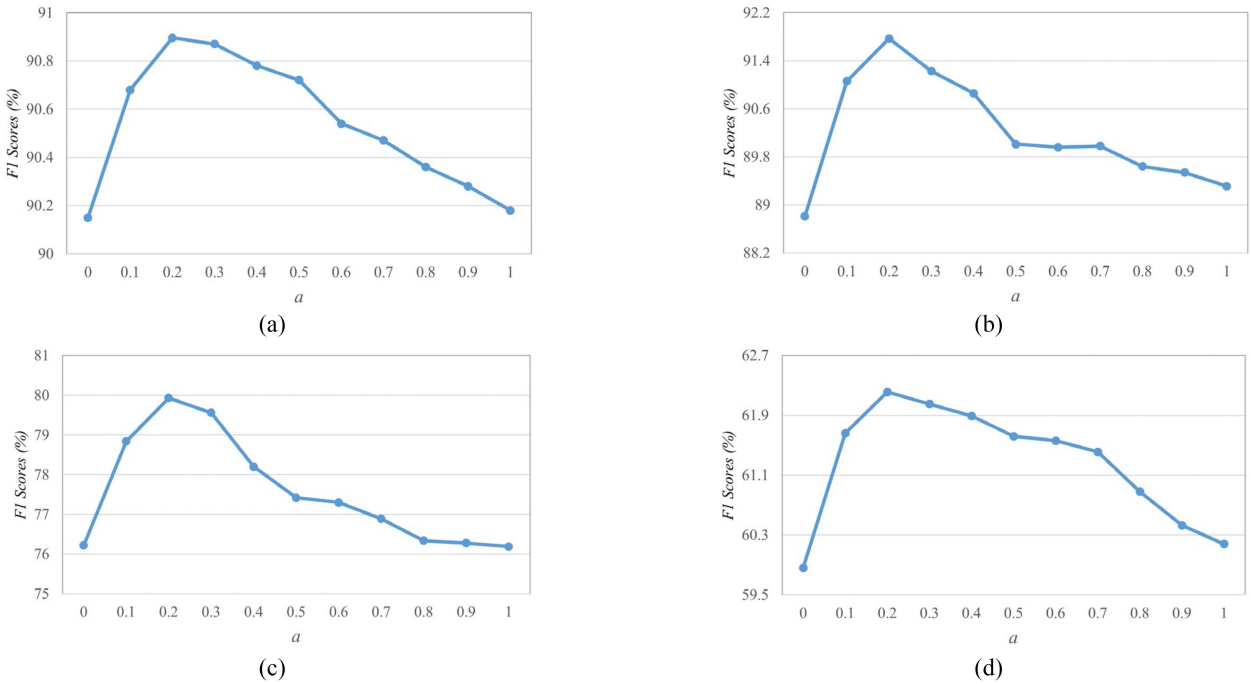

Fig. 14. F1 scores of CGFINet based on different weight parameter α. (a) LEVIR-CD. (b) WHU-CD. (c) DSIFN-CD. (d) S2Looking.

of changing targets in different datasets is inconsistent, and the changes in the evaluation results are within a reasonable range.

2) Effects of the Weight Parameter in Loss Function: There is a weight parameter (α) of semantic segmentation auxiliary loss L_s in the loss function. The objective of this experiment is to analyze the impact of α on the performance of CGFINet. The specific method is to set the parameter value α from 0 to 1 in steps of 0.1, and then count the F1 value of CGFINet on different datasets. The other configurations in the comparison experiment are exactly the same. As shown in Fig. 14, we can find the following conclusions. First, the performance of CGFINet has the same trend on different datasets, that is, there is an obvious peak when $\alpha = 0.2$. This shows that the change loss function is relatively more important. Second, when $\alpha = 0$, the indicator of CGFINet decreases, which indicates that the semantic segmentation auxiliary loss function term is indispensable. Third, the performance of CGFINet is relatively stable for the LEVIR-CD dataset. However, change trend is more obvious on other datasets, which may be due to the different distribution of change instances in different datasets.

3) Parameters and Running Efficiency: We analyze the efficiency of all comparison approaches on S2Looking, and display assessment results in Table IV. The evaluation indexes included the model parameters (Params), floating point operations per second (FLOPs), training time, and inference time. The experimental outcomes indicate that the CGFINet achieves the excellent F1 score on S2Looking using a limited number of parameters, and a relatively low computational complexity. Specifically, the CNN-based methods (FC-EF [39], DDCNN [12], CLNet [40], SILICD [20]) have less Params, FLOPs and faster training speed. However, this is often at the expense of their performance. Although the Transformer-based methods (BiT [42], ChangeFormer [43], and VcT [27]) improve the performance, they require a lot of training time. It is worth noting that our CGFINet achieve excellent performance with much smaller Params, FLOPs and running time than CGNet [45] and C2FNet [46].

