BD-MSA: Body Decouple VHR Remote Sensing Image Change Detection Method Guided by Multiscale Feature Information Aggregation

Yonghui Tan, Xiaolong Li, Yishu Chen, and Jinquan Ai

***Abstract*—The purpose of remote sensing image change detection (RSCD) is to detect differences between bitemporal images taken at the same place. Deep learning has been extensively used to RSCD tasks, yielding significant results in terms of result recognition. However, due to the shooting angle of the satellite, the impacts of thin clouds, and certain lighting conditions, the problem of fuzzy edges in the change region in some remote sensing photographs cannot be properly handled using current RSCD algorithms. To solve this issue, we proposed *a body decouple multiscale by feature aggregation change detection*, a novel model that collects both global and local feature map information in the channel and space dimensions of the feature map during the training and prediction phases. This approach allows us to successfully extract the change region's boundary information while also divorcing the change region's main body from its boundary. Numerous studies have shown that the assessment metrics and evaluation effects of the model described in this article on the publicly available datasets DSIFN-CD, S2Looking, and WHU-CD are the best when compared to other models.**

***Index Terms*—Body decouple, change detection (CD), multiscale information aggregation, very-high-resolution (VHR) images.**

I. Introduction

CHANGE detection (CD) is a technique for determining whether a change has occurred in the same area by examining images of that location at different times [1], [2], [3]. Binary CD is a popular technique that analyzes information between two images to determine whether a pixel in one image has changed. It then categorizes the pixels in the image as either changed or unchanged. One of the fundamental and essential issues in remote sensing (RS) is the interpretation of very-high-resolution (VHR) RS images. VHR remote sensing image change detection (RSCD) is useful for a variety of RS applications, including urban land use analysis [4], [5], [6], building detection [7], [8], [9], deforestation monitoring [10], [11], urban planning [12], [13], urban sprawl analysis [14], [15], disaster assessment [16], [17], and so on. All of the aforementioned are necessary for local governments to effectively manage their local urban development, and precise and effective RSCD procedures enable cities to be assessed and planned for in order to minimize or prevent adverse effects.

The challenge in RSCD is capturing the connections between regions of interest between bitemporal images while disregarding interference from other regions. At the same time, several irritating elements such as seasonality in the bitemporal and image quality issues such as noise and contrast are not of importance and should be ignored when performing CD.

The two primary streams of CD in RS images are the traditional method and the deep learning method, which has gained popularity in the last decade. Several strategies for detecting changes in RS images have been proposed before using deep learning to RS images on a broad scale. Coppin and Bauer [18] employed a pixel-based CD method for RSCD, which detects changes in gray values or colors by comparing images from two points in time pixel by pixel. Deng et al. [19] discovered and quantified land use change using PCA and a hybrid classifier that includes both unsupervised and supervised classification. He et al. [20] combined texture change information with standard spectral-based change vector analysis, resulting in integrated spectral and texture change information. Wu et al. [21] used slow feature analysis to isolate the most time-undeformed section of a multitemporal image and migrate it to a new feature space, effectively concealing the image's unaltered pixels.

Although these techniques have produced good results, they have certain drawbacks because they rely on standard image processing.

1) Traditional techniques often necessitate the manual design of features, which may necessitate domain expertise and experience.
2) When dealing with complicated sceneries, varied lighting conditions, and multicategory changes, traditional approaches have rather weak generalization capacity.

Manuscript received 25 December 2023; revised 27 February 2024 and 18 March 2024; accepted 19 April 2024. Date of publication 24 April 2024; date of current version 1 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42261078, in part by Jiangxi Provincial Key R&D Program under Grant 20223BBE51030, in part by the Science and Technology Research Project of Jiangxi Bureau of Geology under Grant 2022JXDZKJKY08, in part by the Open Research Fund of Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources under Grant MEMI-2021-2022-31, and in part by the Graduate Innovative Special Fund Projects of Jiangxi Province under Grant YC2023-S556. *(Corresponding author: Xiaolong Li.)*

Yonghui Tan, Xiaolong Li, and Jinquan Ai are with the Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake, Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China, and also with the CNNC Engineering Research Center of 3D Geographic Information, East China University of Technology, Nanchang 330013, China (e-mail: cv_tyh@ecut.edu.cn; lixiaolong@ecut.edu.cn; jinquan@ecut.edu.cn).

Yishu Chen is with the Ningbo Alatu Digital Technology Company, Ltd., Ningbo 315000, China (e-mail: 2817161223@qq.com).

Digital Object Identifier 10.1109/JSTARS.2024.3392917

3) For supervised learning, traditional methods often necessitate enormous amounts of manually labeled data.

Deep learning is a technique that has evolved tremendously quickly in the past decade, and deep-learning-based computer vision has achieved exceptional performance in RSCD tasks because of CNNs' robust feature extraction capabilities. Deep learning computer-vision-based RSCD techniques can be divided into three categories based on the structure of these models: pure convolution based, attention mechanism based, and Transformer based. The aforementioned can be categorized as follows.

1) Fully convolutional (FC-EF, FC-Siam-Di, and FC-Siam-Conc) [22], improved UNet++ [23], IFNet [24], CDNet [25], DTCDSCN [26], and TINY-CD [27], these models simply extract features from RS images using CNNs, which make it difficult to capture long-term dependencies across images and may be insensitive to complicated scene changes.
2) MSPSNet [28], DSAMNet [29], HANet [30], STANet [31], SNUNet [32], ADS-Net [33], DARNet [34], SR-CDNet [35], and TFI-GR [36], the strategies described previously boost the model's sensitivity to crucial regions and help improve CD accuracy, but it is challenging to collect global information in bitemporal images.
3) BIT [37], ChangeFormer [38], RSP-BIT [39], Swin-SUNet [40], MTCNet [41], TransUNetCD [42], DMAT-Net [43], FTN [44], AMTNet [45], and Hybrid-transcd [46], when compared to traditional convolutional approaches, Transformer can handle long-range relationships better, but its ability to extract local contextual information is poor and computationally expensive.

Although the methods described previously produced outstanding results in the RSCD task, they have certain flaws. Because of their narrow local perceptual domain and susceptibility to spatial fluctuations, pure convolution-based approaches have limited degrees of feature extraction for RSCD. Second, when executing RSCD, the approach based on the attention mechanism can only take into consideration the local information in the feature map and cannot aggregate the global information. Third, the Transformer-based solution lacks the link between contexts in the details, and the arithmetic need is excessively large.

It is worth mentioning that the changing camera angles for different time phases, as well as the fact that most RS photographs are not shot at an angle perpendicular to the ground, result in shadows on various features in the enormous number of RS images. Furthermore, thin clouds appear in some RS photographs as a result of meteorological conditions. As illustrated in Fig. 1, the majority of the buildings in the image are inclined and cast long shadows on the ground, and thin clouds can be seen in some of the photographs. When executing RSCD, the margins of the change region are likely to get blurred due to the aforementioned issue. As a result, we prefer to address this issue during model training.

Targeting the two aforementioned primary issues—that is, the inability of current RSCD methods to effectively aggregate global and local feature information simultaneously and the blurring of change region edges as a result of feature shadowing in RS images—we proposed *body decouple multiscale by feature aggregation change detection (BD-MSA)*, a model that can simultaneously aggregate global and local information in multiscale feature maps and decouple the change region's center from its edges during training.

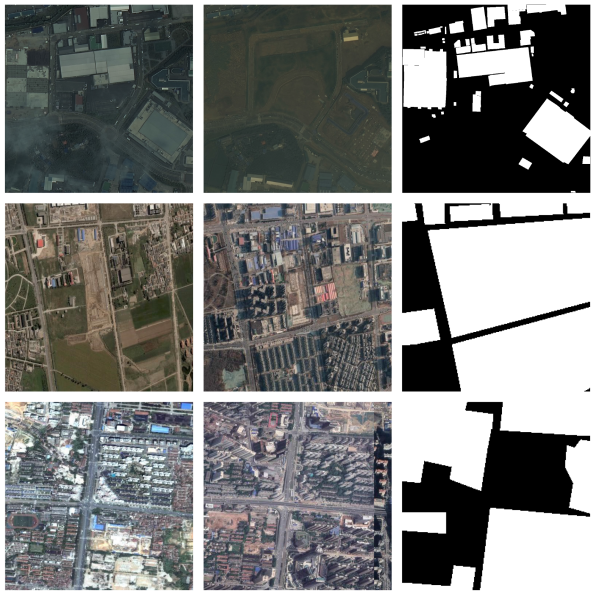

Fig. 1. Section of the images in the DSIFN-CD and S2Looking, with the first column representing the prechange images, the second representing the postchange images, and the third representing the change mask. The photographs in the figure's top row are from S2Looking, while those in the second and third rows are from DSIFN-CD.

The contributions of this article are as follows.

1) The overall feature aggregation module (OFAM), which we proposed in this article, is a technique that can simultaneously aggregated global and local information in both channel and spatial dimensions. It can adapt feature information at different scales in the backbone part while effectively increasing the model's accuracy.
2) Given the large difference in recognition accuracy between the main body and the edge of the changing region in the RSCD task, this article designs a decouple module in the prediction head part that can effectively separate the main body and the edge of the changing region, and the experimental results show that using this module improves the model's recognition accuracy for the edge.
3) Since the MixFFN module in SegFormer can capture intricate feature representations in the network, this article presents the module in the network decoder, enhancing the feature extraction and generalization capabilities of the model.
4) Extensive studies show that the technique presented in this work outperforms existing models on the public datasets DSIFN-CD and S2Looking, achieving the SOTA (state-of-the-art) performance.

The rest of this article is structured as follows. The prior approaches are introduced in Section II. The model's detail

described in this article is introduced in Section III. Section IV conducts experiments to compare this article's method with related methods. The discussion is shown in Section V. Finally, Section VI concludes this article.

II. Related Work

In this section, we present an overview of existing RSCD works, including: pure convolution based, attention mechanism based, and Transformer based.

A. Pure Convolutional-Based Model

Deep CNNs have achieved amazing performance in the field of computer vision [47] due to their powerful feature extraction capabilities. RS image interpretation is essentially an image processing in which deep learning plays an important role such as image classification [48], object detection [49], [50], semantic segmentation [51], [52], and CD [53].

In the field of RSCD, the first attempt to use fully convolutional networks was by Daudt et al. [22], who divided it into three methods, namely FC-EF, FC-Siam-Di, and FC-Siam-Conc, proposed a CD architecture for the Siamese Network, and demonstrated that this architecture is effective. In [23], an improved UNet++ [54] has been proposed, which adopts the MSOF strategy that can effectively combine multiscale information and help to detect objects with large size and scale variations on VHR RS images. Zhang et al. [24] proposed a depth-supervised image fusion network for CD in high-resolution bitemporal RS images, which combines the attention module and depth supervision to provide an effective new way for CD in RS images. For better industrial applications, Andrea et al. [27] proposed TINY-CD, which employs the Siamese U-Net architecture and an innovative mixed and attention masking block (MAMB) to achieve better performance than existing models while being smaller in size .

