BiFA: Remote Sensing Image Change Detection with Bitemporal Feature Alignment

Haotian Zhang, Hao Chen, Chenyao Zhou, Keyan Chen,
Chenyang Liu, Zhengxia Zou, and Zhenwei Shi*, *Senior Member, IEEE*

Abstract—**Despite the success of deep learning-based change detection methods, their existing insufficiency in temporal (channel, spatial) and multi-scale alignment have rendered them insufficient capability in mitigating external factors (illumination changes and perspective differences, etc.) arising from different imaging conditions during change detection. In this paper, a Bi-temporal Feature Alignment (BiFA) model is proposed to produce a precise change detection map in a lightweight manner by reducing the impact of irrelevant factors. Specifically, for the temporal alignment, the Bi-temporal Interaction (BI) module is proposed to realize the alignment of the bi-temporal image channel level. Our intuition is introducing the bi-temporal interaction in the feature extraction stage may benefit suppressing the interference, such as illumination changes. Simultaneously, the Alignment module based on Differential Flow Field (ADFF) is proposed to explicitly estimate the offset of the bi-temporal image and realize their spatial level alignment to mitigate the inadequate registration resulting from different perspectives. Furthermore, for the multi-scale alignment, we introduce the Implicit Neural alignment Decoder (IND) to produce more refined prediction maps achieving precise alignment of multi-scale features by learning continuous image representations in coordinate space. Our BiFA outperforms other state-of-the-art methods on six datasets (such as the F1/IoU scores are improved by 2.70%/3.91%, 2.01%/2.94% on LEVIR+-CD and SYSU-CD, respectively) and displays greater robustness in cross-resolutions change detection. Our code is available at https://github.com/zmoka-zht/BiFA.**

Index Terms—**Change detection (CD), high-resolution optical remote sensing image, feature alignment, bi-temporal interaction, flow field, implicit neural representation.**

I. Introduction

CHANGE detection (CD) is designed to monitor changes occurring in the same region at different times. With the increasing accessibility of remote sensing data, many multi-temporal high-resolution remote sensing images are used for urban expansion surveys [1–4], land management [5–7], and damage assessment [8], as comprehensively summarized in [9]. However, acquiring multi-temporal images may result in irrelevant interference due to variations in imaging conditions, which manifest as color differences arising from disparities in illumination intensity or building renovation and inadequate registration caused by variations in imaging perspectives. Consequently, effectively utilizing these images to accurately detect change regions becomes challenging.

Till now, the investigation of conventional change detection techniques, including algebra-based, transformation-based, and classification-based methods, has been extensive. The algebra-based methods utilize arithmetic operations, such as image rationing [10], image differencing [11], and image regression [12], and image adaptive region measuring [13, 14], to generate a change map by directly comparing the pixel values of bi-temporal images. The transformation-based approaches, such as principal component analysis (PCA) [15], tasselled cap transformation [16], and change vector analysis [17], employ feature space mapping to distinguish change information. The classification-based methods, such as k-nearest neighbors [18] and support vector machines [19], distinguish changing pixels by spatiotemporal or differential features. Nonetheless, these techniques are heavily dependent on empirical design. They are susceptible to noise interference, particularly for high-resolution remote sensing images with intricate texture features and fine image details.

With the continuous development of deep learning-based methods in computer vision, remote sensing communities have noticed the significant feature extraction capability of convolutional neural networks (CNN), such as FCN [20] and U-Net [21], and effectively applied them to CD tasks. Peng et al. [22] proposed a U-Net++ network based on multilateral branch fusion. Fang et al. [23] combined the Siamese network and NestedUNet to build SNUNet. Daudt et al. [24] proposed three kinds of fully convolutional networks established on U-Net, the image-level-based FC-Siam-Conc and the feature-level-based FC-EF, FC-Siam-Diff. While the feature-level-based technique utilizes two weighted shared networks to learn single-temporal features individually, image-level-based networks input bi-temporal images as an entire set. Subsequently, the majority of techniques have adhered to these two approaches for extracting bi-temporal image features [25–30]. Recently, some researchers have introduced specific attention mechanisms to obtain more discriminative differential features. For example, spatial attention [31, 32], self-attention [29, 33, 34], and cross-attention [27, 28, 30], are utilized for spatial alignment between bi-temporal features. In [30], a

The work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160401), the National Natural Science Foundation of China under the Grants 62125102, the Beijing Natural Science Foundation under Grant JL23005, and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Haotian Zhang, Chenyao Zhou, Keyan Chen, Chenyang Liu, and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Hao Chen is with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

CD network is proposed for intra-scale interaction and inter-scale feature fusion. Chen et al. [34] propose the bi-temporal image transformer module (BIT) to realize spatial-temporal alignment through the Transformer encoder and Transformer decoder. In addition, many recent CD models employ bilinear interpolation or deconvolution methods to attain multi-scale feature alignment and aggregate high-level semantic and low-level detail information (such as low-resolution depth features and high-resolution shallow features) to achieve precise detection [28, 30].

Despite the aforementioned methods have achieved promising performance, they still exhibit certain limitations in temporal (channel, spatial) and multi-scale alignment. **For the temporal alignment**, firstly, the aforementioned methods exhibit a deficiency in realizing bi-temporal interaction during the feature extraction stage. Specifically, in the feature extraction stage, image-level methods are deficient in fully leveraging the distinctive attributes of the change detection task and in profound feature interaction (e.g., by solely concatenating images). On the other hand, the methods at the feature level absent bi-temporal feature interaction (bi-temporal image feature extraction and bi-temporal feature interaction are separated). Consequently, these methods may be difficult to suppress the interference of unrelated factors such as illuminative and seasonal variation. Secondly, the majority of previous models utilized attention-based approaches (such as spatial attention, self-attention, etc.) to achieve spatial alignment without a dedicated emphasis on predicting pixel-level offsets. Consequently, these methods failed to explicitly represent the precise pixel offset. This limitation poses a challenge in addressing the issue of insufficient registration resulting from variations in perspective. **For the multi-scale alignment**, the off-the-shelf methods usually utilize bilinear interpolation or deconvolution to align feature maps of different scales. However, the utilization of bilinear interpolation may result in a loss of precision in the contextual information, while the implementation of deconvolution introduces additional parameters, thereby raising a challenge to the attainment of precise alignment of multi-scale features in a lightweight manner.

In response to the limitations of the aforementioned methods, we present the BiFA that addresses the inadequacy of bi-temporal feature alignment by incorporating temporal alignment (channel, spatial), and multi-scale alignment. The network comprises three critical modules, namely the bi-temporal interaction (BI) module, the alignment module based on differential flow field (ADFF), and the implicit neural alignment decoder (IND). For temporal alignment, the BI module has been developed to address requiring deep bi-temporal feature interaction during the feature extraction stage. The incorporation of feature interaction into the backbone network facilitates the achievement of channel alignment for bi-temporal image features. In our intuition, with the guidance of another temporal feature, it can suppress the interference of unrelated factors, such as illuminative and seasonal changes, and extract more robust features. The ADFF module is employed to address the issue of inadequate registration that arises due to divergent perspectives and explicitly estimates pixel-level offset between bi-temporal images to achieve spatial alignment of bi-temporal images, thereby mitigating erroneous detections due to offset. For multi-scale alignment, the IND, which aims to acquire a continuous representation of images within the coordinate space, is utilized to solve the problem of multi-scale feature alignment and achieve the precise alignment between multi-scale features through a lightweight approach for facilitating the performance of CD.

The contribution of our work can be summarized as follows:

- A novel CD method BiFA is proposed, which realizes bi-temporal feature alignment at the temporal (channel, spatial), and multi-scale levels, respectively, to identify the changes of interest and exclude extraneous changes achieving more accurate detection.
- For temporal alignment, we propose the BI module and ADFF, the former is utilized to perform deep interaction of bi-temporal features (channel alignment) to suppress interference of irrelevant factors, and the latter is employed to explicitly estimate pixel-level offset between bi-temporal images (spatial alignment) to alleviate the problem of insufficient registration caused by the different perspectives.
- For multi-scale alignment, we incorporate the implicit neural representation learning method into the CD task. By acquiring the continuous representation of images within the coordinate space, it realizes the accurate alignment of multi-level bi-temporal differential feature maps in a lightweight manner.
- Qualitative and quantitative studies on six datasets show that our proposed BiFA outperforms state-of-the-art methods. Further, our experiments show that BiFA has advantages over other methods in cross-resolution CD tasks.

The rest of this paper is organized as follows. Section II describes the related work. Section III gives the details of our proposed method. Some experimental results are reported in section IV. And the conclusion is made in Section V.

II. Related Work

A. Deep Learning based Remote Sensing Image Change Detection

According to the characteristics of bi-temporal input in CD tasks, the existing deep learning methods can be roughly divided into image-level and feature-level categories. The image-level method directly concatenates the bi-temporal images in the channel dimension, which are sent into a single segmentation network for detection [22–24, 26, 35–38]. Alcantarilla et al. [38] first concatenate bi-temporal street view images with three channels into one image with six channels. Then, the six-channel image is input into FCN to realize street view image change detection. Peng et al. [22] utilize the same concatenation method to obtain a binary change map by the U-net++ network with multi-side branch fusion. Fang et al. [23], based on previous work, employ dense connections between encoders and decoders to supplement the middle layer information while adding channel attention to enhance distinguishable features. Jiang et al. [37] propose a weighted rich-scale inception coder network that dynamically assigns

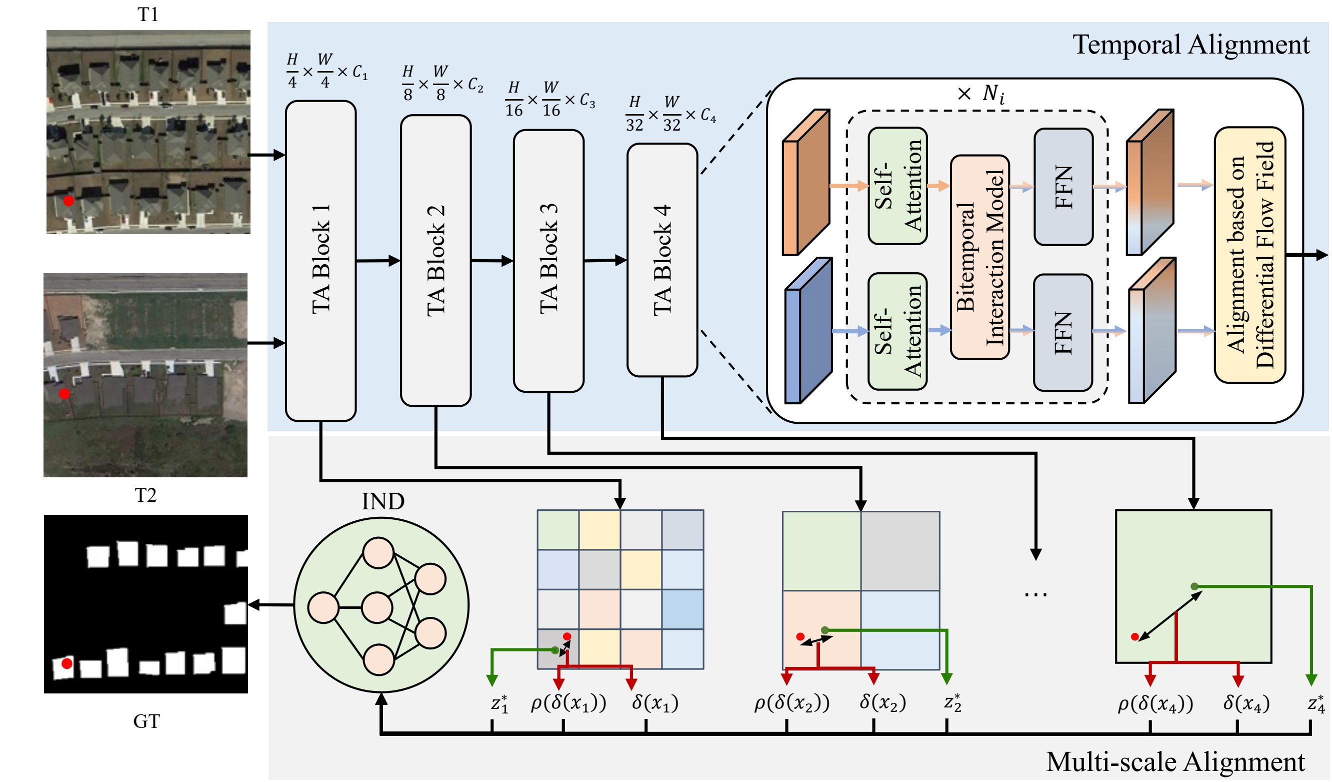

Fig. 1. The architecture of the proposed BiFA. The bi-temporal images T1 and T2 are initially fed to Temporal Alignment (TA) Bolcks, which mainly comprise the Bi-temporal Interaction (BI) module and the alignment module based on differential flow field (ADFF), to realize the channel-level and spatial-level alignment and produce multi-stage differential features. Subsequently, the multi-stage differential features are sent to the implicit neural alignment decoder (IND) and aligned in a lightweight manner by learning the continuous representation of the images in the coordinate space to generate accurate prediction maps. The red point refers to a query coordinate, and the green point is the nearest coordinate from the query coordinate on the feature.

appropriate weights to features of different scales to achieve accurate detection.