TABLE IV
COMPARISON OF PARAMETERS AND OPERATION EFFICIENCY

Method	Params (M)	FLOPs (G)	Training time (1 epoch/min)	Inference time (s)	F1
FC-EF [39]	1.35	3.58	8.02	119.61	37.62
DDCNN [12]	9.84	8.72	8.59	111.33	54.47
CLNet [40]	8.00	7.36	10.41	145.23	55.36
DTCDSCN [41]	31.26	13.22	10.03	129.84	57.42
STANet [36]	16.93	13.16	24.80	194.38	46.75
SNUNet [17]	12.03	54.83	87.44	254.06	55.62
BiT [42]	3.55	10.92	20.92	367.66	61.81
ChangeFormer [43]	20.75	21.18	27.17	264.92	56.50
SILICD [20]	13.06	9.18	7.73	144.92	60.58
HANet [44]	3.03	17.67	142.32	221.41	58.54
VcT [27]	3.57	10.98	34.13	479.69	61.37
CGNet [45]	33.68	62.95	26.23	147.96	64.33
C2FNet [46]	16.17	47.55	39.49	266.91	63.38
Ours	3.32	10.20	8.65	119.34	62.21

V. CONCLUSION

In this article, we proposed a new VHR remote sensing image change detection method. Our aim is to enhance the guidance of cross-scale spatial information on the process of deep semantic information reconstruction, alleviate the interference of background noise on real change information in images, and extract accurate changing region by fusing bitemporal features.

Therefore, we project a DED network with cross-scale guided enhancement decoders to filter the intervention of background noise and enhance edge characteristics of targets. At the same time, we apply a high-order feature interaction module to mine and fuse multiscale features, and employ a series of bitemporal feature alignment fusion module to alleviate the pseudo change problem caused by temporal and spatial deviation. We have carried out extensive experimental verification on four classical change detection datasets, and proved the effectiveness and advantages of our proposed CGFINet.

Although CGFINet shows excellent performance, it still has some limitations. For example, as an end-to-end supervised learning network, our CGFINet depends on the complete and annotated change detection dataset, which is a very time-consuming work. Therefore, in addition to the construction of more complete change detection datasets, it is of great significance to mine self-supervised or semi-supervised methods, or to seek the pretraining model for optical remote sensing image to improve the generalization ability of network. Therefore, our next work will focus on these aspects.

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, Jun. 1989.

[2] B. Fang, G. Chen, L. Pan, R. Kou, and L. Wang, "GAN-based siamese framework for landslide inventory mapping using bi-temporal optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 391–395, Mar. 2021.

[3] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "ChangeNet: A deep learning architecture for visual change detection," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018.

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

[5] F. Gao, X. Wang, Y. Gao, J. Dong, and S. Wang, "Sea ice change detection in SAR images based on convolutional-wavelet neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1240–1244, Aug. 2019.

[6] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, May 2020, Art. no. 1688.

[7] Z. Zheng, S. Du, H. Taubenböck, and X. Zhang, "Remote sensing techniques in the investigation of aeolian sand dunes: A review of recent advances," *Remote Sens. Environ.*, vol. 271, Mar. 2022, Art. no. 112913.

[8] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "FFCA-YOLO for small object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5611215.

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

[10] F. I. Diakogiannis, F. Waldner, and P. Caccetta, "Looking for change? Roll the dice and demand attention," *Remote Sens.*, vol. 13, no. 18, Sep. 2021, Art. no. 3707.

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

[12] 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.

[13] 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, May 2022.

[14] H. Zheng et al., "HFA-Net: High frequency attention Siamese network for building change detection in VHR remote sensing images," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108717.

[15] X. Zhao, K. Zhao, S. Li, and X. Wang, "GeSANet: Geospatial-awareness network for VHR remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402814.

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

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

[18] L. Zhang, X. Hu, M. Zhang, Z. Shu, and H. Zhou, "Object-level change detection with a dual correlation attention-guided detector," *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 147–160, Jul. 2021.

[19] 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.

[20] H. Chen et al., "Continuous cross-resolution remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5623320.