B. Attention Mechanism-Based Model

Attention mechanisms were first introduced in the context of natural language processing [55]. Later, computer vision researchers presented attentional processes that could be applied to images [56], [57], [58].

One can apply attention techniques in the field of RSCD, just like in most other computer vision jobs. In order to address issues like illumination noise and scale variations in aerial image CD, Shi et al. [29] introduced a deeply supervised attentional metric network for RSCD, this network incorporates a metric learning module and a convolutional block attentional module (CBAM) to enhance feature differentiation. In order to increase detection accuracy, Guo et al. [28] suggested a deep multiscale twin network for RSCD. This network is based on a deep multiscale twin neural network and incorporates a self-attention module and a parallel convolutional structure. Li et al. [34] proposed a dense attention refinement network that combines dense hopping connections, a hybrid attention module that combines a channel attention module and a spatial-temporal attention module, and a recursive refinement module to effectively improve the accuracy of CD in high-resolution bitemporal RS images. In order to overcome the resolution disparity between bitemporal images, Liu et al. [35] created SRCDNet, which learns superresolution images using adversarial learning and enriches multiscale features with a stacked attention module made up of five CBAMs.

Even though attention-based RSCD is more adept at identifying local contextual information from bitemporal RS images, it is less effective at capturing the global information.

C. Transformer-Based Model

Transformer is crucial to RSCD because of its potent global feature extraction capacity. For the first time, BIT [37] brings Transformer, which effectively describes context in the spatial-temporal domain to the RSCD domain. In order to model context and improve features, BIT converts the input image into a limited set of high-level semantic tokens using Transformer encoders and decoders; based on BIT, RSP-BIT [39] primarily focuses on using remote sensing pretraining (RSP) to analyze aerial images. It has been observed that RSP enhances performance on the scene identification test and helps comprehend the semantics related to RS; by fusing a multiscale Transformer with a CBAM, Wang et al. [41] developed MTCNet. It creates a multiscale module to create the multiscale Transformer after extracting the bitemporal image features using the Transformer module.

While Transformer performs RSCD tasks effectively in terms of global information extraction, its huge number of parameters makes prediction more time consuming, and it struggles to extract the semantics across local contexts.

III. Methodology

In this section, we proposed BD-MSA, a novel approach in which we first provide a brief overview of the general structure, followed by a full description of the modules in our approach in each subsection.

A. Overall Structure

The siamese network is presently a commonly utilized structure in RSCD, which uses two weight-sharing backbones in the feature extraction phase to extract features from the input. In BD-MSA, we feed $\mathrm{I} = \{\mathrm{I}_1, \mathrm{I}_2\}$ into the CNN backbone to extract the respective deep features of the bitemporal images, which are then sent successively through the decouple decoder and the prediction mask, and the output is compared with the Mask.

Fig. 2 depicts the general architecture diagram of BD-MSA. The diagram is divided into three primary sections: CNN backbone, decouple decoder, and prediction mask. The following equation can illustrate the model training process:

$$\hat{\mathrm{Y}} = \mathrm{Predict}\left(\mathrm{Decoder}\left(\mathrm{Backbone}\left\{\mathrm{I}_1, \mathrm{I}_2\right\}\right)\right) \tag{1}$$

where Backbone, Decoder, and Predict represent different parts of the model diagram, $\hat{\mathrm{Y}}$ represents the training result graph, and I_1 and I_2 represent the input bitemporal images. In Algorithm 1, we have expressed the model training procedure as pseudocode to help the reader comprehend.

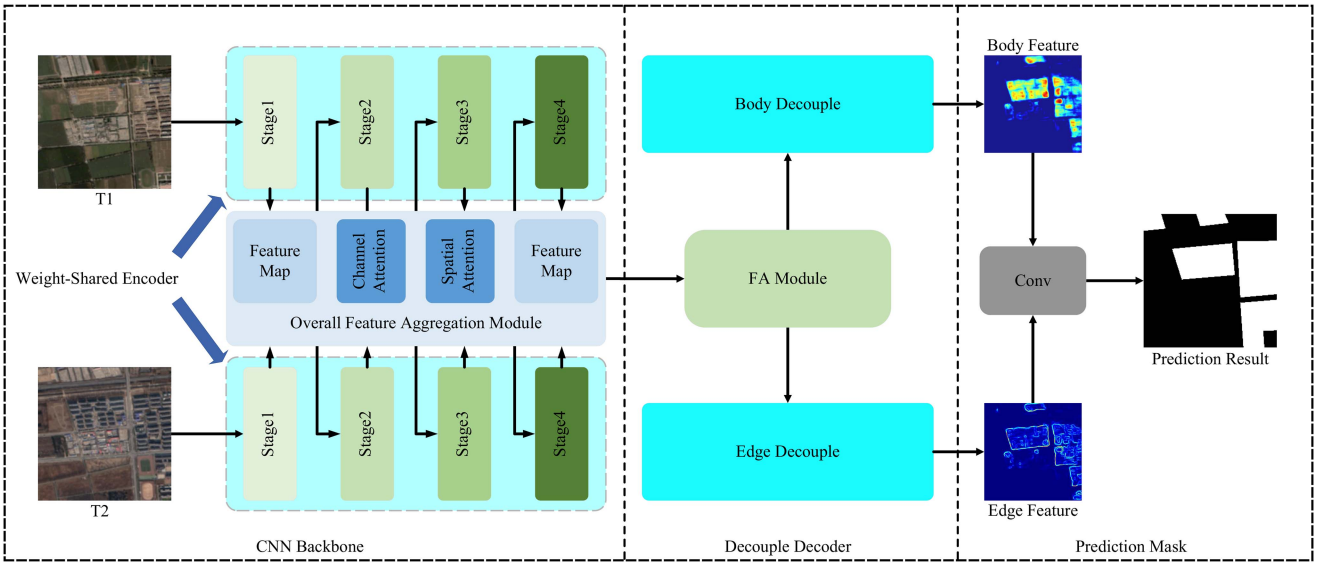

Fig. 2. Schematic diagram of BD-MSA.

Algorithm 1: Inference of BD-MSA for Change Detection.

Input: $\mathbf{I} = \{(\mathbf{I}^1, \mathbf{I}^2)\}$ (a pair of bi-temporal image)
Output: M (a prediction change mask)

// step1: extract high-level features by MiT backbone and OAFM
for i *in* $\{1, 2\}$ **do**
  for n *in* $\{1, 2, 3, 4\}$ **do**
    $\mathbf{MiT}_n^i = \text{MiT\_Backbone}(\mathbf{T}^i)$
    $\mathbf{F}_n^i = \text{OAFM}(\mathbf{MiT}_n^i)$
  end for
end for
// step2: Concat high-level feature to FA Module
$\mathbf{F}_{\text{FA}} = \text{FA\_Module}(\mathbf{F}_4^1, \mathbf{F}_4^2)$
// step3: Decoupling $\mathbf{F}_{\text{FA}}$ into $\mathbf{F}_{\text{body}}$ and $\mathbf{F}_{\text{edge}}$ by Body Decouple and Edge Decouple
$\mathbf{F}_{\text{body}} = \text{Body\_Decouple}(\mathbf{F}_{\text{FA}})$
$\mathbf{F}_{\text{edge}} = \text{Edge\_Decouple}(\mathbf{F}_{\text{FA}})$
$\mathbf{M} = \text{Conv}(\text{Concat}(\mathbf{F}_{\text{body}}, \mathbf{F}_{\text{edge}}))$

B. Overall Feature Aggregation Module (OFAM)

In the feature extraction section, we utilize a feature encoder with shared weights to send the input diachronic phase through two identical Backbones with the same weights during the training process. The backbone we designed, in particular, can be divided into four stages, in which the input features are first made to pass through the MiT [59], because of its greater success in the field of semantic segmentation in recent years, and the output results are then pass through the OFAM while extracting both local and global features in the channel dimension and spatial dimension of the feature map. Fig. 3 depicts the OFAM module that we designed.

Our designed OFAM is divided into three parts. First, we divide the output of MiT in channel dimension into two branches, one of which, local channel attention, is used to extract local features and the other branch, global channel attention, is used to extract global features; the related computational formula is as follows:

$$\mathbf{F}_c^l = \text{MiT}\left(\mathbf{F}_n^i\right) + \text{LCA}\left(\text{MiT}\left(\mathbf{F}_n^i\right)\right) \times \text{MiT}\left(\mathbf{F}_n^i\right) \quad (2)$$

$$\mathbf{F}_c^g = \text{GCA}\left(\text{MiT}\left(\mathbf{F}_n^i\right)\right) \times \text{MiT}\left(\mathbf{F}_n^i\right) \quad (3)$$

where LCA and GCA denote local channel attention and global channel attention, respectively, and $\mathbf{F}_c^l$ and $\mathbf{F}_c^g$ denote locally and globally extracted channel dimension features.

Following channel attention, the obtained $\mathbf{F}_c^l$ and $\mathbf{F}_c^g$ are sent to spatial attention, where they are used to construct global and local attention feature maps $\mathbf{F}_s^g$, $\mathbf{F}_s^l$ in channel dimension. The relevant formulas are as follows:

$$\mathbf{F}_s^l = \text{LSA}\left(\text{MiT}\left(\mathbf{F}_n^i\right)\right) \times \mathbf{F}_c^l + \mathbf{F}_c^l \quad (4)$$

$$\mathbf{F}_s^g = \text{GSA}\left(\text{MiT}\left(\mathbf{F}_n^i\right)\right) \times \mathbf{F}_c^g + \mathbf{F}_c^g \quad (5)$$

where LSA and GSA correspond to local spatial attention and global spatial attention in Fig. 3. $\mathbf{F}_s^l$ and $\mathbf{F}_s^g$ are the two output feature layers of spatial attention, which weight local and global information in the spatial dimension, respectively. Unlike channel attention, the topic part of spatial attention is symmetric, with only local spatial attention and global spatial attention differing.

Following the extraction of global and local information in the channel and spatial dimensions, the features are fused to produce the final output feature map

$$\mathbf{F}_{n+1}^i = \text{MiT}\left(\mathbf{F}_n^i\right) \times \mathbf{F}_s^l \times \mathbf{F}_s^g + \mathbf{F}_s^l + \mathbf{F}_s^g. \quad (6)$$

Each attention module in OFAM is detailed in depth in Fig. 4. The processing of the feature maps in each section is shown as follows.

1) Fig. 4(a) depicts a simple convolutional neural network that incorporates the layers of convolution, pooling, and

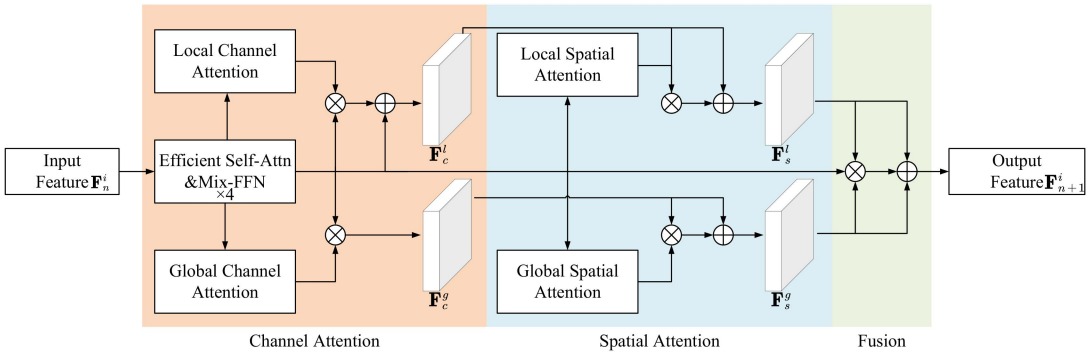

Fig. 3. Graphic depicts our OFAM, which is separated into three major portions: channel attention, spatial attention, and fusion, which are distinguished by various colored backgrounds.