The feature-level approach utilizes a pair of weight-shared networks to acquire single-temporal features independently [1, 24, 27, 28, 30, 33, 34, 39–42]. Zhan et al. [39] use the Siamese network combining measurement learning strategies to update the initial results of threshold segmentation through the k-nearest neighbor method. Daudt et al. [24] propose the U-Net based fully convolutional networks for CD. On the basis of the work [24], Guo et al. [43] improve the performance by combining the fully convolutional Siamese network with contrast loss. Shi et al. [40] impose constraints on the differential features of the output at each stage of the encoder to produce a finer representation. Li et al. [33] propose the dense skip connection module to realize multi-level feature aggregation between encoders and decoders. Some methods introduce the Transformer module, which exhibits solid aptitude in modeling global relationships, to address the limited capacity of CNN models to capture long-term dependencies. Zhang et al. [44] employ many Swin-Transformer blocks to construct a pure Transformer Siamese CD network. Feng et al. [30] have presented a novel approach for feature fusion in intra-scale and inter-scale, utilizing the Transformer module. The methodology employs two weight-shared backbones to extract distinct features from two images. Subsequently, it facilitates information aggregation between different backbone features and scale features based on single temporal features. Chen et al. [34] propose a bi-temporal image transformer (BIT) composed of one Transformer encoder and two Transformer decoders to capture temporal and spatial context.

However, the aforementioned model exhibits a deficiency in bi-temporal feature interaction during feature extraction. This limitation may result in the model having difficulty suppressing the interference from irrelevant factors such as illuminative and seasonal change and ignoring the critical change information. The objective of this study is to improve the capacity of feature resistance against interference by incorporating feature information from the other phase during the feature extraction stage. To achieve this, a bi-temporal interaction (BI) module is proposed.

B. Optical Flow

Optical Flow is widely used in video processing tasks [45] to represent the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by relative motion. Gadde et al. [46] realize video semantic segmentation by warping the internal features of the network. Nilsson et al. [47] wrap the features of adjacent frames along the optical flow to predict the final segmentation map. Simonyan et al. [48] employ continuous multi-frame optical flow stacking for video action recognition. Furthermore, the idea of optical flow has also been incorporated into the image semantic segmentation task. Li et al. [49] propose the concept of semantic flow to align feature maps of different levels. In [50], the flow field is learned to warp image features and enhance the consistency of object features. Very recently, the optical flow has been applied in remote sensing CD task [29, 51]. For instance, Liu et al. [29] introduce it to develop a prior feature extraction module utilizing the flow field. The module is applied to the single-temporal image during feature extraction to enhance the structure prior.

Different from existing methods that only enhancement of individual temporal features by optical flow, we believe that pixel-level registration of bi-temporal images can be represented by the "motion" of each pixel from one phase to another. We call the field describing this motion the differential flow field. Then, an alignment module based on the differential flow field (ADFF) is proposed to explicitly estimate the pixel-level offset between bi-temporal images to alleviate the problem of insufficient registration caused by the difference in perspective.

C. Implicit Neural Representation

In recent 3D reconstruction methods, shapes, scenes, and objects can be represented by a multi-layer perception (MLP) that maps coordinate to signals, known as implicit neural representation [52–55]. It is essentially a continuously differentiable function that maps the properties of spatial points (e.g., amplitude, color, depth) to functions of related coordinates. For example, DeepSDF [52] learns a group of continuous signed distance functions for shape representation. NeRF [53] learns the implicit representation of complex scenes to achieve view synthesis.

Due to the notable advancements of implicit neural representation in 3D tasks, many researchers have incorporated this technique into 2D tasks [56–61]. Inspired by the implicit neural representation, LIIF [56] designs a local implicit image function to achieve continuous image superresolution, which takes coordinates and nearby feature representations as inputs, and outputs RGB values of corresponding positions. Chen et al. [60] improve on this basis and propose a dual path decoder based on the implicit function, which generates high-resolution images by parsing coordinates of global and local levels to achieve image super-resolution. IFA [57] proposes an implicit feature alignment function and applies it to the alignment of multilevel feature maps. Implicit PointRend [58] focuses on instance segmentation with point supervision, where the implicit function generates different parameters of the point head for each object. Qi et al. [61] utilize implicit functions to construct implicit 3D scene representation and realize semantic segmentation from new views.

In contrast to the literature above, our study incorporates implicit neural representation into the remote sensing CD task. The attainment of precise multi-level differential feature map alignment in a lightweight manner is made possible by acquiring a continuous representation of images in coordinate space.

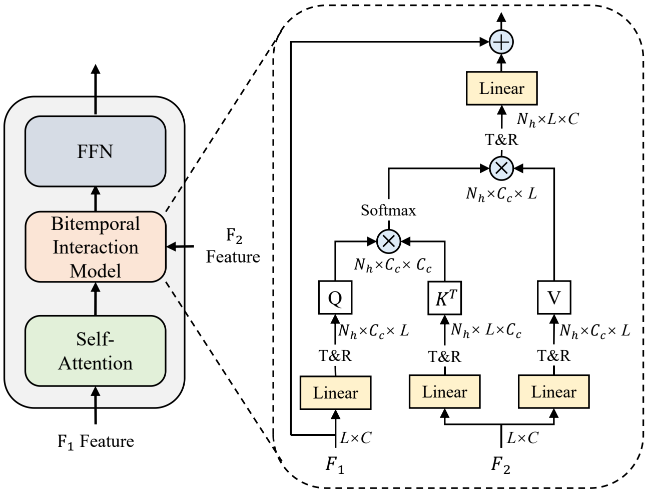

Fig. 2. Illustration of our Bi-temporal Interaction module (BI). T&R means transpose and reshape operations. F_1 and F_2 represent bi-temporal features, respectively.

III. Bitemporal Feature Alignment Change Detection Model

A. Overview

The architecture of the proposed BiFA is shown in Fig. 1, which is mainly composed of the bi-temporal interaction (BI) module, the alignment module based on differential flow field (ADFF), and the implicit neural alignment decoder (IND). Given the bi-temporal remote sensing images T$_1$ and T$_2$, firstly, the multi-stage image feature $\left\{\mathbf{I_1^i}\right\}_{i=1}^4$ and $\left\{\mathbf{I_2^i}\right\}_{i=1}^4$ are extracted by the Temporal Alignment (TA) Blocks. It is noteworthy that instead of using the original transformer-based method [62], we added the BI module into Self-attention and FFN. The bi-temporal features are sent to BI, and the alignment of the bi-temporal feature channel level is achieved by utilizing the guidance of the features from another phase to counteract the impact of unrelated factors. Subsequently, $\left\{\mathbf{F_1^i}\right\}_{i=1}^4$ and $\left\{\mathbf{F_2^i}\right\}_{i=1}^4$, which are at channel-level alignment, are transmitted to the ADFF. And the spatial alignment of bi-temporal features is achieved by predicting the differential flow field between $\left\{\mathbf{F_1^i}\right\}_{i=1}^4$ and $\left\{\mathbf{F_2^i}\right\}_{i=1}^4$. Then, the bi-temporal differential features of various stages $\{\mathbf{diff}\}_{i=1}^4$ are obtained utilizing absolute subtraction. Finally, IND is employed to convert the differential features of various stages into continuous feature maps, thereby enabling the lightweight alignment of the differential features of these stages and facilitating the production of the precise prediction map.

B. Bi-temporal Interaction Module

Due to the different imaging periods of multi-temporal images, it is easy to interfere with irrelevant factors such as illuminative and seasonal changes. By aligning bi-temporal features at the channel level, guided by the feature of another phase in the feature extraction stage, we believe the impacts above can be effectively mitigated. Therefore, the bi-temporal interaction (BI) module is being proposed.

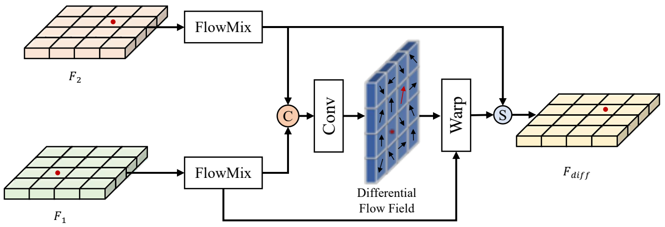

Fig. 3. Illustration of our ADFF. The red dots represent the location of the object. The FlowMix is used for feature enhancement. C and S mean concatenation and absolute subtraction, respectively. Conv is the convolution layer. Warp operation represents warping the F_1 in space to align with F_2.

Take $\mathbf{F_2}$ guiding $\mathbf{F_1}$ as an example (for convenience, stage labeling is omitted), as shown in Fig. 2, given the bi-temporal features $\mathbf{F_1} \in \mathbb{R}^{L\times C}$ and $\mathbf{F_2} \in \mathbb{R}^{L\times C}$, where $L = H \times W$ represents the sequence length. In contrast to the previous spatial alignment method using self-attention and cross-attention, our approach is to achieve alignment of the bi-temporal feature at the channel level. Therefore, $\mathbf{F_1}$ and $\mathbf{F_2}$ are linearly mapped to different sub-spaces. Subsequently, they are transposed and reshaped to obtain the query vectors $\mathbf{Q} \in \mathbb{R}^{N_h \times C_c \times L}$, key vectors $\mathbf{K} \in \mathbb{R}^{N_h \times C_c \times L}$ and value vectors $\mathbf{V} \in \mathbb{R}^{N_h \times C_c \times L}$. N_h represents the number of heads, and C_c is the channel dimension of each head. $\mathbf{Q}$ and $\mathbf{K}$ calculate the correlation scores between channels by scaled dot-product attention to generate attention map. Then the attention map and $\mathbf{V}$ are aggregated by matrix multiplication to obtain the feature $\mathbf{F}_{align}$, which means the bi-temporal image aligned in the channel dimension. The calculation process is as follows:

$$\mathbf{Q} = \text{Transpose}(\text{Reshape}(\mathbf{F_1 W}^q)) \tag{1}$$

$$\mathbf{K} = \text{Transpose}(\text{Reshape}(\mathbf{F_2 W}^k)) \tag{2}$$

$$\mathbf{V} = \text{Transpose}(\text{Reshape}(\mathbf{F_2 W}^v)) \tag{3}$$

$$\begin{aligned} \mathbf{F}_{align} &= \text{Concat}(\text{head}_1, .., \text{head}_h)\mathbf{W}^O, \\ where \quad \text{head}_j &= \sigma\left(\frac{\mathbf{QK}^T}{\sqrt{C_c}}\right)\mathbf{V} \end{aligned} \tag{4}$$

where $\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v \in \mathbb{R}^{C\times C}$ are the learnable parameters of three linear projection layers and C is the channel dimension. $\mathbf{W}^O \in \mathbb{R}^{N_h C_c \times C}$ are the linear projection matrices. σ indicates the softmax operation.

Finally, $\mathbf{F}_{align}$ is restored to the size of $L \times C$ by transposing, reshaping, and linear transformation, then it is added to $\mathbf{F_1}$ by residual connection to carry out the next feature extraction stage.

C. Alignment Module based on Differential Flow Field

Inspired by the optical flow, we describe the field that illustrates the spatial offset between the bi-temporal features as the differential flow field. To address the issue of inadequate registration resulting from varying imaging perspectives, we introduce the alignment module based on the differential flow field, which facilitates the explicit acquisition of pixel-level offset between bi-temporal images, as shown in Fig. 3.

Taking the first stage as an example, after obtaining the bi-temporal features $\mathbf{F_1}$ and $\mathbf{F_2}$ (to simplify stage annotation), they are sent into the FlowMix module composed of three layers of convolution and GeLU to enhance the representation ability of features. Then, the enhanced features are joined together and sent into a sub-network composed of a convolution layer to generate the differential flow field $\boldsymbol{\Delta} f \in \mathbb{R}^{2\times H\times W}$. The process can be expressed as follows:

$$\mathbf{F}_1^{'} = \text{FlowMix}(\mathbf{F_1}) \tag{5}$$

$$\mathbf{F}_2^{'} = \text{FlowMix}(\mathbf{F_2}) \tag{6}$$

$$\boldsymbol{\Delta} f = \text{Conv}(\text{C}(\mathbf{F}_1^{'}, \mathbf{F}_2^{'})) \tag{7}$$

where C(· , ·) represents the concatenation operation and Conv is a convolution layer with the kernel size of 3. $\boldsymbol{\Delta} f$ stores the offset of each position in the T_1 phase (p_1) on the standard spatial grid. Then p_1 is added to $\boldsymbol{\Delta} f$ to get the coordinates of the wrapped point $\hat{p}$, and the missing points due to the position offset are completed by a differentiable bilinear sampling mechanism [63] to obtain the wrapped feature $\mathbf{F}_{warp}^{'}$, which enables explicit alignment at the pixel level. Finally, $\mathbf{F}_{warp}^{'}$ and $\mathbf{F}_2^{'}$ are absolute subtracted to obtain the bi-temporal differential feature $\mathbf{F}_{diff}$.

$$\mathbf{F}_{warp}^{'}(p_w) = \sum_{p\in N(\hat{p})} \mathbf{w_p}\mathbf{F_1}^{'}(p) \tag{8}$$

$$\mathbf{F}_{diff} = \text{S}(\mathbf{F}_{warp}^{'}, \mathbf{F_2}^{'}) \tag{9}$$

where p_w represents the position of the point in $\mathbf{F}_{warp}^{'}$, $N(\hat{p})$ means the four neighborhood position of $\hat{p}$ (top left, bottom left, top right, bottom right), $\mathbf{w_p}$ is bilinear kernel weights on warped spatial gird, and S(· , ·) represents the absolute subtracted.