[21] H. Zhang et al., "BiFA: Remote sensing image change detection with bitemporal feature alignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5614317.

[22] S. Zhao et al., "Rs-mamba for large remote sensing image dense prediction," 2024, *arXiv:2404.02668*.

[23] H. Zhang et al., "CDMamba: Remote sensing image change detection with mamba," 2024, *arXiv:2406.04207*.

[24] P. Chen et al., "A region-based feature fusion network for VHR image change detection," *Remote Sens.*, vol. 14, no. 21, Nov. 2022, Art. no. 5577.

[25] S. Liang, Z. Hua, and J. Li, "Enhanced feature interaction network for remote sensing change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Sep. 2023, Art. no. 7505605.

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

[27] B. Jiang et al., "VcT: Visual change transformer for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Oct. 2023, Art. no. 2005214.

[28] S. Zhao, X. Zhang, P. Xiao, and G. He, "Exchanging dual encoder-decoder: A new strategy for change detection with semantic guidance and spatial localization," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Oct. 2023, Art. no. 4508016.

[29] J. Pan et al., "M-Swin: Transformer-based multi-scale feature fusion change detection network within cropland for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Mar. 2024, Art. no. 4702716.

[30] L. Song, M. Xia, J. Jin, M. Qian, and Y. Zhang, "SUACDNet: Attentional change detection network based on Siamese U-shaped structure," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102597.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[32] Z. Su et al., "Pixel difference networks for efficient edge detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5097–5107.

[33] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[34] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S.-N. Lim, and J. Lu, "HorNet: Efficient high-order spatial interactions with recursive gated convolutions," 2022, *arXiv:2207.14284*.

[35] N. Zheng, M. Zhou, C. Zhou, and C. C. Loy, "Rubik's cube: High-order channel interactions with a hierarchical receptive field," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, vol. 36, pp. 18377–18390.

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

[37] 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.

[38] L. Shen et al., "S2Looking: A satellite side-looking dataset for building change detection," *Remote Sens.*, vol. 13, no. 24, Dec. 2021, Art. no. 5094.

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

[40] Z. Zheng, Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang, "CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 247–267, May 2021.

[41] 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.

[42] H. Chen, Z. Qi, and Z. Shi, "Remote Sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607514.
[43] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 207–210.
[44] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3867–3878, Apr. 2023.
[45] C. Han, C. Wu, H. Guo, M. Hu, J. Li, and H. Chen, "Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8395–8407, Aug. 2023.
[46] C. Han, C. Wu, M. Hu, J. Li, and H. Chen, "C2F-SemiCD: A coarse-to-fine semi-supervised change detection method based on consistency regularization in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Feb. 2024, Art. no. 4702621.

Qichao Han received the M.E. degree in optical engineering from Harbin Institutes of Technology (HIT), Harbin, China, in 2021. He is currently working toward the Ph.D. degree with the Research Center for Space Optical Engineering, HIT.

His research interests include remote sensing image change detection, multisource fusion, and image registration.

Xiyang Zhi received the Ph.D. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2012.

He is currently a Full Professor at the HIT. His current research interests including remote sensing image acquisition and processing, multisource remote sensing image intelligent interpretation, and optical target detection and identification.

Jianming Hu (Member, IEEE) received the M.Eng. and Ph.D. degrees in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2017 and 2022, respectively.

He was a Visiting Scholar with the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, in 2019. He is currently an Assistant Professor with HIT. His research interests include remote sensing image processing, image characteristic analysis, and sea-aero target detection and identification.

Yuanxin Huang received the M.E. degree in optical engineering in 2018 from the Harbin Institutes of Technology, Harbin, China, where he is currently working toward the Ph.D. degree.

His research interests include object detection, instance segmentation, and object tracking.

Wenbin Chen received the M.E. degree in optical engineering in 2018 from the Harbin Institutes of Technology, Harbin, China, where he is currently working toward the Ph.D. degree.

His research interests include small target detection and hyperspectral image processing.

Shikai Jiang (Member, IEEE) received the M.Eng. and Ph.D. degrees in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2018 and 2022, respectively.

He is currently an Assistant Professor of Optical Engineering with the School of Astronautics, HIT. His research interests include image processing of space optical remote sensing satellite including image inversion restoration, image fusion, super-resolution, and target detection and identification.

Jinnan Gong received the Ph.D. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2016.

He is currently an Assistant Professor with HIT. His research interests include optical image processing, processing algorithm optimization in engineering level, and performance evaluation.