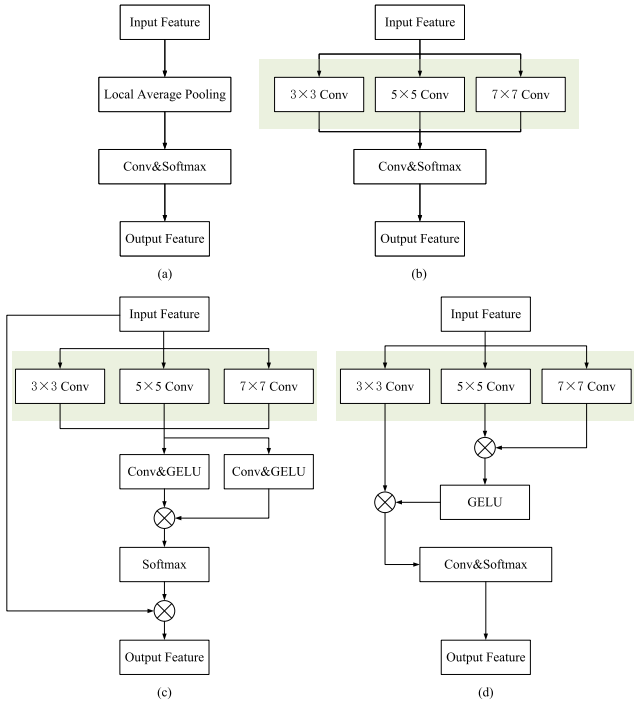

Fig. 4. (a)–(d) of the OFAM schematic diagrams depict local channel attention, global channel attention, local spatial attention, and global spatial attention, respectively, in Fig. 3.

so on by linking them in sequence, shown as follows:

$$\mathbf{F}_{\text{out}} = \sigma\left(\text{Conv}^{3\times3}\left(\text{LAP}\left(\mathbf{F}_{\text{in}}\right)\right)\right) \tag{7}$$

where σ denotes the Softmax activation function, LAP denotes local channel attention, $\text{Conv}^{3\times3}(\cdot)$ is a convolutional layer with a convolutional kernel size of 3×3, and $\mathbf{F}_{\text{in}}$ and $\mathbf{F}_{\text{out}}$ denote the input and output, respectively.

2) Fig. 4(b) tends to extract global features compared to Fig. 4(a) in the design of the pooling layer, and we picked three different sizes of convolution to extract the input features, which are 3×3, 5×5, and 7×7. Fig. 4(b) can be written as follows:

$$\mathbf{F}_{\text{out}} = \sigma\left(\text{Conv}^{3\times3}\left(\text{Concat}\left(\text{Conv}\left(\mathbf{F}_{\text{in}}\right)\right)\right)\right)$$

$$\text{Conv}\left(\cdot\right) = \left\{\text{Conv}^{3\times3}\left(\cdot\right), \text{Conv}^{5\times5}\left(\cdot\right), \text{Conv}^{7\times7}\left(\cdot\right)\right\} \tag{8}$$

where Concat denotes the splicing of the input feature $\mathbf{F}_{\text{in}}$ in the channel dimension after three different convolutions and scaling to a uniform size.

3) Fig. 4(a) and (b) weights the feature maps solely in the channel dimension, but they do not examine the relationship between the convolution kernel and the input feature maps for different convolution sizes; therefore, we devised Fig. 4(c) to address this issue. This can be stated mathematically as follows:

$$\mathbf{F}_{\text{mid}} = \sigma\left(\prod_{i=1}^{2}\text{ConvG}_i\left(\text{Conv}\left(\mathbf{F}_{\text{in}}\right)\right)\right)$$
$$\mathbf{F}_{\text{out}} = \mathbf{F}_{\text{in}} \times \mathbf{F}_{\text{mid}} \tag{9}$$

where ConvG_i denotes that the features are first subjected to a convolution operation with a convolution kernel size of 3×3, followed by the GeLU activation function [60].

4) We created the module depicted in Fig. 4(d) to use the ability of the interaction between different convolutional kernels for the extraction of global information, with the goal of weight extraction of global information at the spatial level. The following are the calculating formulas:

$$\mathbf{F}_{\text{mid}} = \gamma\left(\text{Conv}^{5\times5}\left(\mathbf{F}_{\text{in}}\right) \times \text{Conv}^{7\times7}\left(\mathbf{F}_{\text{in}}\right)\right)$$
$$\mathbf{F}_{\text{out}} = \text{ConvS}\left(\mathbf{F}_{\text{mid}} \times \text{Conv}^{3\times3}\left(\mathbf{F}_{\text{in}}\right)\right) \tag{10}$$

where γ denotes the GeLU activation function, and $\text{Conv}^{5\times5}$ and $\text{Conv}^{7\times7}$ denote convolutional layers with convolutional kernel sizes of 5×5 and 7×7, respectively. ConvS indicates that the previous feature map is first convolved by a convolution kernel size of 3×3, followed by a softmax activation function.

In the model feature extraction section, we combine the MiT feature extractor with OFAM. The global and local information in the feature map is retrieved simultaneously in both channel and spatial dimensions, thereby aggregating the positional and spectral information in the RS image.

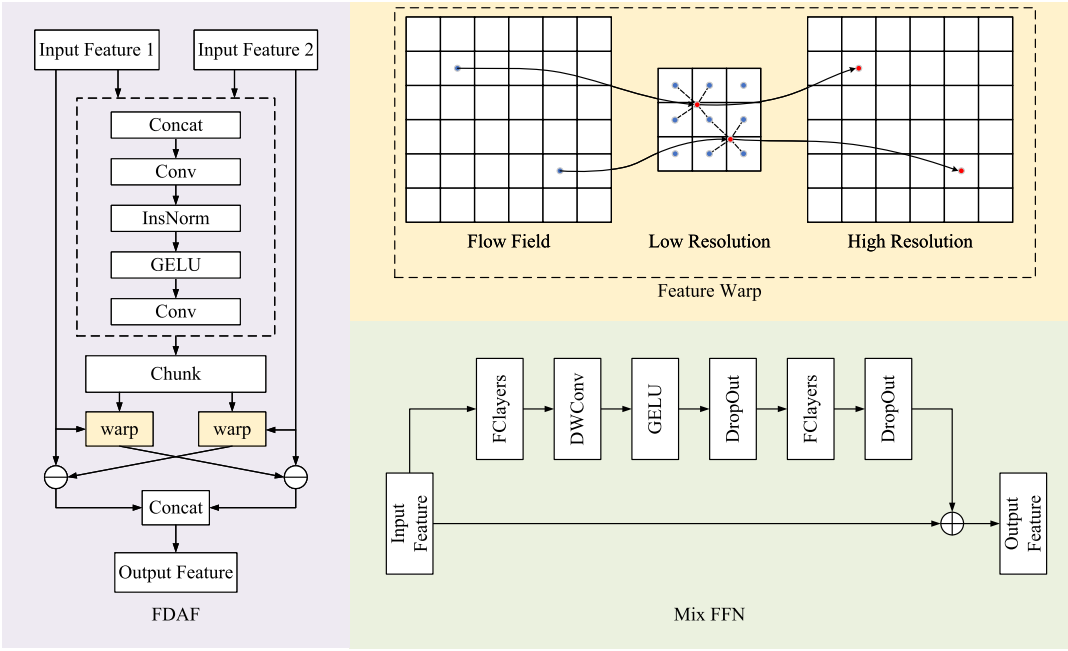

Fig. 5. Schematic representation of our FA module, which is separated into two main portions, FDAF and MixFFN, which are distinguished by various colored backgrounds.

C. Feature Alignment (FA) Module

After backbone, we created a feature aggregation module called FA module to better aggregate the deep features produced by feature extraction for bitemporal images. The FA module construction is depicted in Fig. 5. We integrate MixFFN from SegFormer [59] after FDAF in ChangerEX [61] to improve feature representation and contextual comprehension when performing feature extraction in image altering regions. The following are the relevant formulas:

$$\begin{aligned}\mathbf{F}_{\text{con}} &= \text{Concat}\left(\mathbf{F}_{\text{in1}}, \mathbf{F}_{\text{in2}}\right)\\ \mathbf{F}_{\text{flow}} &= \text{Conv}\left(\gamma\left(\text{InsNorm}\left(\text{Conv}\left(\mathbf{F}_{\text{con}}\right)\right)\right)\right)\\ \mathbf{F}_{\text{FDAF}} &= \text{Concat}\left(\mathbf{F}_{\text{in}} - \text{warp}\left(\mathbf{F}_{\text{flow1}}, \mathbf{F}_{\text{flow2}}\right)\right)\end{aligned} \tag{11}$$

where $\mathbf{F}_{\text{in1}}$ and $\mathbf{F}_{\text{in2}}$ denote the feature maps generated by backbone, respectively; InsNorm denotes the instance normalization method [62]; γ denotes the GeLU activation function; and warp is the feature warp in the upper right corner of the Fig. 5.

In FDAF, we first splice the two input features in channel dimension, and then, insert them into the dashed box on the left side of the figure. Borrowing the idea of flow field in the field of video processing [63], the authors design an FA method, i.e., warp in Fig. 5, to correct the feature offset problem caused by the dimensional change of the input feature maps after feature extraction is performed.

In warp, the semantic flow field Δ_{l-1} is generated by bilinear interpolating $\mathbf{F}_{l-1}$ to the same size as $\mathbf{F}_l$, then concatenating the two in the channel dimension, and finally, a convolutional layer. Following that, using a simple addition operation, each position p_{l-1} is mapped to a point p_l in the preceding layer l. Finally, using a bilinear sampling method, the values of the four nearby pixels are linearly interpolated to approximate the FAM's final output $\mathbf{F}_l(p_{l-1})$. The following are the relevant formulas for the aforementioned calculations:

$$\begin{aligned}\Delta_{l-1} &= \text{Conv}_l\left(\text{Concat}\left(\mathbf{F}_l, \mathbf{F}_{l-1}\right)\right)\\ p_l &= p_{l-1} + \frac{\Delta_{l-1}\left(p_{l-1}\right)}{2}\\ \mathbf{F}_l\left(p_{l-1}\right) &= \mathbf{F}_l\left(p_l\right) = \sum_{p \in N(p_l)} \omega_p \mathbf{F}_l(p)\end{aligned} \tag{12}$$

where $N(p_l)$ denotes the neighborhood of the deformation point p_l in $\mathbf{F}_l$, and ω_p denotes the bilinear kernel weights.