D. Implicit Neural Alignment Decoder

In the task of CD, it is essential to integrate high-level semantic information and low-level detail features. Hence, most CD models employ bilinear upsampling or deconvolution to integrate feature maps with different resolutions into the same resolution to achieve multi-scale feature alignment. However, the utilization of bilinear upsampling may blur the precise information learned from feature maps, and deconvolution incurs additional computational expenses. Therefore, we introduce the implicit neural alignment decoder to realize the accurate alignment of multi-level differential feature maps in a lightweight way by learning the continuous representation of images in coordinate space.

Implicit neural representation defines a decoding function f_θ (typically an MLP) over a discrete feature map to get the continuous feature map $\mathbf{S}$. Taking a single feature map as an example, in the discrete feature mapping stage, each feature vector is regarded as an implicit code evenly distributed in 2D spaces, and a spatial coordinate is assigned to each vector. The feature of $\mathbf{S}$ at x_i is defined as:

$$\mathbf{S}(x_i) = f_\theta(\mathbf{z}^*, x_i - x^*) \tag{10}$$

where $\mathbf{z}^*$ means the nearest implicit code to x_i, x^* is the coordinate value of implicitly code $\mathbf{z}^*$. In addition, to further improve the learning ability of the decoding function, learnable position coding is introduced [57]. Therefore, the final definition of the implicit neural representation is:

$$\mathbf{S}(x_i) = f_\theta(\mathbf{z}^*, \beta(x_i - x^*), x_i - x^*) \tag{11}$$

where $\beta(x_i - x^*)$ represents the positional coding of the relative coordinates. To achieve the alignment of multilevel differential feature maps, a continuous map $\mathbf{S}$ is defined on multilevel discrete differential features with different resolutions. In particular, the value of $\mathbf{S}$ at x_i is defined as:

$$\mathbf{S}(x_i) = f_\theta(\{\mathbf{z}_{\mathbf{l}}^*\}_{l=1}^4, \{\beta(\Delta x_l)\}_{l=1}^4, \{\Delta x_l\}_{l=1}^4) \tag{12}$$

where l represents different levels, $\mathbf{z}_{\mathbf{l}}^*$ means the nearest implicit code of the L layer to x_i, Δx_l indicates relative position coordinates $x_i - x_l^*$, x_l^* is the coordinates of $\mathbf{z}_{\mathbf{l}}^*$.

As shown in Fig.1, for a query coordinate (red point), we obtain from each layer the nearest hidden code $\mathbf{z}_{\mathbf{l}}^*$ (green point), relative coordinate Δx_l, and the positional coding [57] of the relative coordinates $\beta(\Delta x_l)$. Then, they are joined and input into the decoding function f_θ, composed of three layers of MLP. The f_θ decodes the features of each stage and concurrently models the interconnections between various stages. In this way, multilevel differential maps can be aligned in a lightweight way to obtain high-quality prediction maps.

IV. Experimental Results and Analysis

A. Data description

Extensive experiments are conducted on six representative CD datasets to verify the practical performance of the proposed BiFA.

Wuhan University (WHU-CD) [64] is a building change detection dataset containing a pair of aerial images with a space size of 32507 × 15354 and the resolution of 0.2m/pixel. Since no segmentation strategy is provided in [64], we cut the image into a 256 × 256 size patch and randomly divided them into 6096/762/762 for training/validation/testing.

Learning, VIsion, and Remote sensing (LEVIR-CD) [1] and its recent extended version LEVIR+-CD are publicly available datasets for buildings CD, containing 637 and 985 pairs of 1024 × 1024 remote sensing images with a 0.5m/pixel resolution, respectively. Due to the limitation of GPU capacity, we divided the image into 256 × 256 patches in a non-overlapping manner. Specifically, LEVIR-CD utilizes 7120/1024/2048 patches for training/validation/testing, and LEVIR+-CD uses 10192/5568 patches for training/testing, respectively.

Sun Yat-sen University Dataset (SYSU-CD) [40] is a large dataset developed recently, including 20000 pairs of images with a 0.5m/pixel resolution and a size of 256 × 256. There are 8000/4000/8000 pairs of samples for training/validation/testing. Notably, the SYSU-CD dataset covers buildings and many other types of targets, such as ships, roads, and vegetation, which is a challenge for CD tasks.

Deeply supervised image fusion network (DSIFN-CD) [25] is a publicly binary CD dataset. It includes six pairs of 2m/pixel resolution images of six major cities in China. The dataset contains changes in various land cover objects, such as buildings, roads, farmland, etc. According to the method in [25], we cut the image into 512 × 512, with 3600/340/48 pairs of samples for training/validation/testing.

Cropland Change Detection dataset (CLCD) [65] consists of 600 pairs of farmland transformation samples with a size of 512 × 512 and a resolution of 0.5m/pixel-2m/pixel, including buildings, roads, lakes, and bare land. According to the method [65], we divided 320/120/120 pairs of samples for training/validation/testing.

B. Experimental setup

1) Architecture details: In the temporal alignment stage, we followed the setting of Segformer-B0 [62]. The four stages were downsampled to 1/4, 1/8, 1/16, and 1/32 of the original image size. The number of blocks in each stage N_i is set to 2, and the number of channels C_i is 32, 64, 160, and 256. The number of channels in the FlowMix middle layer is 4 × C_i. Before sending the multilevel differential features into the IND, the differential features at each stage are unified into 256 dimensions. The dimensions of the three layers in MLP that make up IND are set to 512, 256, and 256.

2) Training details: The proposed BiFA model is implemented based on the Pytorch framework and runs on an NVIDIA RTX 3090ti. For optimization, we use the Adam optimizer with an initial learning rate of 1e-4, β_1 and β_2 are 0.9, 0.999, respectively. The mini-batch size is set to 8. The total training epochs are 200. The loss function is an addition of the cross-entropy loss and the dice loss [66].

$$L_{total} = \lambda_1 L_{ce} + \lambda_2 L_{dice} \tag{13}$$

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N} y_i log(\hat{y_i}) \tag{14}$$

$$L_{dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y_i}}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y_i}} \tag{15}$$

where λ_1 and λ_2 denote the coefficients of loss function, y_i is the ground truth in the ith pixel, $\hat{y_i}$ represents the probability in the ith pixel. N indicates the number of pixels.

3) Evaluation metrics: For the performance measurement of similarity between the predictions and the ground truth, we introduce seven metrics, including Precision (Pre.), Recall (Rec.), F1-score (F1), Intersection over Union (IoU), overall accuracy (OA), false alarm (FA) and Kappa. The metrics can be individually defined as follows.

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1 = \frac{2}{Recall^{-1} + Precision^{-1}} \tag{18}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{19}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$FA = \frac{FP}{TN + FP} \tag{21}$$

$$Kappa = \frac{OA - P}{1 - P} \tag{22}$$

$$P = \frac{(TP + FP)(TP + FN) + (TN + FN)(TP + TN)}{(TP + FP + TN + FN)^2} \tag{23}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively. P in Kappa denotes the hypothetical probability of the chance agreement between reference and prediction. It is worth noting that F1 and IoU can better reflect the generalization ability of the model.

C. Performance comparison

To verify the validity of BiFA in the remote sensing CD task, in this section, several state-of-the-art models are chosen as the competitors, including three methods based on pure convolution (FC-EF [24], FC-Siam-Diff [24], FC-Siam-Conc [24]), two attention-based methods (IFNet [25] and SNUNet [23]), two methods based on pure Transformer (SwinUnet [44], Changeformer [67]), and some approaches combining CNN and Transformer simultaneously (BIT [34], MSCANet [65], Paformer [29], DARNet [33], ACABFNet [27], and DMINet [28]).

1) FC-EF [24]: is a CD model based on the U-net [21], which takes the concatenation of bi-temporal images along the channel dimension as the input to the model.
2) FC-Siam-Diff [24]: is a variant of FC-EF that employs a weight-shared Siamese architecture to extract multi-level features, facilitating the fusion of bi-temporal information through feature subtraction.
3) FC-Siam-Conc [24]: represents a variant of FC-EF, which utilizes a weight-shared Siamese architecture to acquire multi-level features and integrate bi-temporal information through feature concatenation.
4) IFNet [25]: utilizes a weight-shared VGG-16 [68] to extract multi-level features, integrating bi-temporal information through a concatenation approach. Additionally, spatial attention and channel attention are applied at each stage of the decoder. Furthermore, deep supervision, such as calculating supervised loss for each decoder level, is employed to enhance the training of intermediate layers.
5) SNUNet [23]: employs a weight-shared NestedUNet [69] to extract multi-level features, with channel attention applied to features at various levels of the decoding stage. Furthermore, deep supervision is also utilized to enhance the training of intermediate layers.
6) SwinUnet [44]: uses a weight-shared SwinTransformer [70] for the extraction of multi-level features, whereby the features from the final layer are merged through concatenation to integrate bi-temporal information before being fed into the decoder. Simultaneously, an Unet-like connection is established, facilitating the fusion of the extracted multi-level features with those at the corresponding decoder layer through concatenation, further enhanced through channel attention mechanisms.
7) Changeformer [67]: utilizes the weight-shared Segformer-B1 to extract multi-level features. For each level, the features undergo differencing and convolution operations to facilitate bi-temporal feature fusion. Subsequently, the differential features from various levels are concatenated and input into a decoder comprised of fully connected layers to achieve change detection.
8) BIT [34]: employs weight-shared ResNet18 to extract bi-temporal features and uses the Semantic Tokenizer to condense them into a reduced set of semantic tokens. Subsequently, different tokens are concatenated and input into a native Transformer Encoder/Decoder to learn spatial-temporal relationships.
9) MSCANet [65]: utilizes weight-shared ResNet18 for multi-level feature extraction. It integrates spatial attention and Transformer to establish multi-level spatial-temporal relationships, with inter-level information fusion achieved through concatenation. Finally, three decoders are employed to process different-level features constructing a multi-scale supervision framework.
10) Paformer [29]: employs weight-shared ResNet18 to extract bi-temporal shallow features and deep semantic features. The bi-temporal shallow features are separately fed into a prior interpreter composed of a flow field to enhance the structural priors of a specific temporal, followed by fusion through addition. Subsequently, the fused shallow features and fused deep features by addition are jointly input into a Transformer Decoder for change detection with structural prior awareness.
11) DARNet [33]: uses weight-shared convolutional networks to extract multi-level features. Features from different levels are utilized in conjunction with Transformers and channel attention mechanisms to learn spatial-temporal correlations. Additionally, features from different levels undergo dense fusion with each other. Ultimately, detection results are obtained through the deep supervision approach.
12) ACABFNet [27]: employs different branches (ResNet34 and Transformer) for the extraction of multi-level features. Simultaneously, axial attention is employed to fuse the height (H) or width (W) dimensions of features with the channel (C) dimension, enabling efficient learning of spatial-temporal relationships in bi-temporal features.
13) DMINet [28]: utilizes weight-shared ResNet18 to extract multi-level features. Subsequently, these features are concatenated along the channel dimension and fed into a Transformer, utilizing shared queries, to learn multi-level spatial-temporal relationships. Finally, different-level feature decoders are utilized to construct the multi-level supervision mechanism.
14) Baseline: uses weight-shared Segformer-B0 to extract multi-level features. Then, these features are processed by absolute subtraction to generate multi-level differential features. Finally, the multi-level differential features are unified through concatenation and fed into the decoder composed of fully connected layers to achieve change detection.

TABLE I
COMPARISON RESULTS ON THE THREE CD TEST SETS. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, BLUE. ALL RESULTS ARE DESCRIBED IN PERCENTAGE (%).

	Backbone	WHU-CD	LEVIR-CD	LEVIR+-CD
		Pre. / Rec. / F1 / IoU / OA	Pre. / Rec. / F1 / IoU / OA	Pre. / Rec. / F1 / IoU / OA
FC-EF$_{18}$ [24]	UNet	92.10 / 90.64 / 91.36 / 84.10 / 99.32	90.64 / 87.23 / 88.90 / 80.03 / 98.89	76.49 / 76.32 / 76.41 / 61.82 / 98.08
FC-Siam-Diff$_{18}$ [24]	UNet	87.39 / **92.36** / 89.81 / 81.50 / 99.16	90.81 / 88.59 / 89.69 / 81.31 / 98.96	80.88 / 77.65 / 79.23 / 65.61 / 98.34
FC-Siam-Conc$_{18}$ [24]	UNet	86.57 / 91.11 / 88.78 / 79.83 / 99.08	91.41 / 88.43 / 89.89 / 81.64 / 98.98	81.12 / 77.16 / 79.09 / 65.42 / 98.33
IFNet$_{20}$ [25]	VGG16	91.51 / 88.01 / 89.73 / 81.37 / 99.20	89.62 / 86.65 / 88.11 / 78.75 / 98.81	**81.79** / 78.40 / 80.06 / 66.76 / 98.41
SNUNet$_{21}$ [23]	NestedUNet	84.70 / 89.73 / 87.14 / 77.22 / 98.95	89.73 / 87.47 / 88.59 / 79.51 / 98.85	78.90 / 78.23 / 78.56 / 64.70 / 98.26
SwinUnet$_{22}$ [44]	Swin-T	92.44 / 87.56 / 89.93 / 81.71 / 99.22	89.11 / 86.47 / 87.77 / 78.21 / 98.77	77.65 / 78.98 / 78.31 / 64.35 / 98.22
BIT$_{22}$ [34]	ResNet18	91.84 / 91.95 / 91.90 / 85.01 / 99.35	**92.07** / 88.08 / 90.03 / 81.87 / 99.01	80.50 / 81.41 / 80.95 / 68.00 / 98.43
ChangeFormer$_{22}$ [67]	SegformerB1	93.73 / 87.11 / 90.30 / 82.32 / 99.26	90.68 / 87.04 / 88.83 / 79.90 / 98.88	77.32 / 77.75 / 77.54 / 63.31 / 98.16
MSCANet$_{22}$ [65]	ResNet18	93.47 / 89.16 / 91.27 / 83.94 / 99.32	90.02 / 88.71 / 89.36 / 80.77 / 98.92	76.92 / **83.69** / 80.16 / 66.89 / 98.31
Paformer$_{22}$ [29]	ResNet18	**94.28** / 90.38 / 92.29 / 85.69 / 99.40	91.34 / 88.07 / 89.68 / 81.29 / 98.96	79.89 / **82.96** / **81.40** / **68.63** / 98.45
DARNet$_{22}$ [33]	-	91.99 / 91.17 / 91.58 / 84.46 / 99.33	**92.19** / **88.99** / **90.56** / **82.76** / **99.05**	77.84 / 78.42 / 78.13 / 64.11 / 98.21
ACABFNet$_{23}$ [27]	ResNet34	91.57 / 90.86 / 91.21 / 83.84 / 99.31	90.11 / 88.27 / 89.18 / 80.48 / 98.91	72.85 / 80.91 / 76.67 / 62.17 / 97.99
DMINet$_{23}$ [28]	ResNet18	**94.89** / **92.02** / **93.43** / **87.68** / **99.48**	**92.16** / **88.83** / **90.46** / **82.59** / **99.04**	81.61 / 80.91 / 81.26 / 68.44 / **98.48**
Baseline	SegformerB0	94.19 / 91.15 / **92.64** / **86.30** / **99.42**	90.75 / 88.54 / 89.63 / 81.21 / 98.95	**84.63** / 78.65 / **81.53** / **68.82** / **98.54**
BiFA	SegformerB0	**95.15** / **93.60** / **94.37** / **89.34** / **99.56**	91.52 / **89.86** / **90.69** / **82.96** / **99.06**	**83.85** / **84.06** / **83.96** / **72.35** / **98.69**

TABLE II
COMPARISON RESULTS ON THE THREE CD TEST SETS. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, BLUE. ALL RESULTS ARE DESCRIBED IN PERCENTAGE (%).

	Backbone	SYSU-CD	DSIFN-CD	CLCD-CD
		Pre. / Rec. / F1 / IoU / OA	Pre. / Rec. / F1 / IoU / OA	Pre. / Rec. / F1 / IoU / OA
FC-EF$_{18}$ [24]	UNet	78.94 / 78.58 / 78.76 / 64.96 / 90.01	63.84 / 68.83 / 66.24 / 49.52 / 88.34	65.83 / 66.50 / 66.16 / 49.43 / 94.93
FC-Siam-Diff$_{18}$ [24]	UNet	79.82 / 78.75 / 79.28 / 65.68 / 90.29	57.17 / 75.11 / 64.92 / 48.06 / 86.51	66.37 / 65.59 / 65.98 / 49.23 / 94.96
FC-Siam-Conc$_{18}$ [24]	UNet	80.15 / 77.39 / 78.75 / 64.95 / 90.15	52.07 / **76.14** / 61.85 / 44.77 / 84.38	71.30 / 62.45 / 66.58 / 49.91 / 95.33
IFNet$_{20}$ [25]	VGG16	**85.46** / 75.60 / 80.23 / 66.98 / 91.21	62.40 / 73.90 / 67.67 / 51.13 / 88.26	59.08 / 66.39 / 62.52 / 45.48 / 94.08
SNUNet$_{21}$ [23]	NestedUNet	**90.12** / 58.01 / 70.59 / 54.54 / 88.60	70.95 / 66.01 / 68.39 / 51.96 / 89.86	71.42 / 53.78 / 61.36 / 44.25 / 94.96
SwinUnet$_{22}$ [44]	Swin-T	82.31 / 73.73 / 77.78 / 63.64 / 90.07	63.47 / 66.92 / 65.15 / 48.31 / 88.10	70.66 / 59.19 / 64.42 / 47.52 / 95.13
BIT$_{22}$ [34]	ResNet18	76.88 / **80.68** / 78.73 / 64.93 / 89.72	63.14 / 68.27 / 65.61 / 48.81 / 88.11	72.47 / 66.93 / 69.59 / 53.36 / 95.64
ChangeFormer$_{22}$[67]	SegformerB1	77.29 / 78.75 / 78.01 / 63.95 / 89.53	**71.08** / 64.01 / 67.36 / 50.78 / 89.69	72.37 / 63.12 / 67.43 / 50.87 / 95.46
MSCANet$_{22}$ [65]	ResNet18	81.75 / 74.41 / 77.91 / 63.81 / 90.04	55.93 / **76.42** / 64.59 / 47.70 / 86.07	67.09 / 67.91 / 67.50 / 50.94 / 95.13
Paformer$_{22}$ [29]	ResNet18	84.84 / 72.51 / 78.19 / 64.20 / 90.46	60.84 / 70.85 / 65.47 / 48.66 / 87.57	76.20 / 60.13 / 67.22 / 50.63 / 95.63
DARNet$_{22}$ [33]	-	83.33 / 78.72 / 80.96 / 68.01 / 91.27	67.53 / 70.23 / **68.85** / **52.50** / 89.44	68.73 / 62.75 / 65.60 / 48.81 / 95.10
ACABFNet$_{23}$ [27]	ResNet34	83.78 / 78.94 / 81.29 / 68.48 / 91.42	70.40 / 71.96 / **71.17** / **55.25** / **90.31**	**78.16** / 64.25 / 70.53 / 54.47 / **96.00**
DMINet$_{23}$ [28]	ResNet18	84.99 / **79.41** / **82.11** / **69.65** / **91.84**	60.17 / **78.51** / 68.13 / 51.66 / 87.79	73.24 / **70.11** / **71.64** / **55.81** / 95.87
Baseline	SegformerB0	85.42 / 78.78 / **81.97** / **69.45** / **91.82**	**71.65** / 64.99 / 68.16 / 51.70 / **89.91**	**81.12** / **74.96** / **77.92** / **63.83** / **96.83**
BiFA	SegformerB0	**86.98** / **81.44** / **84.12** / **72.59** / **92.75**	**73.99** / 68.87 / **71.34** / **55.45** / **90.80**	**84.02** / **74.98** / **79.23** / **65.61** / **97.08**

For a fair comparison, all methods are trained under the same conditions based on the officially published Pytorch code.

1) Quantitative results: Numerically, Table I and Table II show the overall performance of all methods on the WHU-CD, LEVIR-CD, LEVIR+-CD, SYSU-CD, DSIFN-CD, and CLCD-CD test sets. Bold numbers represent the best results. Evidently, the BiFA exhibits superior performance in comparison to alternative methods. In contrast to the pure Transformer-based methods, SwinUnet and Changeformer, BiFA demonstrates substantial superiority outcomes across all six datasets. Furthermore, the F1-score of BiFA surpassed the most recent DMINet by varying percentages across different datasets. Specifically, on the WHU-CD, LEVIR+-CD, SYSU-CD, and DSIFN-CD datasets, the F1-score of BiFA exceeded that of DMINet by 0.94%, 3.21%, 2.01%, and 2.7%, respectively. Notably, on the CLCD-CD dataset, the F1-score of BiFA even exceeded that of DMINet by 7.59%. While BiFA does not exhibit a significant improvement over DMINet on the LEVIR-CD dataset, it does attain satisfactory outcomes. The results indicate that while BiFA's Rec. metric on the DSIFN-CD is lower than that of DMINet, the former outperforms the latter in other metrics, suggesting its superior accuracy in detecting changes in the area. Similarly, it is noteworthy that on the SYSU-CD dataset, while the precision metric of BiFA is comparatively lower than that of SNUNet, BiFA outperformed SNUNet in other metrics. This suggests that BiFA is more comprehensive in detecting changing regions. In summary, the above quantitative analysis proves that BiFA implementing bi-temporal feature alignment at the temporal (channel, spatial), and multi-scale levels is essential for change detection tasks.

2) Qualitative results: To further illustrate the validity of our proposed method, qualitative analyses are conducted on WHU-CD, LEVIR-CD, LEVIR+-CD, SYSU-CD, DSIFN-CD, and CLCD-CD test sets (Fig. 4-9), where distinct colors are assigned to identify the correctness or incorrectness of the detection, including TP (white), TN (black), FP (red) and FN (green).

Visualization on WHU-CD (Fig. 4): Several representative samples are selected for visualization comparison, such as the surface disturbance caused by the cement surface covering in Fig. 4(a), the densely distributed small buildings in Fig. 4(b), the offset of buildings affected by the different perspectives in Fig. 4(c) and the intense illumination variation in Fig. 4(d). And the superiority of our BiFA over other competitors is evident from the data presented in Fig. 4(a). While the ACABFNet exhibits a satisfactory detection effect, it reveals more conspicuous omissions compared with our method. The results obtained from Fig. 4(b) demonstrate that our BiFA can detect changes in scenes with complex structures, as evidenced by the reduced occurrence of omission faults and commission errors. Moreover, comparing the false identifications of each model in Fig. 4(c) and (d), our BiFA attains a surprisingly precise discriminability despite the interference of differences in perspective and illumination. This verifies the advantage of BiFA for bi-temporal feature alignment at the channel and spatial levels. Shortly, the qualitative analysis results are consistent with the quantitative analysis in Table I, and the proposed BiFA achieves state-of-the-art performance on WHU-CD.

Visualization on LEVIR-CD (Fig. 5): A similar scheme is utilized to compare the LEVIR-CD, and several representative examples are selected. Fig. 5(a) represents the changing scene of large buildings, Fig. 5(b) represents the scene of dense small buildings, and Fig. 5(c) and (d), respectively, represent the influence of perspective difference and illumination intensity. Fig. 5(a) shows that BiFA achieved far better results than other methods in detecting large irregular buildings. Similarly, our method achieves good results in the complex scene of dense small buildings in Fig. 5(b). This is because the IND, which learns the continuous representation of the image in coordinate space, precisely aligns the multi-level differential features. By looking at the results in Fig. 5(c), (d), it can be seen that BiFA is more robust to perspective difference and illumination distinctions.

Visualization on LEVIR+-CD (Fig. 6): In the LEVIR+-CD, we also selected several representative samples. Fig. 6(a) shows the changes in large buildings, Fig. 6(b) shows the scene of dense small buildings, Fig. 6(c) and (d) respectively show the difference in perspective and the influence of illumination. As can be seen, our BiFA achieved better results in all cases, especially in small buildings, where our approach has less connectivity.

Visualization on SYSU-CD (Fig. 7): It is worth noting that the above three datasets focus on building change. To further verify the detection performance for various category changes, we validated the BiFA on SYSU-CD, which has a broader category. As can be seen, the BiFA has an outstanding visualization effect. For example, in the large area of surface vegetation change in Fig. 7(c), our method has almost achieved the same visual effect as the ground truth. Regarding apparent color difference interference (Fig. 7(d)), BiFA has far superior performance compared with BIT, DMINet, and other methods. This further demonstrates the robustness of BiFA against independent interference.

Visualization on DSIFN-CD (Fig. 8): We also extend the proposed BiFA to a broader range of change scenarios. Fig. 8(a), (b) represents the changes in large objects, and Fig. 8(c), (d) represents the changes in small objects. It is difficult to distinguish different types of changes in these complex scenarios. Encouragingly, even though the results are limited relative to other datasets, our BiFA has the best visuals on the DSIFN-CD.

TABLE III

COMPARISON RESULTS ON MODEL EFFICIENCY. WE REPORT THE NUMBER OF PARAMETERS (PARAMS.), FLOATING-POINT OPERATIONS PER SECOND (FLOPS), AND THE TRAINING TIME (TIME) FOR A SINGLE EPOCH ON LEVIR+-CD. THE SIZE OF THE INPUT IMAGE TO THE MODEL IS $256 \times 256 \times 3$ TO CALCULATE THE FLOPS.

Model	Params. (M)	FLOPs (G)	Time (Min)
FC-EF$_{18}$	1.35	3.57	1.61
FC-Siam-Diff$_{18}$	1.34	4.72	1.50
FC-Siam-Conc$_{18}$	1.54	5.33	1.77
IFNet$_{20}$	50.71	41.1	5.02
SNUNet$_{21}$	1.35	4.72	1.65
SwinUnet$_{22}$	30.28	11.83	3.01
BIT$_{22}$	3.04	8.75	2.8
Changeformer$_{22}$	41.02	202.78	20.45
MSCANet$_{22}$	16.42	14.80	6.28
Paformer$_{22}$	16.13	10.85	2.12
DARNet$_{22}$	15.09	64.48	12.53
ACABFNet$_{23}$	102.32	28.28	5.12
DMINet$_{23}$	6.24	14.55	3.53
BiFA	5.58	53.00	12.73

Visualization on CLCD-CD (Fig. 9): We have also extended BiFA to the vast range of change scenarios CLCD-CD. As shown in Fig. 9, BiFA can achieve good results in large-area road changes (Fig. 9(a)) and rural road changes(Fig. 9(c)). And, BiFA can well identify woodland changes even in the presence of significant illumination differences (Fig. 9(d)).