Considering the information interaction between bitemporal RS images and inspired by ChangerEx, we introduce FDAF into the method of this article and simultaneously insert MixFFN after FDAF to improve the feature expression ability after information fusion between bitemporal phases.

D. Feature Decouple Module

Some of the image change edges in the RSCD datasets were found to be blurred. This is due in part to the long shadows cast by tilt photography on ground buildings and in part to the blurring of image regions of interest caused by image quality issues in RS photographs such as overexposure, thin clouds, and so on.

Meanwhile, in the RSCD datasets, detection accuracy is high relative to the edges of the modified region due to consistent semantic information throughout the building, indicating homogeneity. In order to solve the aforementioned challenges, we expect to decouple the changing region interior and edges throughout the training process, which will allow us to extract the region boundary on the one hand and effectively minimize the computation on the other.

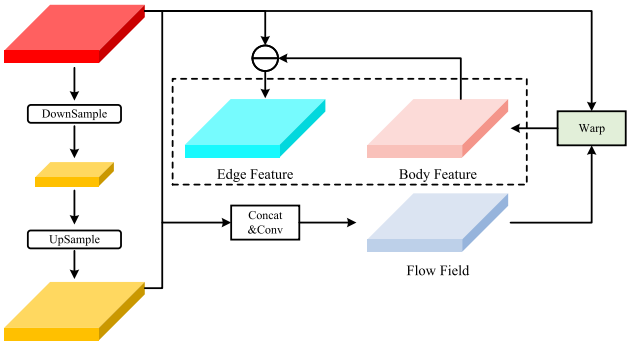

Fig. 6. Illustration of our proposed decouple module. Decoupling edge and body is feasible using the previously described flow, as well as the deep semantic characteristics that have been sampled twice.

As a result, we use the flow field concept and add the decouple module after feature decoding in the model to successfully extract the boundary of the changing region throughout the training process. The decouple module is depicted in Fig. 6.

We initially sample the input feature map $\mathbf{F}_{\text{in}}$ twice (DownSample and UpSample) in Fig. 6 to boost its semantic information without affecting the feature size. In Section III-C, we use Warp to correct the features of $\mathbf{F}_{\text{in}}$ to get $\mathbf{F}_{\text{body}}$, and then, subtract $\mathbf{F}_{\text{in}}$ from $\mathbf{F}_{\text{body}}$ to produce $\mathbf{F}_{\text{edge}}$.The following are the relevant formulas:

$$\begin{aligned}\mathbf{F}_{\text{flow}} &= \text{ConcatConv}\left(\mathbf{F}_{\text{in}}, \text{DownUp}\left(\mathbf{F}_{\text{in}}\right)\right)\\ \mathbf{F}_{\text{body}} &= \text{Warp}\left(\mathbf{F}_{\text{flow}}, \mathbf{F}_{\text{in}}\right)\\ \mathbf{F}_{\text{edge}} &= \mathbf{F}_{\text{in}} - \mathbf{F}_{\text{body}}\end{aligned} \tag{13}$$

where DownUp indicates that $\mathbf{F}_{\text{in}}$ is downsampled before being upsampled. ConcatConv denotes concatenation in the channel dimension first, followed by convolution through a convolution kernel of size 3×3.

After passing the features via the decouple module during the model training process, the features are successfully classified as edge features and body features. To the best of our knowledge, we are the first in the field of RSCD to do so. This substantially enhances the model's prediction capacity, and to some extent, reduces the number of parameters in the model.

IV. Experimental Results and Analysis

In this section, we first introduce the dataset, experimental environment, and validation metrics used in this article's experiments, then compare the model of this article to other models, conduct ablation experiments to evaluate the effect of each module, and finally, visualize some of the feature maps generated during the model's training process.

A. Experimental Setup

For this experiment, the three public RSCD datasets listed as follows were employed.

1) *DSIFN-CD* [24] is derived from six Chinese cities, including Beijing, and was manually collected in Google Earth. It is a publicly available binary CD dataset with a spatial resolution of 2 m that includes changes to roads, buildings, agriculture, and water bodies. During the experimental process, we cropped each image to 512×512, and the test set in the original dataset was of lower quality, so we divided the original training set into a training set and a validation set, and we used the original validation set as the test set, and the dataset now has 3000/600/340 training/validation/test, respectively.
2) *S2Looking* [64] is a publicly available dataset of 5000 pairs of bitemporal RS images broken into 3500/500/1000 training/validation/test sets with a spatial resolution of 0.5–0.8 m and a size of 1024×1024 for each image.
3) *WHU-CD* [65] is a publicly available CD dataset of RS image that covers the area of Christchurch, New Zealand that was struck by a magnitude 6.3 earthquake in February 2011 and rebuilt in subsequent years. The dataset consists of aerial imagery acquired in April 2012 and contains 12 796 buildings in 20.5 square kilometers (16 077 buildings in the same area in the 2016 dataset). The original size of the dataset was $32\,507 \times 15\,345$ with a resolution of 0.075 m, and was cropped to 256×256 for the experiments and the dataset now has 5947/743/744 training/validation/test, respectively. The conditions of this dataset such as illumination are desirable, so it is used for the validation of the generalizability of our model.

Some of the images in DSIFN-CD, S2Looking, and WHU-CD are shown in Fig. 7. The three columns in the figure are prechange image, postchange image, and change Mask, respectively.

B. Implementation Details

This experiment was deployed under PyTorch 2.0.1 and Python 3.8.13. For hardware, we used Intel Xeon E5-2678 v3 at 2.50 GHz $\times$ 2, 32 GB of RAM as well as used an NVIDIA RTX 4090 GPU. And for hyperparameters, we used BCE Loss as the article's loss function for our experiments and use AdamW as the optimizer, which is formally defined as

$$\begin{aligned}\mathcal{L}_{\text{BCE}} = -\frac{1}{H \times W}\sum_{h=1,w=1}^{H,W}\Big[&Y(h,w)\\ &+ (1 - Y(h,w)) \cdot \log\Big(1 - \hat{Y}(h,w)\Big)\Big]\end{aligned} \tag{14}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{\upsilon}_t} + \varepsilon}\hat{m}_t - \alpha\lambda\theta_t \tag{15}$$

where $H \times W$ is the size of the image to be predicted, $Y(h,w)$ is the predicted value of the point (h,w) in the image, $\hat{Y}(h,w)$ is the true value of the point, θ_t and θ_{t+1} denote the parameter values at time steps t and $t+1$, respectively, α is the learning rate, $\hat{m}_t$ and $\hat{\upsilon}_t$ are the exponential moving averages of the first-order and second-order moments, respectively, and ε is a very small value.

In this article, we use the Open-CD development kit [61] based on OpenMMLab [66] in order to compare the training results of different models in the same experimental environment.

Evaluation Metrics: We used the following metrics for validation to validate the training effect of our proposed BD-MSA:

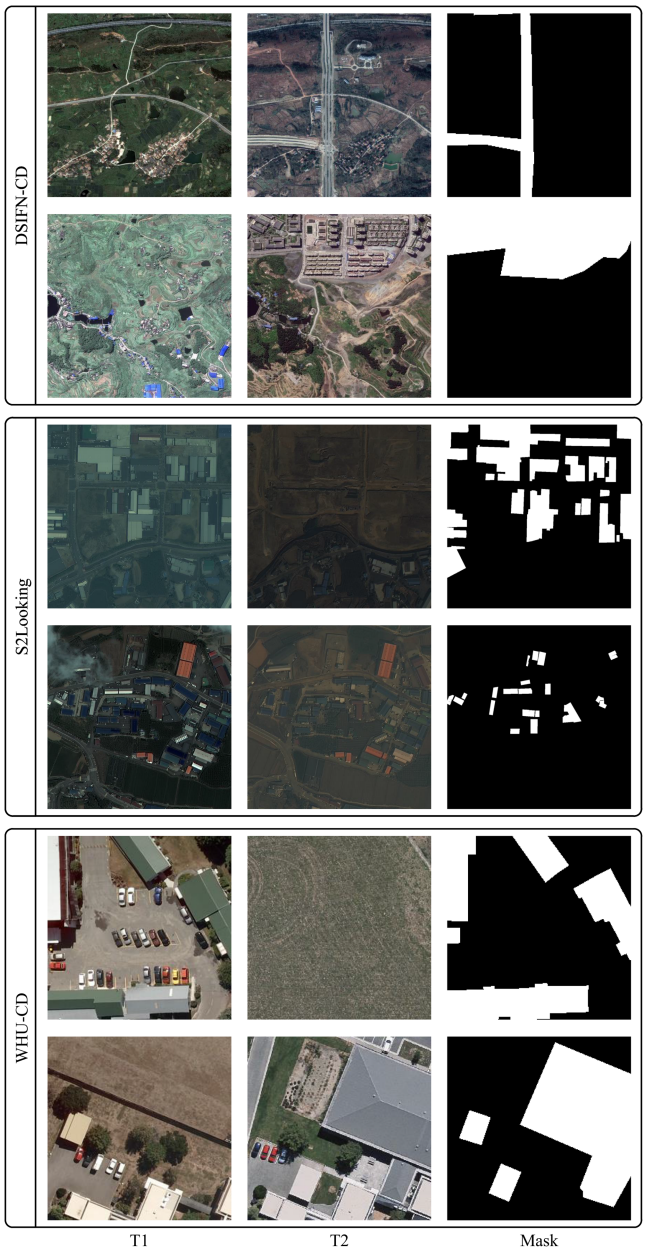

Fig. 7. Some of the images in DSIFN-CD, S2Looking, and WHU-CD.

F1-score (F1), Precision (Prec.), Recall (Rec.), and IoU, which are defined as follows:

$$\text{Pre.} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$\text{Rec.} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

$$\text{F1} = 2 \times \frac{\text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \tag{18}$$

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

where TP, FP, and FN represents the number of true positive, false positive, and false negative pixels, respectively.

C. Comparison With SOTA Methods

We compared the approaches mentioned in this work to some SOTA methods, which are listed as follows.

1) *FC-EF, FC-Siam-Di, and FC-Siam-Conc* [22] are built on fully convolutional networks [67] with a model structure similar to that of U-Net [68], and they use distinct methodologies to analyze paired image data.
2) *BIT* [37] introduces transformer [55] to classic CNN CD networks, which can more effectively capture long-distance interdependence and complex spatial dynamics.
3) *ChangeFormer* [38], unlike typical fully convolutional network-based techniques, ChangeFormer combines a hierarchically structured transformer encoder and a multilayer perceptron decoder to efficiently capture long range information at multiscale, enhancing CD accuracy.
4) *ChangerEx-MiT* [61] emphasizes the significance of feature interaction and presents simple but effective interaction mechanisms—AD and feature "exchange."
5) *HANet* [30] addresses the challenge of data imbalance between changed and unchanged pixels in the CD task by proposing a stepwise foreground-balanced sampling strategy to improve model learning for changed pixels and employing a concatenated network structure with hierarchical attention to integrate multiscale features for finer detection.
6) *IFNet* [24] collects deep features using a fully convolved two-stream architecture, and then, uses a difference discrimination network and an attention module to identify changes, highlighting the significance of deep supervision in improving border integrity and object internal compactness.
7) *SNUNet* [32], through tight hopping connections between the encoder and decoder as well as between decoders, SNUNet is able to maintain high-resolution fine-grained features while mitigating pixel uncertainty at the borders of changing targets and deterministic missingness of small targets.
8) *STANet* [31] captures spatial-temporal correlations via a self-attentive method in order to generate more discriminative features. It was divided into three variants, STANet-Base, STANet-Bam, and STANet-Pam.
9) *TINY-CD* [27] employs the Siamese U-Net architecture and a new feature mixing method to optimally utilize low-level information for spatial and temporal domains, while also offering a new spatial-semantic attention mechanism via its MAMB.