3) Model efficiency: To further verify the efficiency of the proposed model, the model parameters (Params.), floating-point operations per second (FLOPs), and the training time (Time) for a single epoch on LEVIR+-CD are presented in Table III, where the input size is 256 × 256 × 3. Compared to DARNet, our BiFA structure is lightweight and delicately designed, which is superior to DARNet in terms of efficiency and parameters. Compared with most other state-of-the-art methods, BiFA has a small number of parameters. Notwithstanding, since each decoding stage is queried on the entire map (256 × 256), the computational demands are significant, resulting in higher FLOPs and training time than alternative approaches. This presents a space for potential improvement in future research.

D. Ablation studies

To further verify the influence of each module on the performance of CD, ablation experiments are conducted in this section, as shown in Table IV-VII.

1) Effects of Different Components in BiFA: To verify the validity of key modules in the proposed BiFA, eight ablation experiments are designed, and the Segformer-B0 is selected to build the baseline. The experimental results, as shown in Table IV, whether each critical module is added separately or in different combinations, the experimental results are superior to the baseline. The F1/IoU on WHU-CD, LEVIR-CD, and DSIFN-CD increased by 1.72%/3.04%, 1.06%/1.75%, and 3.18%/3.75%, respectively. These improvements indicate that

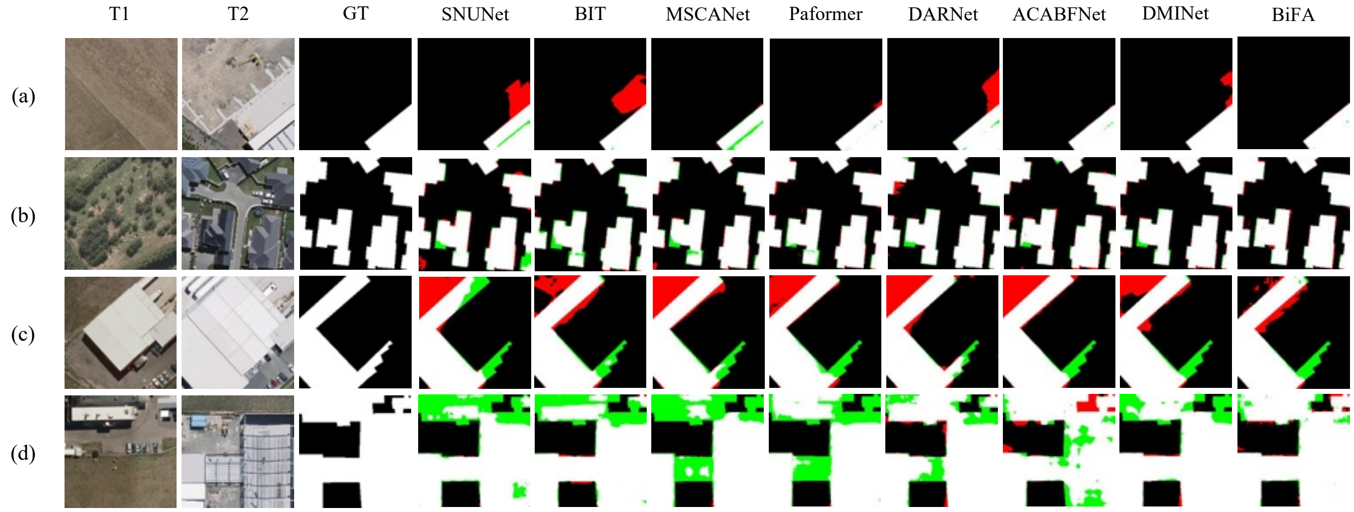

Fig. 4. Visualization results of different methods on the WHU-CD test set. (a)-(d) are representative samples. White represents a true positive, black is a true negative, red indicates a false positive, and green stands as a false negative.

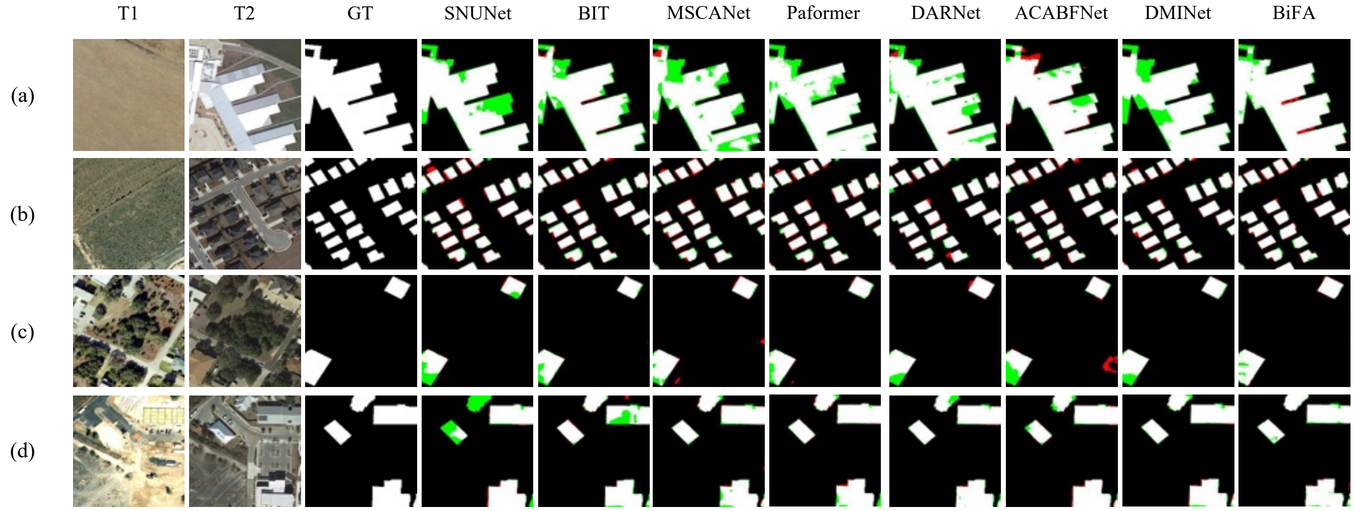

Fig. 5. Visualization results of different methods on the LEVIR-CD test set. (a)-(d) are representative samples. White represents a true positive, black is a true negative, red indicates a false positive, and green stands as a false negative.

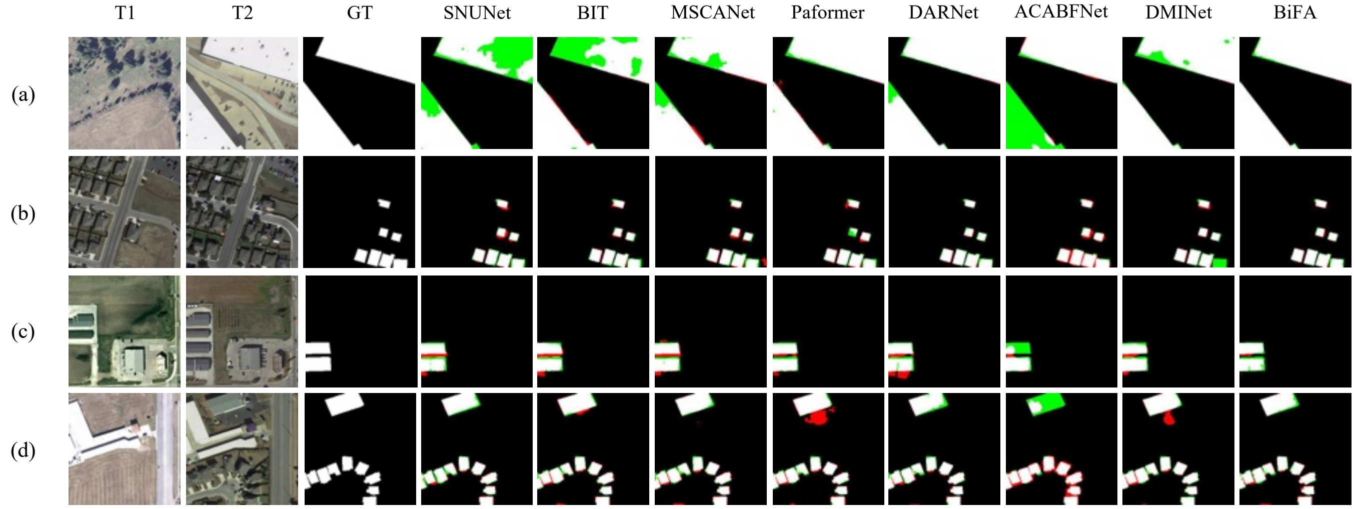

Fig. 6. Visualization results of different methods on the LEVIR+-CD test set. (a)-(d) are representative samples. White represents a true positive, black is a true negative, red indicates a false positive, and green stands as a false negative.

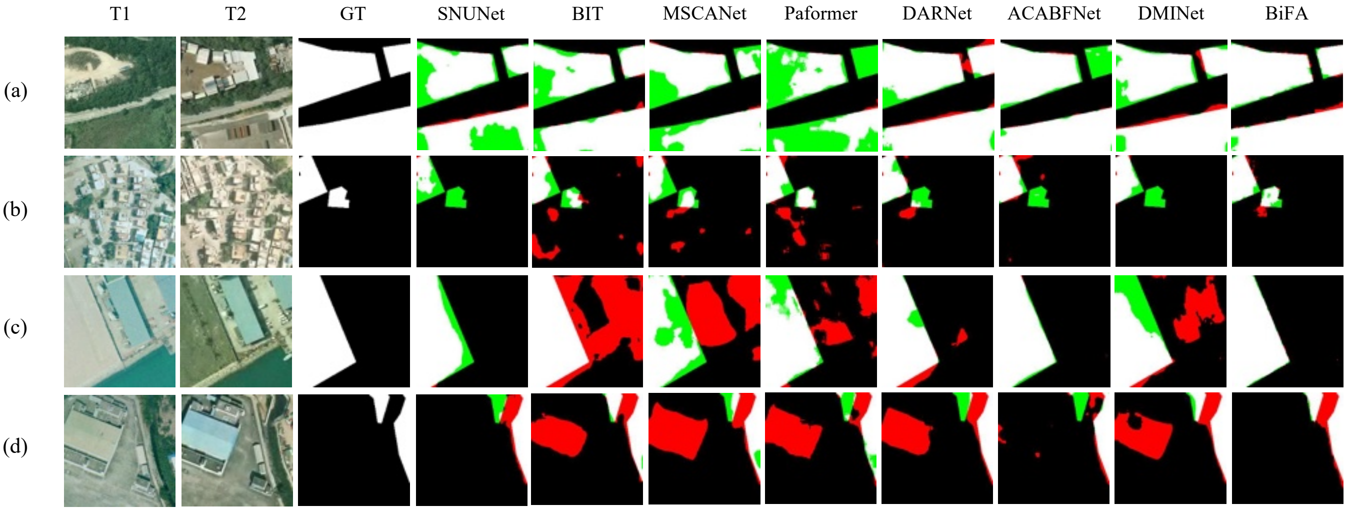

Fig. 7. Visualization results of different methods on the SYSU-CD test set. (a)-(d) are representative samples. White represents a true positive, black is a true negative, red indicates a false positive, and green stands as a false negative.

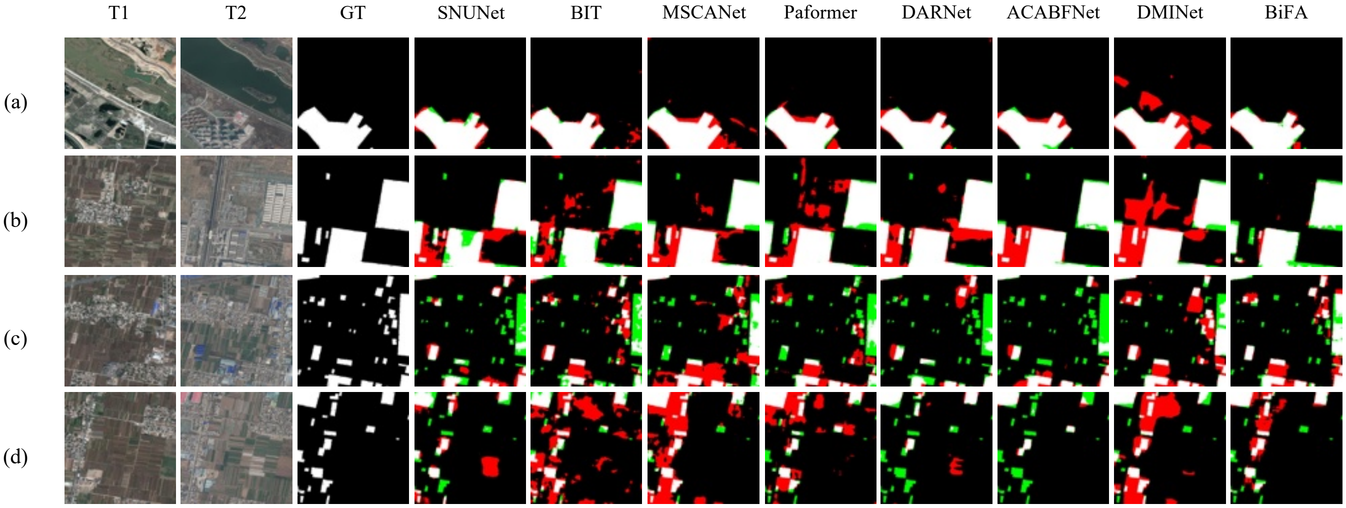

Fig. 8. Visualization results of different methods on the DSIFN-CD test set. (a)-(d) are representative samples. White represents a true positive, black is a true negative, red indicates a false positive, and green stands as a false negative.