D. Main Results

On the DSIFN-CD and S2Looking datasets, we compared the outcomes of our proposed BD-MSA with previous SOTA approaches in Table I. The top, second best, and third best performers in each evaluation metric are shown in red, blue, and bolded black, respectively. The results reveal that our proposed BD-MSA outperforms the second-best model ChangerEx-MiT on the DSIFN-CD dataset, with an F1 score and IoU of 83.98% and 72.38%, respectively, which is 3.11% and 4.49% higher. Our

TABLE I
COMPARISON OF OUR PROPOSED BD-MSA WITH OTHER SOTA METHODS ON DSIFN-CD, S2LOOKING, AND WHU-CD DATASETS

Method	Backbone	#Param (M)	FLOPs(G)	DSIFN-CD				S2Looking				WHU-CD			
				F1	Prec.	Rec.	IoU	F1	Prec.	Rec.	IoU	F1	Prec.	Rec.	IoU
FC-EF [22]	-	1.353	12.976	63.44	75.97	54.47	46.46	7.65	81.36	8.95	8.77	69.73	80.32	61.6	53.52
FC-Siam-Di [22]	-	1.352	17.54	63.41	73.23	55.92	46.43	13.19	83.29	15.76	15.28	64.42	55.87	76.07	47.51
FC-Siam-Conc [22]	-	1.548	19.956	67.68	66.83	68.56	51.15	13.54	68.27	18.52	17.05	59.62	46.88	81.9	42.47
BIT [37]	ResNet18	2.99	34.996	71.04	77.22	65.78	55.09	44.51	67.41	33.23	28.63	91.08	91.55	90.61	83.62
ChangeFormer [38]	MiT-b1	3.847	11.38	**80.23**	84.4	76.46	**66.99**	**60.92**	**75.79**	50.93	43.8	92.25	95.39	89.31	85.62
ChangerEx-MiT [61]	MiT-b0	3.457	8.523	80.87	87.93	74.87	67.89	60.01	67.52	54.0	**42.87**	89.04	90.92	87.24	80.25
HANet [30]	-	3.028	97.548	75.91	75.9	**75.92**	61.18	43.67	44.89	42.51	27.93	79.18	86.65	72.9	65.54
IFNet [24]	VGG-16	35.995	323.584	79.21	**85.54**	73.75	65.58	61.98	64.96	59.27	44.91	91.09	90.74	91.45	83.64
SNUNet [32]	-	3.012	46.921	76.08	78.26	74.02	61.4	48.25	60.8	39.99	31.79	74.1	66.51	83.64	58.86
STANet-Base [31]	ResNet18	12.764	70.311	66.28	76.07	58.71	49.56	26.92	15.87	88.54	15.55	65.35	50.43	**92.8**	48.53
STANet-Bam [31]	ResNet18	12.846	391.168	61.48	70.83	54.31	44.39	27.27	16.11	88.68	15.79	67.88	52.47	96.12	51.38
STANet-Pam [31]	ResNet18	13.356	512	37.84	76.13	25.18	23.34	23.73	13.85	**82.79**	13.46	76.64	63.64	96.31	62.13
TINY-CD [27]	EfficientNet	0.285	5.791	74.71	76.37	73.12	59.63	54.5	63.81	47.56	37.46	**91.17**	**92.15**	90.21	**83.77**
BD-MSA (Ours)	MiT-b0	3.465	12.658	83.98	88.01	80.3	72.38	64.08	70.44	58.77	47.14	93.41	94.3	92.53	87.63

We use different colors to indicate: best, second best, and third best.

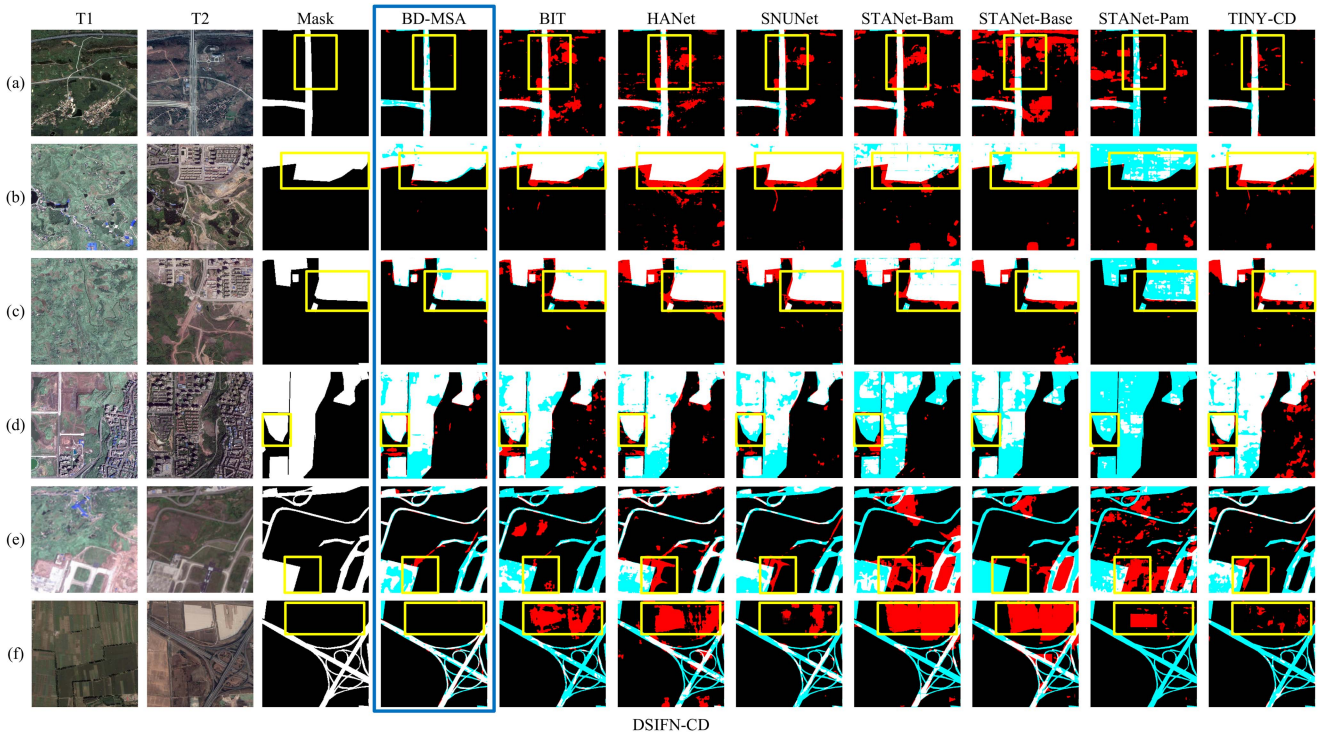

Fig. 8. Comparative experimental visualization results for each model on the DSIFN-CD test sets, which different colored regions denote FP, FN, and **TN**, respectively, and where the white region is TP.

suggested BD-MSA achieves an F1 score and IoU of 64.08% and 47.17% on the S2Looking dataset, which is 2.1% and 2.23% higher than the second-best model IFNet. The results demonstrate that our proposed BD-MSA performs well in the field of RSCD. In the column of #Param (M), we can see that our proposed BD-MSA has a modest number of parameters, which is 3.465 M; while this indication is not the smallest, it is a comparatively small number of parameters compared to many other techniques.

We also performed the same experiments on WHU-CD to confirm the proposed model's generalizability on other datasets. The findings indicate that, BD-MSA, like DSIFN-CD and S2Looking, achieves the highest metrics of both F1 and IoU on the WHU-CD test set, with an enhancement of 1.16% and 2.01%, respectively, over the second-best model. The aforementioned findings demonstrate BD-MSA's superior generalization capability.

We visualized the prediction results in the DSIFN-CD and S2Looking datasets to compared the method of this research with other methods in prediction results, as shown in Figs. 8 and 9. Varied hues in the graphic represent the model's varied prediction results for each pixel during the prediction phase. Simply said, the greater the proportion of white and black patches in the figure to the total image, the better the model's prediction outcome.

We specifically chose six photographs at random from each of DSIFN-CD and S2Looking as a test, and it is evident that the method in this work outperforms the other methods in terms of

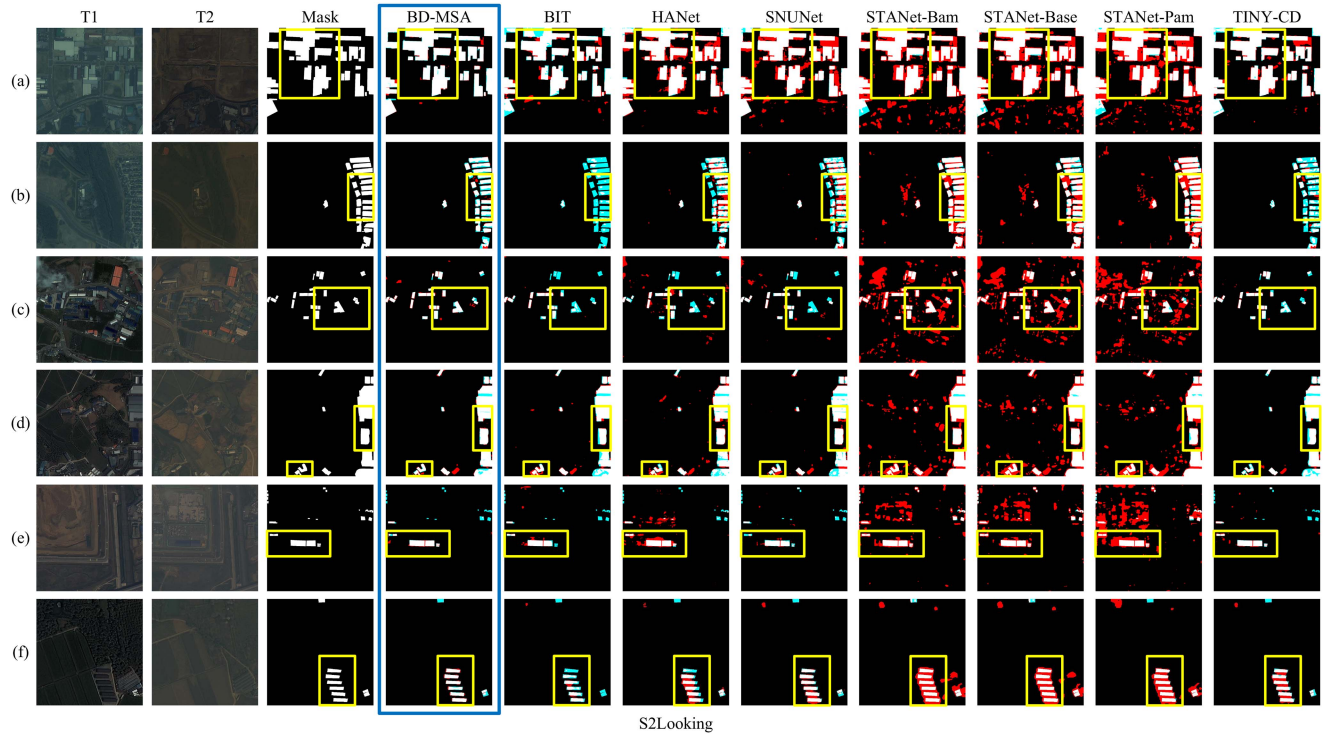

Fig. 9. Comparative experimental visualization results for each model on the S2Looking test sets, which different colored regions denote FP, FN, and **TN**, respectively, and where the white region is TP.

prediction outcomes. In Fig. 8(a), (e), and (f), our proposed approach effectively mitigates misclassification for nonchanging regions when making predictions; other models' predictions for the boundary of the changing regions are generally confusing in Fig. 8(b)–(d), however the model in this research solves the problem to a degree. Although certain models, such as STANet-Pam, have fewer mispredictions within the limits of the change region, they have a high missed detection rate, implying that the model cannot identify the boundaries well. The improvement of this article's model over other models, for S2Looking, is mostly in the precision of modifying the region's boundary and the effective decrease of the adhesion phenomenon between buildings. Refer to Fig. 9(d) and (e), BD-MSA predicts the edges of changing zones more accurately; in Fig. 9(a) and (f), BD-MSA successfully mitigates the adhesion phenomena between buildings with more compact layouts.

To confirm the generalizability of BD-MSA, we conducted tests akin to the ones described previously, randomly selecting six images from the WHU-CD test set to test each model. The experimental outcomes are displayed in Fig. 10. In comparison to the other models, BD-MSA demonstrates satisfying test results that are extremely near to the mask in many images, similar to the findings on the DSIFN-CD and S2Looking test sets.

To compare IoU as well as Params. between multiple models at the same time, we plotted the color mapping for the test results of different models on both datasets, as shown in Fig. 11. Each point in the graph represents a model, with the horizontal axis representing the model's parameters and the vertical axis representing the IoU of each model on the three datasets. The closer the model is to the upper left corner of the figure, the higher the accuracy detection, while the model takes less arithmetic power. Our proposed BD-MSA may be seen in the upper left corner, suggesting that the IoU reaches its maximum value and the number of parameters is lower than in most models.

Furthermore, the preceding conclusions show that the model in this study migrates better across devices than alternative models, particularly for machines with weaker arithmetic capability.

E. Ablation Studies

We conduct ablation tests on OFAM, MixFFN, and the decouple module, respectively, to validate the influence of different modules on our proposed model.

The nomenclature of the models in the ablation experiments is as follows.

1) *Baseline*: MiT + FDAF + Predict layer.
2) *BD-MSA-1-1*: Baseline + MixFFN.
3) *BD-MSA-1-2*: Baseline + Decouple.
4) *BD-MSA-1-3*: Baseline + OFAM.
5) *BD-MSA-2-1*: Baseline + MixFFN + Decouple Module.
6) *BD-MSA-2-2*: Baseline + MixFFN + OFAM.
7) *BD-MSA-2-3*: Baseline + Decouple Module + OFAM.
8) *BD-MSA*: Baseline + MixFFN + Decouple Module + OFAM.

The results of each ablation experiments are shown in Tables II–IV.

This show that adding each module improves the assessment metrics F1 and IoU when compared to the baseline, with F1 being able to synthesize Prec. and Rec. When only one module is added, adding OFAM results in the greatest improvement in

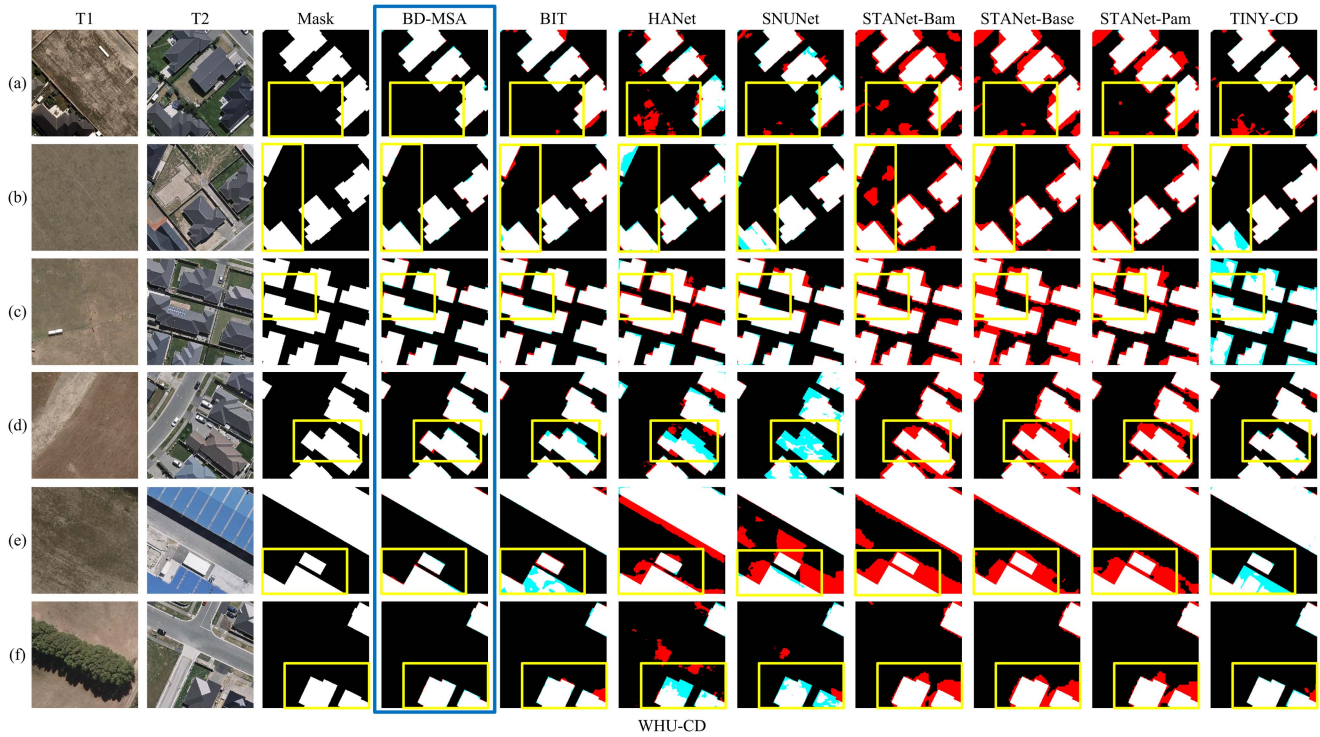

Fig. 10. Comparative experimental visualization results for each model on the WHU-CD test sets, which different colored regions denote FP, FN, and **TN**, respectively, and where the white region is TP.

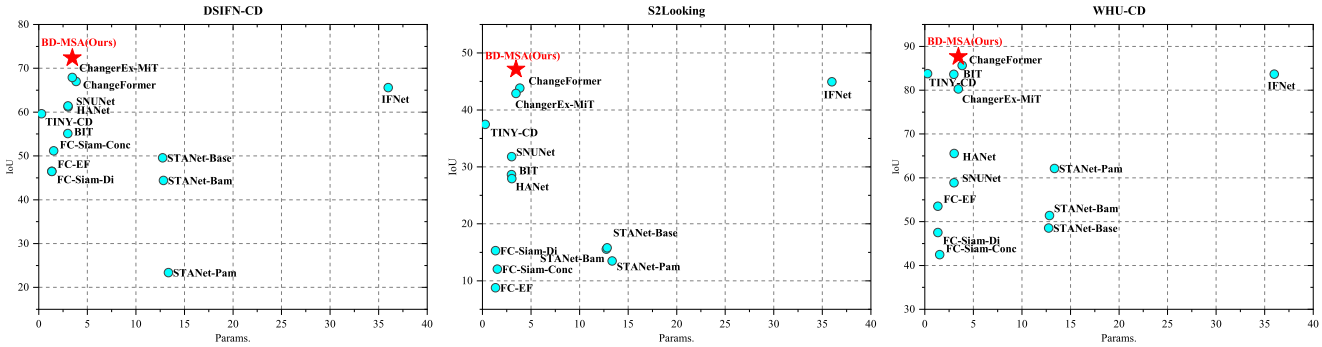

Fig. 11. Params. and IoU of different models on the three datasets, the top and bottom parts of the figure show the evaluation results of each model on DSIFN-CD, S2Looking, and WHU-CD, respectively.

TABLE II
RESULTS OF ABLATION EXPERIMENTS ON DSIFN-CD TEST

Module	+MixFFN	+Decouple	+OFAM	F1	Prec.	Rec.	IoU
Baseline				80.64	84.73	76.92	67.56
BD-MSA-1-1	✓			80.87	**87.93**	74.87	67.89
BD-MSA-1-2		✓		80.87	87.94	74.85	67.89
BD-MSA-1-3			✓	81.7	86.64	77.3	69.06
BD-MSA-2-1	✓	✓		81.04	86.19	76.46	68.12
BD-MSA-2-2	✓		✓	82.82	86.92	**79.1**	70.68
BD-MSA-2-3		✓	✓	**82.77**	86.29	79.52	**70.6**
BD-MSA	✓	✓	✓	83.98	88.01	80.3	72.38

We use different colors to indicate: best, second best, and third best.

TABLE III
RESULTS OF ABLATION EXPERIMENTS ON S2LOOKING TEST

Module	+MixFFN	+Decouple	+OFAM	F1	Prec.	Rec.	IoU
Baseline				56.72	70.21	47.58	39.59
BD-MSA-1-1	✓			60.01	67.52	54.0	42.87
BD-MSA-1-2		✓		59.73	70.33	51.9	42.58
BD-MSA-1-3			✓	61.94	70.38	55.31	45.93
BD-MSA-2-1	✓	✓		63.27	75.36	54.52	46.27
BD-MSA-2-2	✓		✓	63.31	73.49	**55.61**	46.32
BD-MSA-2-3		✓	✓	**63.29**	**71.23**	56.94	**46.28**
BD-MSA	✓	✓	✓	64.08	70.44	58.77	47.14

We use different colors to indicate: best, second best, and third best.

assessment metrics, which we assume is related to the fact that OFAM is added to all four phases of the backbone.