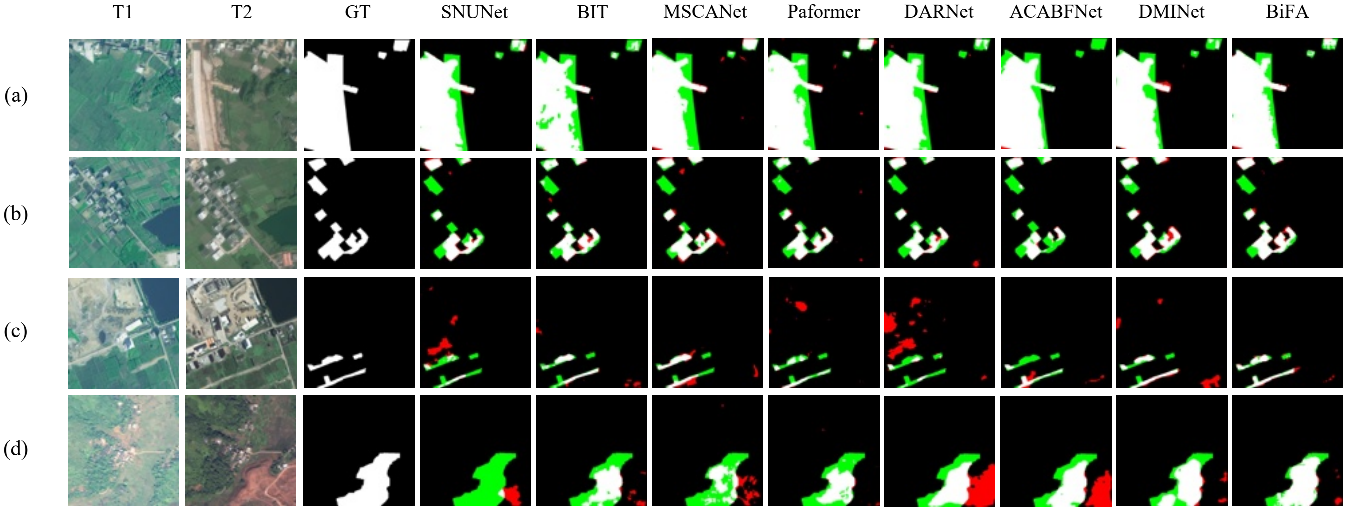

Fig. 9. Visualization results of different methods on the CLCD-CD test set. (a)-(d) are representative samples. White represents a true positive, black is a true negative, red indicates a false positive, and green stands as a false negative.

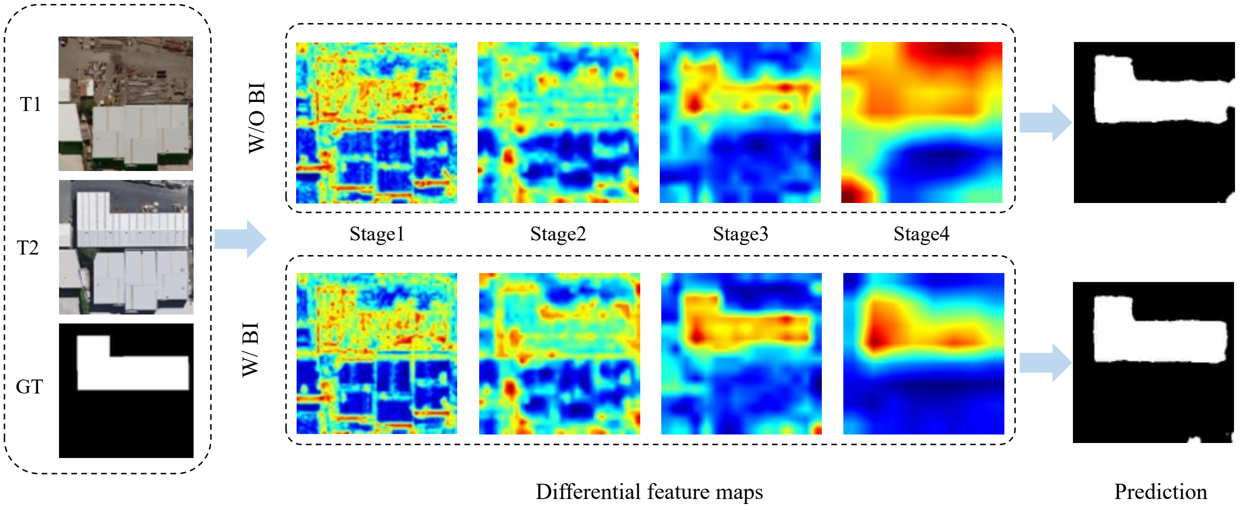

Fig. 10. Visualization results of the differential feature maps on the WHU-CD test set. W/O BI is the baseline model, and W/ BI represents the baseline model adding BI. Red denotes higher attention values, and blue denotes lower values.

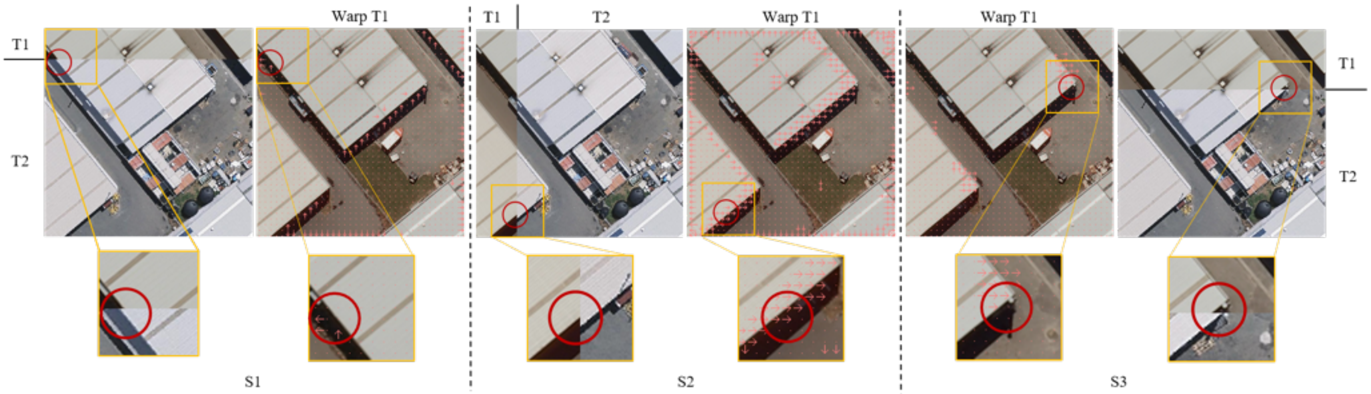

Fig. 11. Visualization results of position offset from T1 to T2 on the WHU-CD test set. S1, S2, and S3 represent different stages. Warp T1 indicates the offset of the T1 image. The red arrow indicates the offset direction.

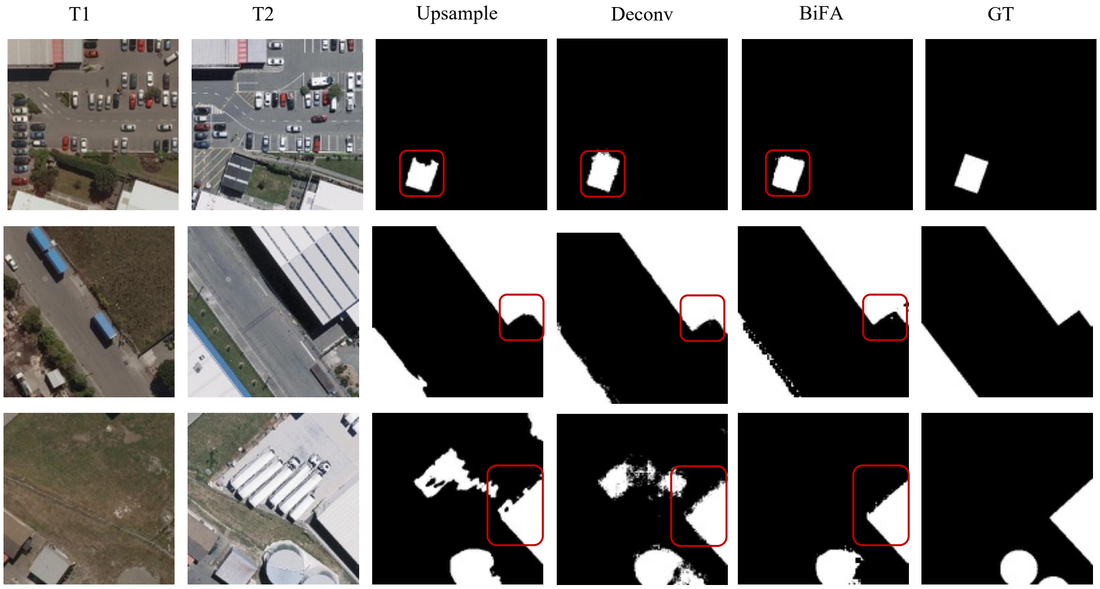

Fig. 12. Visualization results of prediction results. T1 and T2 mean bi-temporal images on the WHU-CD test set. Upsampel represents the method utilizing bilinear interpolation. Deconv is deconvolution. BiFA is our method. GT means the ground truth. Red squares denote the challenging regions that can be resolved by our proposed BiFA.

BiFA can reasonably apply to CD tasks by implementing bi-temporal feature alignment at the temporal (channel, spatial), and multi-scale levels.

2) Effects of different stages of BI: To further study the influence of BI at different stages, we conducted the experiment as shown in Table V. BiFA w/o BI represents the BiFA model without the BI, and S1-S4 represents the four feature extraction stages. With the gradual addition of the BI in different stages, the performance of the model (e.g., F1, IoU, and Kappa) generally increases. This indicates that the channel-level alignment of the bi-temporal features at different stages can effectively alleviate the interference caused by irrelevant factors such as illumination. To provide a more intuitive explanation and eliminate the interference of other modules, we visualized the differential features between baseline (w/o BI) and baseline-added BI (w BI) in the four stages on the test set of WHU-CD. As shown in Fig. 10, adding BI modules at both shallow and high levels obtains better visualization. Specifically, adding BI to shallow layers (such as S1) produces fewer noise points on the building and more apparent structural details. After BI is added in the deep layers (such as S3 and S4), the model pays more attention to changing regions, thus obtaining more accurate prediction results.

3) Effects of different stages of ADFF: In addition, we also studied the influence of ADFF at different stages, as shown in Table VI. The BiFA w/o ADFF represents the BiFA without the ADFF module, and S1-S4 represents the four stages. It can be seen that the performance of the model (e.g., F1, IoU, and Kappa) is gradually improved when ADFF is added in the first three stages. However, when ADFF is added in the fourth stage, the performance of the model will decline. This can be interpreted as the abstract semantics of the feature map in the fourth stage, which makes it difficult to learn effective position offset information and affects the performance of the model. In order to show the position offset intuitively, we visualized the offset image on the test set of WHU-CD. As shown in Fig. 11, the red circle represents the unregistered phenomenon of pixel level in the bi-temporal image, and the arrow in the red circle means the direction of pixel offset in the T1 image, and the length represents the offset. Obviously, the unaligned region in T1 will shift toward T2 after being sent to the ADFF, which is consistent with our expectations.

4) Effects of IND: We also compare IND with standard upsampling methods, such as bilinear interpolation and deconvolution. The Base indicates utilizing the head that is used in Segformer-B0. Upsample and Deconv indicate utilizing the same structure as IND, respectively, except that the upsampling method is replaced with bilinear interpolation and deconvolution. Specifically, the Upsample approach unifies the feature maps from the four stages into 256 dimensionalities, subsequently transforming to the original image size by bilinear interpolation. Ultimately, these resized feature maps are fed into the decoder with the same structure as IND for predicting. Similarly, the Deconv method replaces bilinear interpolation with deconvolution whereby the deconvolution kernel/step size is set as 6/4, 12/8, 24/16, and 48/32, respectively, to handle feature maps across different stages. As shown in Table VII, compared with Upsample, IND obtains better results with a slight increase in the number of parameters (for example, F1/IoU respectively increased by 0.43%/0.77% on the WHU-CD). And there are varying degrees of improvement over Deconv and Base. This implies that IND is a lighter and more effective approach. We further provide comparisons of visualization results on the test set of WHU-CD in Fig. 12. The BiFA can produce more precise boundaries by effectively aligning low-level and high-level feature maps. Particularly, the FLOPs of BiFA increase significantly compared with the Base, which we will improve in the future.