To visualize the outcomes of each module's ablation experiments, we exhibit its effect on the test set evaluation of DSIFN-CD, S2Looking, and WHU-CD in Fig. 12. Although the prediction effect of each ablation experimental model for the bitemporal images prediction in Fig. 12 is mostly right. BD-MSA outperforms the other models in predicting the edges

Fig. 12. Results of ablation experiments for each model on the DSIFN-CD, S2Looking, and WHU-CD test sets.

TABLE IV
RESULTS OF ABLATION EXPERIMENTS ON WHU-CD TEST

Module	+MixFFN	+Decouple	+OFAM	F1	Prec.	Rec.	IoU
Baseline				89.04	90.92	87.24	80.25
BD-MSA-1-1	✓			89.96	90.63	89.3	81.75
BD-MSA-1-2		✓		89.32	93.73	85.31	80.7
BD-MSA-1-3			✓	91.02	94.14	88.11	83.52
BD-MSA-2-1	✓	✓		92.25	95.39	89.31	85.62
BD-MSA-2-2	✓		✓	**92.48**	94.41	**90.63**	**86.01**
BD-MSA-2-3		✓	✓	92.59	**94.33**	90.9	86.2
BD-MSA	✓	✓	✓	93.41	94.3	92.53	87.63

We use different colors to indicate: best, second best, and third best.

TABLE V
DIFFERENT STAGES IN BACKBONE ARE FOLLOWED BY THE RESULTS OF THE OFAM ABLATION EXPERIMENTS ON THE DSIFN-CD TEST SETS

Methods	stage1	stage2	stage3	stage4	F1	Prec.	Rec.	IoU
OFAM-1	✓				82.6	87.68	78.08	70.36
OFAM-2	✓	✓			**79.65**	**84.81**	**75.09**	**66.18**
OFAM-3	✓	✓	✓		76.32	83.73	70.11	61.71
OFAM-4	✓	✓	✓	✓	83.98	88.01	80.3	72.38

We use different colors to indicate: best, second best, and third best.

TABLE VI
DIFFERENT STAGES IN BACKBONE ARE FOLLOWED BY THE RESULTS OF THE OFAM ABLATION EXPERIMENTS ON THE S2LOOKING TEST SETS

Methods	stage1	stage2	stage3	stage4	F1	Prec.	Rec.	IoU
OFAM-1	✓				63.66	74.29	55.7	46.7
OFAM-2	✓	✓			57.49	66.91	50.4	40.34
OFAM-3	✓	✓	✓		**60.01**	**67.59**	**53.96**	**42.86**
OFAM-4	✓	✓	✓	✓	64.08	70.44	58.77	47.14

We use different colors to indicate: best, second best, and third best.

of the change region. The parts of the photography where BD-MSA outperforms the predictions of other models have been highlighted in yellow boxes.

In addition to ablation experiments on different modules, we also perform ablation studies on OFAM modules, specifically adding OFAM modules behind different stages in the backbone, as shown in Tables V–VII. The results show that adding OFAM modules to all stages of the backbone has the greatest effect on the evaluation metrics, whereas OFAM-1 in the table is second best in each evaluation index, which we hypothesize is due to the fact that the first stage of the backbone has the largest feature map, and the addition of OFAM modules can effectively

TABLE VII
DIFFERENT STAGES IN BACKBONE ARE FOLLOWED BY THE RESULTS OF THE OFAM ABLATION EXPERIMENTS ON THE WHU-CD TEST SETS

Methods	stage1	stage2	stage3	stage4	F1	Prec.	Rec.	IoU
OFAM-1	✓				92.6	94.39	90.88	86.23
OFAM-2	✓	✓			89.8	90.94	88.69	81.48
OFAM-3	✓	✓	✓		**92.25**	95.39	**89.31**	**85.62**
OFAM-4	✓	✓	✓	✓	93.41	**94.3**	92.53	87.63

We use different colors to indicate: best, second best, and third best.

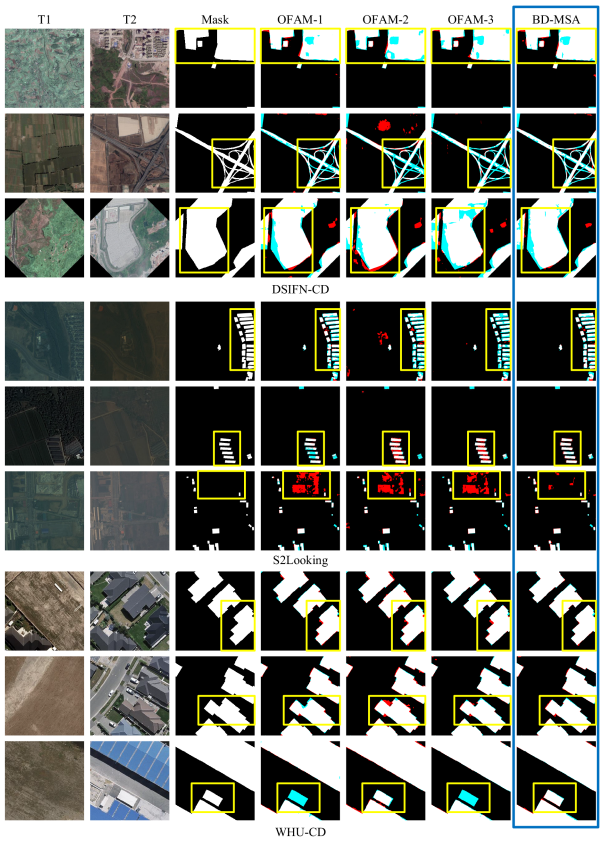

Fig. 13. Visualization of the results of ablation experiments on DSIFN-CD, S2Looking, and WHU-CD test sets for different stages followed by OFAM in backbone.

aggregate information in the feature map, thus reducing the computational cost of the model.

Fig. 13 depicts the experimental outcomes of introducing OFAM behind various phases of the backbone. In general, each model achieves better prediction results, but BD-MSA outperforms the other models in the subtle aspects shown in the figure with yellow boxes, such as edge detection, which is more accurate and can separate buildings with tight layouts very well.

F. Feature Map Visualization

To investigate whether the modules in this article's model are able to aggregate semantic information in the prediction process for bitemporal images, we used Grad-CAM [69] to view some of the feature layers in BD-MSA, and the results are shown in Fig. 14.

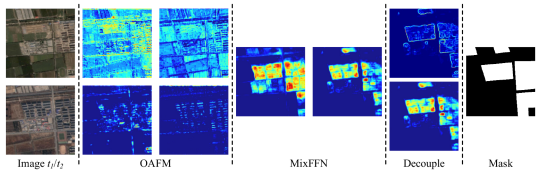

Fig. 14. Visualization of heat maps generated by some modules.

TABLE VIII
RATIO OF POSITIVE AND NEGATIVE PIXEL SAMPLES IN DIFFERENT DATASETS

Datasets	positive sample	negative sample	ratio
DSIFN-CD	361819124	671028236	35.03%
S2Looking	66552990	5176327010	1.27%
WHU-CD	21442501	481873976	4.26%

From left to right, the figure is divided into five sections: the original bitemporal images, feature maps before and after OFAM for stage 1, feature maps before and after MixFFN, boundary and body feature maps generated by decouple module, and change labels.

The figure clearly shows that the OFAM Module can transfer the weight in the feature map from the unimportant road part to the more important building part; MixFFN can focus the features on the changing region while reducing the weight of the nonchanging region; and decouple module can effectively decouple the feature map and extract the edge features.

V. DISCUSSION

In this section, we discuss the following three issues: the effect of different datasets on the experimental results, how different hyperparameters affect the model performance and how semisupervised learning methods affect the overall performance.

A. Effect of Training Set on Experimental Results

An essential factor influencing the experimental outcomes during model training is the quality of the dataset. When we conducted the experiments, we discovered that the final test accuracy varied significantly between datasets. For instance, the IoU on the test set in WHU-CD reached 87.63%, while the IoU on the test set in S2Looking was only 47.14%. Based on our conjectures, we determined that one of the causes of this phenomenon was the datasets' excessively variable proportion of positive and negative samples. To address this, we counted the number of pixels in each dataset as well as the percentage of positive samples overall, as indicated in Table VIII.

The findings demonstrate that while the proportion of positive sample pixels in S2Looking is very low at 1.27% of total pixels, it is also low in WHU-CD, at 4.26%, significantly lower than in DSIFN-CD, which has a proportion of 35.03%. Through an examination of the images in the dataset, we discovered that while the percentage of positive samples in WHU-CD is significantly lower than in DSIFN-CD, the reason for this phenomenon is that the majority of WHU-CD's areas remain unchanged, and in the images that have changed, the altered areas are all buildings, all of which have very regular shapes.

TABLE IX
EXPERIMENTAL RESULTS OF SEMISUPERVISED LEARNING OF BD-MSA WITH DIFFERENT LABELED RATIO ON DIFFERENT DATASETS

Labeled Ratio	DSIFN-CD				S2Looking				WHU-CD			
	F1	Prec.	Rec.	IoU	F1	Prec.	Rec.	IoU	F1	Prec.	Rec.	IoU
5%	50.33	68.03	39.94	33.63	55.07	58.22	52.25	38.0	74.74	**83.3**	67.78	59.67
10%	55.86	69.11	46.87	38.75	56.31	61.37	52.02	39.19	80.11	82.44	77.91	66.82
20%	**57.95**	75.75	**46.93**	**40.8**	**58.73**	**64.6**	**53.84**	**41.58**	**83.94**	79.22	**89.27**	**72.33**
40%	69.59	**69.38**	69.8	53.36	60.73	66.25	56.06	43.61	87.03	84.6	89.6	77.04
100%	83.98	88.01	80.3	72.38	64.08	70.44	58.77	47.14	93.41	94.3	92.53	87.63

We use different colors to indicate: best, second best, and third best.

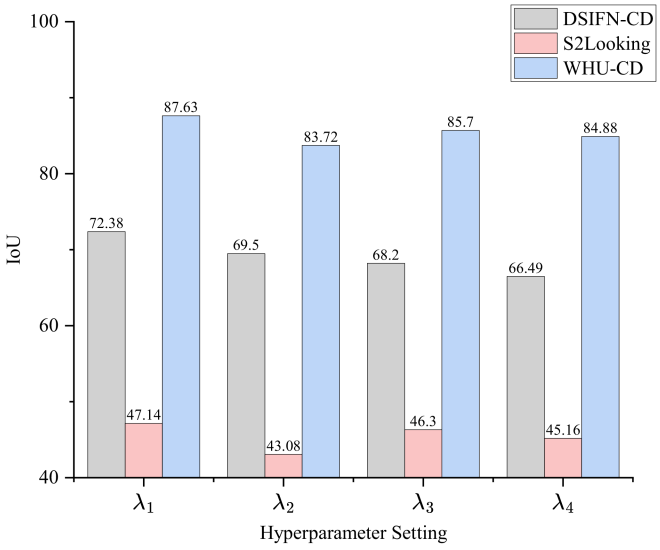

Fig. 15. Experimental results for different hyperparameter settings on different datasets.

The models obtain a reasonably decent level of accuracy in this dataset since the lighting, contrast, and shooting angle are all extremely perfect.