5) Effects of Position Encoding: Some experiments are conducted to explore the influence of different position encodings on IND, as shown in Table VIII. When the BiFA adds position encoding, F1/IoU indicators improve on both datasets. In particular, adding learnable encoding receives the most significant improvement, so we select learnable position encoding in the BiFA.

E. Parameter Analysis

1) Number of Head: To investigate the impact of varying the number of heads within the BI module on model performance, we conducted experiments using different configurations of the BiFA on the WHU-CD and LEVIR-CD datasets. The experimental results are shown in Table IX. As the number of heads gradually increases, the model's performance across both datasets exhibits an ascending trend, with optimal results attained at the number of 8. This phenomenon stems from different heads focusing on different aspects, thereby endowing the model with a more comprehensive focus and consequent performance enhancement. However, as the number of heads continues to mount, the performance tends to decline, attributed to the exacerbation of information redundancy resulting from excessive attentional different aspects. Consequently, the number of heads in BI is set to 8.

2) Kernel Size: To explore the influence of varying convolutional kernel sizes within the ADFF on model performance, we conducted a series of experiments on both the WHU-CD and LEVIR-CD datasets. It can be observed from Table X that the model achieves the best result when the convolution kernel size increases to 3. This result can be attributed to the increase in the receptive field, which can make the model learn the local offset between the bi-temporal images combined with the neighborhood information. However, as the convolutional kernel size continues to increase, a declining trend in model performance begins to appear. This phenomenon is due to the inundation of crucial local information by an excessive inflow of neighboring contextual details with the expansion of the receptive field, which makes it difficult for the model to learn the local offset between the bi-temporal images. So, the size of the convolution kernel in ADFF is set to 3.

3) Coefficients of Loss Function: To validate the impact of various loss function coefficients on model performance, we conducted corresponding experiments on the WHU-CD and LEVIR-CD datasets. The experimental results are presented in Table XI. For the WHU-CD dataset, optimal performance is achieved when the coefficients of cross-entropy loss (λ_1) and dice loss (λ_2) are relatively balanced. As for the LEVIR-CD dataset, the results exhibit relative stability across different

TABLE IV
ABLATION STUDY ON DIFFERENT COMPONENTS.

Model	WHU-CD Pre. / Rec. / F1 / IoU / OA / Kappa / FA	LEVIR-CD Pre. / Rec. / F1 / IoU / OA / Kappa / FA	DSIFN-CD Pre. / Rec. / F1 / IoU / OA / Kappa / FA
Baseline	94.19 / 91.15 / 92.64 / 86.30 / 99.42 / 92.35 / 0.23	90.75 / 88.54 / 89.63 / 81.21 / 98.95 / 89.09 / 0.48	71.65 / 64.99 / 68.16 / 51.70 / 89.91 / 62.18 / 5.12
w/ BI	92.12 / **94.22** / 93.16 / 87.20 / 99.45 / 92.87 / 0.33	91.59 / 88.55 / 90.05 / 81.90 / 99.00 / 89.52 / 0.44	67.79 / 71.71 / 69.65 / 53.43 / 89.63 / 63.35 / 6.36
w/ ADFF	95.75 / 90.85 / 93.24 / 87.32 / 99.48 / 92.96 / 0.17	**91.83** / 88.24 / 90.01 / 81.81 / 99.01 / 89.47 / 0.43	69.86 / 68.28 / 69.07 / 52.75 / 89.83 / 62.98 / 5.87
w/ IND	95.57 / 91.19 / 93.34 / 87.50 / 99.48 / 93.07 / 0.18	91.15 / 88.94 / 90.04 / 81.88 / 98.99 / 89.51 / 0.47	68.56 / 72.14 / 70.31 / 54.22 / 89.87 / 64.21 / 6.59
w/ BI+IND	94.43 / 91.99 / 93.19 / 87.26 / 99.46 / 92.92 / 0.23	91.08 / 89.47 / 90.27 / 82.27 / 99.02 / 89.76 / 0.47	74.05 / 67.57 / 70.66 / 54.64 / 90.67 / 65.14 / 4.72
w/ BI+ADFF	95.18 / 92.69 / 93.92 / 88.54 / 99.52 / 93.68 / 0.19	91.09 / 89.35 / 90.22 / 82.17 / 99.01 / 89.69 / 0.47	**78.32** / 64.74 / 70.89 / 54.91 / **91.16** / 65.74 / **3.57**
w/ ADFF+IND	**96.19** / 91.70 / 93.84 / 88.49 / 99.52 / 93.64 / **0.15**	90.87 / **90.06** / 90.42 / 82.52 / 99.03 / 89.96 / 0.49	68.49 / **72.59** / 70.49 / 54.42 / 89.89 / 64.40 / 6.66
BiFA	95.15 / 93.60 / **94.37** / **89.34** / **99.56** / **94.14** / 0.19	91.52 / 89.86 / **90.69** / **82.96** / **99.06** / **90.19** / **0.43**	73.99 / 68.87 / **71.34** / **55.45** / 90.80 / **65.87** / 4.82

TABLE V
ABLATION STUDY ON DIFFERENT STAGES OF BI. AND S STANDS FOR THE STAGE.

Model	S1	S2	S3	S4	WHU-CD Pre. / Rec. / F1 / IoU / OA / Kappa / FA	LEVIR-CD Pre. / Rec. / F1 / IoU / OA / Kappa / FA
BiFA w/o BI	×	×	×	×	**96.19** / 91.70 / 93.84 / 88.49 / 99.52 / 93.64 / **0.15**	90.87 / **90.06** / 90.42 / 82.52 / 99.03 / 89.96 / 0.49
BiFA	✓	×	×	×	95.81 / 92.46 / 94.10 / 88.87 / 99.54 / 93.87 / 0.17	92.01 / 88.89 / 90.47 / 82.59 / 99.04 / 89.99 / 0.45
BiFA	✓	✓	×	×	92.66 / **95.90** / 94.26 / 89.14 / 99.55 / 94.02 / 0.16	91.63 / 89.53 / 90.57 / 82.77 / 99.05 / 90.07 / 0.44
BiFA	✓	✓	✓	×	96.11 / 92.32 / 94.18 / 89.01 / 99.55 / 93.95 / 0.16	**92.19** / 89.10 / 90.62 / 82.85 / 99.06 / 90.12 / **0.41**
BiFA	✓	✓	✓	✓	95.15 / 93.60 / **94.37** / **89.34** / **99.56** / **94.14** / 0.19	91.52 / 89.86 / **90.69** / **82.96** / **99.06** / **90.19** / 0.43

TABLE VI
ABLATION STUDY ON DIFFERENT STAGES OF ADFF. AND S STANDS FOR THE STAGE.

Model	S1	S2	S3	S4	WHU-CD Pre. / Rec. / F1 / IoU / OA / Kappa / FA	LEVIR-CD Pre. / Rec. / F1 / IoU / OA / Kappa / FA
BiFA w/o ADFF	×	×	×	×	94.43 / 91.99 / 93.19 / 87.26 / 99.46 / 92.92 / 0.23	91.08 / 89.47 / 90.27 / 82.27 / 99.02 / 89.76 / 0.47
BiFA	✓	×	×	×	**95.27** / 91.84 / 93.52 / 87.83 / 99.49 / 93.26 / **0.18**	91.51 / 89.53 / 90.51 / 82.66 / 99.04 / 90.01 / 0.45
BiFA	✓	✓	×	×	95.19 / 92.08 / 93.62 / 87.98 / 99.50 / 93.35 / 0.19	**91.60** / 89.67 / 90.62 / 82.86 / 99.05 / 90.13 / 0.44
BiFA	✓	✓	✓	×	95.15 / **93.60** / **94.37** / **89.34** / **99.56** / **94.14** / 0.19	91.52 / 89.86 / **90.69** / **82.96** / **99.06** / **90.19** / **0.43**
BiFA	✓	✓	✓	✓	95.13 / 93.04 / 94.17 / 88.81 / 99.53 / 93.83 / 0.20	90.87 / **90.47** / 90.67 / 82.94 / 99.05 / 90.17 / 0.49

TABLE VII
ABLATION STUDY ON IND.

Model	Param. (M)	FLOPs. (G)	WHU-CD Pre. / Rec. / F1 / IoU / OA / Kappa / FA	LEVIR-CD Pre. / Rec. / F1 / IoU / OA / Kappa / FA
Base	7.43	11.77	95.18 / 92.69 / 93.92 / 88.54 / 99.52 / 93.68 / 0.19	91.09 / 89.35 / 90.22 / 82.17 / 99.01 / 89.69 / 0.47
Upsample	5.53	49.51	**95.34** / 92.78 / 93.94 / 88.57 / 99.43 / 93.71 / 0.19	91.42 / 89.56 / 90.48 / 82.62 / 99.04 / 89.98 / 0.46
Deconv	206.07	13192.11	95.06 / 92.57 / 93.80 / 88.32 / 99.51 / 93.54 / 0.20	**91.70** / 89.16 / 90.41 / 82.50 / 99.03 / 89.90 / 0.43
BiFA	5.58	53.00	95.15 / **93.60** / **94.37** / **89.34** / **99.56** / **94.14** / **0.19**	91.52 / **89.86** / **90.69** / **82.96** / **99.06** / **90.19** / **0.43**

settings of the loss function coefficients. For ease of model selection, both λ_1 and λ_2 are set to 0.5.

F. Advantages of IND

Different from the previous CD methods, our approach introduces a decoder based on implicit neural representation and realizes alignment between feature maps with different resolutions by learning continuous representation of images in coordinate space. To further explore the advantages of BiFA in the CD task, we conducted experiments on the LEVIR-CD as shown in Table XII. Train Res. and Test Res. indicate the resolution of the train set and test set. Taking the train set of LEVIR-CD as an example, 0.5m/pixel data is obtained by cropping the original image (the image size is 1024 × 1024 and the image resolution is 0.5m/pixel) to 256 × 256 patches. To obtain 1m/pixel data, the original image is first resized 512 × 512, resulting in the resolution of the image becoming 1m/pixel, and then cropped to 256 × 256 patches. In the case of 2m/pixel data, direct resizing of the original image to 256 × 256 is performed. Ultimately, three train sets with differing resolutions can be derived. Similarly, test sets with three resolutions are also available. BiFA w/o IND indicates the BiFA removes the implicit neural alignment decoder.

We choose CNN-based SNUNet, Transformer-based BIT, and DMINet for comparison. According to the experimental results (Table XII), our BiFA obtains the best results when using different-resolution images for training and testing. Specifically, when training at 0.5m/pixel resolution and testing at 1m/pixel resolution, the F1 of our method is 12% higher than SNUNet, 4.85% higher than BIT, and 3.72% higher than DMINet. Compared with a variant that does not use IND (BiFA w/o IND), the F1 of our BiFA also is 1.08% higher than it. These results indicate that BiFA has better robustness for cross-resolution training and testing than other methods.

TABLE VIII
ABLATION STUDY ON DIFFERENT PE.

Model	Learn PE	Fixed PE	**WHU-CD** Pre. / Rec. / F1 / IoU / OA / Kappa / FA	**LEVIR-CD** Pre. / Rec. / F1 / IoU / OA / Kappa / FA
BiFA	×	×	94.88 / 92.58 / 93.72 / 88.18 / 99.50 / 93.46 / 0.21	**91.92** / 89.27 / 90.57 / 82.78 / 99.05 / 90.08 / **0.42**
BiFA	×	✓	**95.55** / 92.26 / 93.88 / 88.47 / 99.52 / 93.62 / **0.18**	91.82 / 89.47 / 90.63 / 82.87 / 99.06 / 90.14 / 0.43
BiFA	✓	×	95.15 / **93.60 / 94.37 / 89.34 / 99.56 / 94.14** / 0.19	91.52 / **89.86 / 90.69 / 82.96 / 99.06 / 90.19** / 0.43

TABLE IX
EFFECT OF THE NUMBER OF HEADS IN BI

Num.	**WHU-CD** Pre. / Rec. / F1 / IoU / Kappa	**LEVIR-CD** Pre. / Rec. / F1 / IoU / Kappa
1	**95.84** / 92.58 / 94.18 / 89.00 / 93.94	91.71 / 89.36 / 90.52 / 82.69 / 90.02
2	95.64 / 92.59 / 94.09 / 88.85 / 93.85	91.95 / 89.39 / 90.65 / 82.90 / 90.15
4	94.75 / 93.06 / 93.89 / 88.50 / 93.64	92.69 / 88.71 / 90.66 / 82.92 / 90.17
8	95.15 / **93.60 / 94.37 / 89.34 / 94.14**	91.52 / 89.86 / **90.69 / 82.96 / 90.19**
16	94.64 / 93.25 / 93.94 / 88.57 / 93.69	91.32 / **89.95** / 90.63 / 82.87 / 90.14
32	95.48 / 92.91 / 94.17 / 88.99 / 93.93	**93.04** / 88.39 / 90.66 / 82.91 / 90.17

TABLE X
EFFECT OF THE KERNEL SIZE IN ADFF

Size	**WHU-CD** Pre. / Rec. / F1 / IoU / Kappa	**LEVIR-CD** Pre. / Rec. / F1 / IoU / Kappa
1	**96.25** / 92.20 / 94.18 / 89.01 / 93.95	**91.84** / 89.32 / 90.56 / 82.76 / 90.07
3	95.15 / **93.60 / 94.37 / 89.34 / 94.14**	91.52 / 89.86 / **90.69 / 82.96 / 90.19**
5	95.40 / 92.56 / 93.96 / 88.61 / 93.72	91.22 / **90.03** / 90.62 / 82.86 / 90.13
7	95.73 / 92.89 / 94.29 / 89.20 / 94.06	91.45 / 89.56 / 90.50 / 82.64 / 89.99
9	94.12 / 93.42 / 93.77 / 88.28 / 93.52	92.12 / 89.08 / 90.57 / 82.77 / 90.07

TABLE XI
EFFECT OF COEFFICIENTS OF LOSS FUNCTION

λ_1	λ_2	**WHU-CD** Pre. / Rec. / F1 / IoU / Kappa	**LEVIR-CD** Pre. / Rec. / F1 / IoU / Kappa
1	0	95.36 / **92.87** / 94.10 / 88.86 / 93.86	92.44 / 89.31 / **90.85 / 83.23 / 90.37**
0	1	92.65 / 87.57 / 90.04 / 81.88 / 89.64	90.55 / 89.30 / 89.93 / 81.69 / 89.39
0.5	0.5	95.15 / 93.60 / **94.37 / 89.34 / 94.14**	91.52 / 89.86 / 90.69 / 82.96 / 90.19
0.5	1	**95.74** / 92.04 / 93.86 / 88.43 / 93.61	90.84 / **90.54** / 90.68 / 82.96 / 90.19
1	0.5	96.30 / 92.49 / 94.36 / 89.32 / 94.13	**92.67** / 88.99 / 90.79 / 83.14 / 90.31

TABLE XII
COMPARISON RESULTS OF CROSS-RESOLUTION ON THE LEVIR-CD TEST SET. **RES.** IS THE SPATIAL RESOLUTION.

Model	**Train** Res. (m/pixel)	**Test** Res. (m/pixel)	**LEVIR-CD** F1	 IoU
SNUNet	0.5	0.5	88.59	79.51
SNUNet	0.5	1	64.45	47.55
SNUNet	2	1	66.40	49.70
BIT	0.5	0.5	90.03	81.87
BIT	0.5	1	71.61	55.77
BIT	2	1	63.65	46.68
DMINet	0.5	0.5	90.46	82.59
DMINet	0.5	1	72.74	57.16
DMINet	2	1	66.31	49.60
BiFA w/o IND	0.5	0.5	90.22	82.17
BiFA w/o IND	0.5	1	75.38	60.49
BiFA w/o IND	2	1	70.00	53.85
BiFA	0.5	0.5	90.69	82.96
BiFA	0.5	1	76.46	61.88
BiFA	2	1	72.15	56.43

V. CONCLUSION

In this paper, we propose a new CD method BiFA which is designed to align the features at different levels: temporal (channel, spatial), and multi-scale, to achieve accurate detection. For temporal alignment, the bi-temporal images are utilized as input for the BI, and their channel-level alignment is achieved through mutual guidance between them during the feature extraction stage. Subsequently, the bi-temporal features following channel alignment are inputted into the ADFF to acquire the spatial offset of the bi-temporal images, achieving their spatial alignment and producing more precise multi-stage differential features. Ultimately, for multi-scale alignment, the differential features of each stage are fed into the IND to precisely align differences across multiple stages in a lightweight manner, resulting in the acquisition of prediction maps of superior quality. Many ablation experiments have verified the effectiveness of each module. Meanwhile, experimental results on six public datasets show that our method is advantageous over other state-of-the-art methods. Furthermore, a cross-resolution CD is performed to delve into the benefits of BiFA, and the experimental results show that BiFA exhibits superior cross-resolution robustness. However, the feature map of each stage is queried by IND at the original image size, resulting in a significant increase in FLOPs. Future works will enhance the efficiency of IND and utilize it in the cross/continuous-resolution change detection tasks.

REFERENCES

[1] 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, p. 1662, 2020.

[2] F. Luo, T. Zhou, J. Liu, T. Guo, X. Gong, and J. Ren, “Multiscale diff-changed feature fusion network for hyperspectral image change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[3] Z. Lv, H. Huang, L. Gao, J. A. Benediktsson, M. Zhao, and C. Shi, “Simple multiscale unet for change detection with heterogeneous remote sensing images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[4] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, “Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[5] X. Li, F. Ling, G. M. Foody, and Y. Du, “A superresolution land-cover change detection method using remotely sensed images with different spatial resolutions,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 3822–3841, 2016.

[6] H. Zhou, F. Luo, H. Zhuang, Z. Weng, X. Gong, and Z. Lin, “Attention multihop graph and multiscale convolutional fusion network for hyperspectral image classification,” *IEEE Transac-*

tions on Geoscience and Remote Sensing, vol. 61, pp. 1–14, 2023.

[7] C. Liu, R. Zhao, J. Chen, Z. Qi, Z. Zou, and Z. Shi, "A decoupling paradigm with prompt learning for remote sensing image change captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[8] J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, "Building damage detection in satellite imagery using convolutional neural networks," *arXiv preprint arXiv:1910.06444*, 2019.

[9] Z. Lv, H. Huang, X. Li, M. Zhao, J. A. Benediktsson, W. Sun, and N. Falco, "Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective," *Proceedings of the IEEE*, vol. 110, no. 12, pp. 1976–1991, 2022.

[10] W. J. Todd, "Urban and regional land use change detected by using landsat data," *Journal of Research of the US Geological Survey*, vol. 5, no. 5, pp. 529–534, 1977.

[11] A. Singh, "Change detection in the tropical forest environment of northeastern india using landsat," *Remote sensing and tropical land management*, vol. 44, pp. 273–254, 1986.

[12] R. D. Jackson, "Spectral indices in n-space," *Remote sensing of environment*, vol. 13, no. 5, pp. 409–421, 1983.

[13] 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 Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–9, 2023.

[14] Z. Lv, P. Zhang, W. Sun, J. A. Benediktsson, J. Li, and W. Wang, "Novel adaptive region spectral–spatial features for land cover classification with high spatial resolution remotely sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[15] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE geoscience and remote sensing letters*, vol. 6, no. 4, pp. 772–776, 2009.

[16] T. Han, M. A. Wulder, J. C. White, N. C. Coops, M. Alvarez, and C. Butson, "An efficient protocol to process landsat images for change detection with tasselled cap transformation," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 1, pp. 147–151, 2007.

[17] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in vhr images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3677–3693, 2019.

[18] Y. Sun, L. Lei, D. Guan, and G. Kuang, "Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images," *IEEE Transactions on Image Processing*, vol. 30, pp. 6277–6291, 2021.

[19] R. G. Negri, A. C. Frery, W. Casaca, S. Azevedo, M. A. Dias, E. A. Silva, and E. H. Alcântara, "Spectral–spatial-aware unsupervised change detection with stochastic distances and support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 2863–2876, 2020.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

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

[23] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[24] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[25] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.

[26] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7296–7307, 2020.

[27] L. Song, M. Xia, L. Weng, H. Lin, M. Qian, and B. Chen, "Axial cross attention meets cnn: Bibranch fusion network for change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 32–43, 2022.

[28] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[29] M. Liu, Q. Shi, Z. Chai, and J. Li, "Pa-former: learning prior-aware transformer for remote sensing building change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[31] X. Yang, S. Li, Z. Chen, J. Chanussot, X. Jia, B. Zhang, B. Li, and P. Chen, "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 177, pp. 238–262, 2021.

[32] 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 Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.

[33] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[34] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[35] P. P. De Bem, O. A. de Carvalho Junior, R. Fontes Guimarães, and R. A. Trancoso Gomes, "Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks," *Remote Sensing*, vol. 12, no. 6, p. 901, 2020.

[36] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 42, no. 2, 2018.

[37] Y. Jiang, L. Hu, Y. Zhang, and X. Yang, "Wricnet: A weighted rich-scale inception coder network for multi-resolution remote sensing image change detection," *arXiv preprint arXiv:2108.07955*, 2021.

[38] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Autonomous Robots*, vol. 42, pp. 1301–1322, 2018.

[39] 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 Geoscience and Remote Sensing*

Letters, vol. 14, no. 10, pp. 1845–1849, 2017.

[40] 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 transactions on geoscience and remote sensing*, vol. 60, pp. 1–16, 2021.

[41] H. Chen, H. Zhang, K. Chen, C. Zhou, S. Chen, Z. Zou, and Z. Shi, "Continuous cross-resolution remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.

[42] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multiscale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[43] E. Guo, X. Fu, J. Zhu, M. Deng, Y. Liu, Q. Zhu, and H. Li, "Learning to measure change: Fully convolutional siamese metric networks for scene change detection," *arXiv preprint arXiv:1810.09111*, 2018.

[44] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[45] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2349–2358.

[46] R. Gadde, V. Jampani, and P. V. Gehler, "Semantic video cnns through representation warping," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4453–4462.

[47] D. Nilsson and C. Sminchisescu, "Semantic video segmentation by gated recurrent flow propagation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6819–6828.

[48] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[49] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, S. Tan, and Y. Tong, "Semantic flow for fast and accurate scene parsing," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 775–793.

[50] X. Li, X. Li, L. Zhang, G. Cheng, J. Shi, Z. Lin, S. Tan, and Y. Tong, "Improving semantic segmentation via decoupled body and edge supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 435–452.

[51] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *arXiv preprint arXiv:2209.08290*, 2022.

[52] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[53] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[54] M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, and A. Eriksson, "Implicit surface representations as layers in neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4743–4752.

[55] K. Genova, F. Cole, D. Vlasic, A. Sarna, W. T. Freeman, and T. Funkhouser, "Learning shape templates with structured implicit functions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7154–7164.

[56] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8628–8638.

[57] H. Hu, Y. Chen, J. Xu, S. Borse, H. Cai, F. Porikli, and X. Wang, "Learning implicit feature alignment function for semantic segmentation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*. Springer, 2022, pp. 487–505.

[58] B. Cheng, O. Parkhi, and A. Kirillov, "Pointly-supervised instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2617–2626.

[59] K. Chen, W. Li, J. Chen, Z. Zou, and Z. Shi, "Resolution-agnostic remote sensing scene classification with implicit neural representations," *IEEE Geoscience and Remote Sensing Letters*, 2022.

[60] K. Chen, W. Li, S. Lei, J. Chen, X. Jiang, Z. Zou, and Z. Shi, "Continuous remote sensing image super-resolution based on context interaction in implicit function space," *arXiv preprint arXiv:2302.08046*, 2023.

[61] Z. Qi, Z. Zou, H. Chen, and Z. Shi, "Remote-sensing image segmentation based on implicit 3-d scene representation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[62] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[63] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[64] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[65] M. Liu, Z. Chai, H. Deng, and R. Liu, "A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4297–4306, 2022.

[66] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[67] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 207–210.

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

[69] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[70] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

Haotian Zhang received his B.S. degree from the School of Computer and Information Technology, Shanxi University in 2019 and his M.S. degree from the School of Computer Science, Inner Mongolia University in 2022. He is currently working toward the PhD degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include remote sensing image processing, machine learning, and pattern recognition.

Hao Chen received his B.S. and Ph.D. degrees from the Image Processing Center, School of Astronautics, Beihang University in 2017 and 2023, respectively. He is currently a Junior Researcher at Shanghai AI Laboratory. His research interests include geospatial machine learning, remote sensing, earth monitoring, and prediction.

Chenyao Zhou received the B.S. degree from the School of Astronautics, Beihang University, Beijing, China, in 2022, where he is currently pursuing the M.S. degree with the Image Processing Center. His research interests include pattern recognition and machine learning.

Keyan Chen (Student Member, IEEE) received the B.S. and M.S. degree from the School of Astronautics, Beihang University, Beijing, China, in 2019 and 2022. He is working toward a Ph.D. in Image Processing Center, School of Astronautics, Beihang University. His research interests include remote sensing image processing, deep learning, pattern recognition, and multimodal.

Chenyang Liu (Student Member, IEEE) received his B.S. degree from the Image Processing Center, School of Astronautics, Beihang University in 2021. He is currently working towards the Ph.D. degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include machine learning, computer vision, and multimodal learning.

Zhengxia Zou is currently a Professor at the School of Astronautics, Beihang University. He received his B.S. degree and his Ph.D. degree from Beihang University in 2013 and 2018. During 2018-2021, he was a postdoc research fellow at the University of Michigan, Ann Arbor. His research interests include remote sensing image processing and computer vision. He has published more than 40 peer-reviewed papers in top-tier journals and conferences, including Nature Communications, Proceedings of the IEEE, IEEE Transactions on Image Processing, IEEE Transactions on Geoscience and Remote Sensing, and IEEE / CVF Computer Vision and Pattern Recognition. His website is https://zhengxiazou.github.io/.

Zhenwei Shi (Senior Member, IEEE) is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing, China. He has authored or co-authored over 200 scientific articles in refereed journals and proceedings. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning. Prof. Shi serves as an Editor for IEEE Transactions on Geoscience and Remote Sensing, Pattern Recognition, ISPRS Journal of Photogrammetry and Remote Sensing, Infrared Physics and Technology, etc. His personal website is http://levir.buaa.edu.cn/.