B. Effect of Different Hyperparameters on Model Performance

In every experiment in this work, we use AdamW as the optimizer and BCE Loss as the loss function. Since we are using the Open-CD development kit, we utilize PolyLR, the default learning rate strategy, for learning rates. We test the model on all three of the publicly accessible datasets for which the aforementioned hyperparameters are modified in order to investigate the effects of various hyperparameters on the model's performance.

The various methods for setting hyperparameters are as follows.

1) λ_1: BCE Loss + AdamW + PolyLR.
2) λ_2: Dice Loss + AdamW + PolyLR.
3) λ_3: BCE Loss + SGD + PolyLR.
4) λ_4: BCE Loss + AdamW + StepLR.

Fig. 15 displays the outcomes of the experiment. The IoU of evaluation metrics on each dataset varies somewhat depending on the hyperparameter settings used. By comparing the data in the image, it can be seen that the hyperparameter setting of λ_1 (BCE Loss + AdamW + PolyLR) yields the greatest IoU across all datasets.

C. Impact of Semisupervised Learning on Model Performance

This study proposes the BD-MSA model, which is mainly based on supervised learning and necessitates a huge quantity of labeled data. It remains challenging to gather a significant number of high-quality bitemporal RS images of the same region in the real world, despite the fact that the experimental setting used in this study can accommodate a sizable number of datasets for training.

To address the aforementioned problems, this article simulates the semisupervised learning approach and investigates how it influences the model's overall performance by randomly sampling the training sets from each public dataset in proportions of 5%, 10%, 20%, and 40%, respectively. The training sets are then configured with the same hyperparameter settings. Table IX presents the experimental outcomes.

It makes logical sense that when the sample ratio rises, as demonstrated by the experimental findings, the model's various assessment metrics across datasets increase. Remarkably, BD-MSA attains relatively good assessment metrics on S2Looking and WHU-CD upon reaching a sample ratio of 40%.

Since this article's methodology is fundamentally a supervised learning strategy, CD in semisupervised learning will necessarily be less accurate. However, utilizing less than half of the training set data volume, a relatively good accuracy was obtained, a phenomenon that captures our interest and can be focused on semisupervised learning in future work.

VI. CONCLUSION

In this study, we suggested a novel approach for RSCD called BD-MSA. In the training and prediction phase, the approach can combine global and local information in both channel and spatial dimensions, as well as decouple the main body of the change region and the edges of the feature maps. The experimental results suggest that the technique in this research outperforms previous models on the public datasets DSIFN-CD, S2Looking, and WHU-CD in terms of SOTA performance. We further demonstrate, through a series of ablation experiments, that all modules in this study are superior to the baseline.

We will continue to investigate the following aspects in the future:

1) the method in this article has only been validated on three public datasets, DSIFN-CD, S2Looking, and WHU-CD,

and it will be validated on more public datasets in the future;

2) the method in this article is essentially a supervised learning method, and we hope to explore unsupervised learning methods for application to tasks such as RSCD and more domain migration in future work.

REFERENCES

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

[2] T. Bai et al., "Deep learning for change detection in remote sensing: A review," *Geo-Spatial Inf. Sci.*, vol. 26, no. 3, pp. 262–288, 2023.

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

[4] I. Onur, D. Maktav, M. Sari, and N. Kemal Sönmez, "Change detection of land cover and land use using remote sensing and GIS: A case study in Kemer, Turkey," *Int. J. Remote Sens.*, vol. 30, no. 7, pp. 1749–1757, 2009.

[5] A. Tariq and F. Mumtaz, "Modeling spatio-temporal assessment of land use land cover of Lahore and its impact on land surface temperature using multi-spectral remote sensing data," *Environ. Sci. Pollut. Res.*, vol. 30, no. 9, pp. 23908–23924, 2023.

[6] R. Ray et al., "Quantitative analysis of land use and land cover dynamics using geoinformatics techniques: A case study on Kolkata metropolitan development authority (KMDA) in West Bengal, India," *Remote Sens.*, vol. 15, no. 4, 2023, Art. no. 959.

[7] J. Li, X. Huang, L. Tu, T. Zhang, and L. Wang, "A review of building detection from very high resolution optical remote sensing images," *GISci. Remote Sens.*, vol. 59, no. 1, pp. 1199–1225, 2022.

[8] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1835–1838.

[9] E. Maltezos, N. Doulamis, A. Doulamis, and C. Ioannidis, "Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds," *J. Appl. Remote Sens.*, vol. 11, no. 4, pp. 042620–042620, 2017.

[10] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 7, pp. 3299–3307, Jul. 2023.

[11] J. V. Solórzano, J. F. Mas, J. A. Gallardo-Cruz, Y. Gao, and A. F.-M. de Oca, "Deforestation detection using a spatio-temporal deep learning approach with synthetic aperture radar and multispectral images," *ISPRS J. Photogramm. Remote Sens.*, vol. 199, pp. 87–101, 2023.

[12] R. E. Kennedy et al., "Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects," *Remote Sens. Environ.*, vol. 113, no. 7, pp. 1382–1396, 2009.

[13] M. Gomroki, M. Hasanlou, and P. Reinartz, "STCD-EffV2T UNet: Semi transfer learning efficientnetV2 T-UNet network for urban/land cover change detection using sentinel-2 satellite images," *Remote Sens.*, vol. 15, no. 5, 2023, Art. no. 1232.

[14] P. Du, S. Liu, P. Gamba, K. Tan, and J. Xia, "Fusion of difference images for change detection over urban areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1076–1086, Aug. 2012.

[15] M. Arif, S. Sengupta, S. Mohinuddin, and K. Gupta, "Dynamics of land use and land cover change in peri urban area of Burdwan city, India: A remote sensing and GIS based approach," *GeoJ.*, vol. 88, pp. 4189–4213, 2023.

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

[17] E. Pedzisai, O. Mutanga, J. Odindi, and T. Bangira, "A novel change detection and threshold-based ensemble of scenarios pyramid for flood extent mapping using Sentinel-1 data," *Heliyon*, vol. 9, no. 3, 2023, Art. no. e13332.

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

[19] J. Deng, K. Wang, Y. Deng, and G. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, 2008.

[20] C. He, A. Wei, P. Shi, Q. Zhang, and Y. Zhao, "Detecting land-use/land-cover change in Rural–Urban fringe areas using extended change-vector analysis," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 13, no. 4, pp. 572–585, 2011.

[21] C. Wu, B. Du, and L. Zhang, "Slow feature analysis for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2858–2874, May 2014.

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

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

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

[25] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603216, doi: 10.1109/TGRS.2021.3066802.

[26] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[27] A. Codegoni, G. Lombardi, and A. Ferrari, "TinyCD: A (not so) deep learning model for change detection," *Neural Comput. Appl.*, vol. 35, no. 11, pp. 8471–8486, 2023.

[28] Q. Guo, J. Zhang, S. Zhu, C. Zhong, and Y. Zhang, "Deep multiscale siamese network with parallel convolutional structure and self-attention for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406512, doi: 10.1109/TGRS.2021.3131993.

[29] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816, doi: 10.1109/TGRS.2021.3085870.

[30] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3867–3878, 2023, doi: 10.1109/JSTARS.2023.3264802.

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

[32] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8007805, doi: 10.1109/LGRS.2021.3056416.

[33] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "ADS-Net: An attention-based deeply supervised network for remote sensing image change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, 2021, Art. no. 102348.

[34] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409818, doi: 10.1109/TGRS.2022.3159544.

[35] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403718, doi: 10.1109/TGRS.2021.3091758.

[36] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628711, doi: 10.1109/TGRS.2022.3199502.

[37] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607514, doi: 10.1109/TGRS.2021.3095166.

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

[39] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5608020, doi: 10.1109/TGRS.2022.3176603.

[40] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713, doi: 10.1109/TGRS.2022.3160007.
[41] W. Wang, X. Tan, P. Zhang, and X. Wang, "A CBAM based multiscale transformer fusion approach for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6817–6825, 2022, doi: 10.1109/JSTARS.2022.3198517.
[42] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519, doi: 10.1109/TGRS.2022.3169479.
[43] X. Song, Z. Hua, and J. Li, "Remote sensing image change detection transformer network based on dual-feature mixed attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5920416, doi: 10.1109/TGRS.2022.3209972.
[44] T. Yan, Z. Wan, and P. Zhang, "Fully transformer network for change detection of remote sensing images," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1691–1708.
[45] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, "An attention-based multiscale transformer network for remote sensing image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 599–609, 2023.
[46] Q. Ke and P. Zhang, "Hybrid-transcd: A hybrid transformer remote sensing image change detection network via token aggregation," *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 4, 2022, Art. no. 263.
[47] S. S. Islam, S. Rahman, M. M. Rahman, E. K. Dey, and M. Shoyaib, "Application of deep learning to computer vision: A comprehensive study," in *Proc. 5th Int. Conf. Inform., Electron. Vis.*, 2016, pp. 592–597.
[48] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017, doi: 10.1109/TGRS.2016.2612821.
[49] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, 2016.
[50] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 3–22, 2018.
[51] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, 2018.
[52] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. With Appl.*, vol. 169, 2021, Art. no. 114417.
[53] R. Zhang, H. Zhang, X. Ning, X. Huang, J. Wang, and W. Cui, "Global-aware siamese network for change detection on remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 199, pp. 61–72, 2023.
[54] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support, 4th Int. Workshop, 8th Int. Workshop, Held Conjunction MICCAI*, 2018, pp. 3–11.
[55] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.
[56] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[58] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489–1500, Feb. 2023.
[59] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–12090.
[60] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[61] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610111, doi: 10.1109/TGRS.2023.3277496.
[62] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1501–1510.
[63] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3899–3908.
[64] L. Shen et al., "S2looking: A satellite side-looking dataset for building change detection," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 5094.
[65] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[66] M. Contributors, "MMCV: OpenMMLab computer vision foundation," 2018. [Online]. Available: https://github.com/open-mmlab/mmcv
[67] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[68] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[69] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

Yonghui Tan received the B.S. degree in surveying and mapping engineering from the Hunan University of Science and Engineering, Yongzhou, China, in 2022. He is currently working toward the M.S. degree in surveying and mapping with the School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang, China.

His research interests include deep learning in remote sensing image processing, such as supervised or semisupervised remote sensing change detection.

Xiaolong Li received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2014.

He has been an Associate Professor with the School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang, China. He has authored or coauthored more than 20 scientific publications in international journals and conferences. His research interests include remote sensing image analysis, computer vision, and knowledge graph.

Yishu Chen received the M.S. degree in cartography and geographic information system from the East China University of Technology, Nanchang, China, in 2023.

She is currently working with Ningbo Alatu Digital Technology Company, Ltd., Ningbo, China, as an Assistant Engineer. Her research interests include the intelligent interpretation of remote sensing images based on deep learning.

Jinquan Ai received the Ph.D. degree in cartography and geographic information systems from East China Normal University, Nanchang, China, in 2018.

He is currently a Lecturer with the East China University of Technology, Nanchang. His main research interests include remote sensing image processing, deep learning, and wetland remote sensing.