A Novel Remote Sensing Image Change Detection Approach Based on Multilevel State Space Model

Zhongyu Zhang, Xuanmei Fan, Xin Wang, Yingxiang Qin, and Junshi Xia, *Senior Member, IEEE*

***Abstract*— Remote sensing image change detection (CD) is crucial for disaster assessment, land use change, and urban management. Most CD methods are realized by CNN and Transformer. However, these methods are not satisfied with modeling global dependencies while keeping a low computational complexity. Recently, the emergence of Mamba architectures based on state space models (SSMs) can remedy the above problems. In this article, we propose a visual Mamba-based multiscale feature extraction network to efficiently interactively fuse global and local information, which is named as MF-VMamba (MF: multiscale feature). First, a VMamba-based encoder is used to extract multiscale semantic features from bitemporal images. Then, a feature enhancement module (FEM) is proposed to capture the difference information between images. In addition, we employ a multilevel attention decoder (MAD) based on large kernel convolution (LKC) to obtain the information in spatial and spectral dimensions to realize the information interaction between global and local features. After the sequential processing of these three modules, the discriminative ability of changing objects is significantly improved. Notably, the computational complexity of our VMamba-based model grows linearly, which can significantly reduce the computational cost. In the experiments, our method performs well on CDD, DSIFN-CD, LEVIR-CD, and SYSU-CD datasets, with *F*1 scores and OA reaching 95.69%/88.05%/90.64%/86.95% and 98.97%/96.01%/99.07%/90.75%, respectively. The code can be accessed at https://github.com/121zzy/MF-Mamba.git.**

***Index Terms*— Change detection (CD), high-resolution remote sensing image, large kernel convolution (LKC), state space model (SSM), VMamba.**

I. Introduction

With the growth of the global population and the intensification of human activities, land cover, and utilization patterns on the Earth's surface have undergone significant changes. These changes not only affect the ecological environment, but also have far-reaching implications for economic development and social stability. For example, accelerated urbanization has led to the development of agricultural land and forests for use as urban land, and there is a need to focus on the sustainable use and management of natural resources. In addition, the frequent occurrence of natural disasters requires timely and effective monitoring and assessment of their impacts in order to develop rational responses. Remote sensing image change detection (CD) refers to the use of remote sensing data from different periods to monitor the same geographic area, to extract and describe the change features of objects or phenomena of interest, and to quantitatively analyze and judge their changes. This method has a wide range of applications in many fields, such as land use change [1], [2], natural disaster detection [3], [4], and urban land management [5], [6], [7].

With the development of remote sensing technology, the emergence of hyperspectral remote sensing images (HRSIs) and very high-resolution (VHR) images has greatly contributed to the advancement of automated remote sensing detection techniques. Lv et al. [8] and [9] used an adaptive region approach to measure the change magnitude between bitemporal HRSIs, while Wu et al. [10] proposed a multitask hyperspectral CD framework. These methods fully utilize the rich spectral information of HRSIs to improve the accuracy of CD. However, HRSIs impose greater demands on computational resources due to their large data dimensions and high computational complexity, while VHR images are still a more commonly used choice in remote sensing due to lower computational requirements and more mature processing techniques in practical applications. Currently, the more mainstream deep learning models in the field of remote sensing are convolutional neural networks (CNNs) [11] and transformers [12]. CNN can effectively extract local features in the image, such as edges, texture, and shape. With the focus on spatial and temporal scales, recent work has focused on increasing the receptive field of the model by using dilation convolution [13] and attention mechanisms such as Siamese-based spatial–temporal attention neural network (STANet) [14], dual attentive fully convolutional Siamese networks (DASNets) [15], a dual-task-constrained deep Siamese convolutional network (DTCDSCN) [16], and the combination of Siamese network and NestedUNet (SNUNet) [17]. Although methods based on attention mechanisms are able to model global information, they are still

Received 12 July 2024; revised 21 October 2024; accepted 12 November 2024. Date of publication 18 November 2024; date of current version 6 December 2024. This work was supported in part by the National Science Fund for Distinguished Young Scholars of China under Grant 42125702; in part by the National Natural Science Foundation of China under Grant 42307263; in part by the New Cornerstone Science Foundation through the XPLORER Prize under Grant XPLORER-2022-1012; in part by the Natural Science Foundation of Sichuan Province under Grant 2024ZDZX0020, Grant 2022NSFSC0003, and Grant 2022NSFSC1083; and in part by the State Key Laboratory of Geohazard Prevention and Geoenvironment Protection Independent Research Project under Grant SKLGP2021Z027 and Grant SKLGP2021Z023. *(Corresponding authors: Xuanmei Fan; Xin Wang.)*

Zhongyu Zhang, Xuanmei Fan, and Xin Wang are with the State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu 610059, China (e-mail: fxm_cdut@qq.com; wangxrs@126.com).

Yingxiang Qin is with the College of Geography and Planning, Chengdu University of Technology, Chengdu 610059, China.

Junshi Xia is with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan.

Digital Object Identifier 10.1109/TGRS.2024.3501303

limited in capturing global contextual information. In contrast, transformer is able to capture global contextual information and model long-range dependencies through the self-attention mechanism. Transformer has amazing semantic expressiveness and has been applied in various computer vision tasks such as image classification [18], [19], semantic segmentation [20], and target detection [21], [22]. vision transformer (ViT) [18] is the first purely transformer method for image classification. Swin Transformer [23] reduces the complexity of self-attention by shifted window. However, transformer-only-based models tend to ignore local detailed features. A hybrid CNN-transformer-based model facilitates the extraction of global and local information. Chen et al. [24] proposed a bitemporal image transformer (BIT) that used ResNet as the backbone network to extract features and proposed BIT to model contexts within spatial–temporal domain, and achieved better results.

Despite the advantages, transformer-based models and hybrid models are more computationally intensive and resource intensive. Recently, the Mamba model constructed based on state space model (SSM) [25] has gained much attention. Mamba [26] aims to achieve linear time complexity and improves the computational speed, and at the same time, solves the memory challenge of transformer. Mamba is suitable for larger data processing and real-time processing, and has certain advantages for high-resolution and large-scale remote sensing images. The current work based on Mamba includes VMamba [27], ChangeMamba [28], RS-Mamba [29], etc. VMamba [27] enhances the effective sensing field in a specific direction by selective scanning in both horizontal and vertical directions through SSM, but it is not suitable for VHR remote sensing images. ChangeMamba [28] proposes three kinds of visual mamba-based architecture networks for three different CD tasks. RS-Mamba [29] employs an omnidirectional selective scanning module for better global modeling.

Although all the above methods can effectively improve the detection accuracy and obtain global information, the detail information of high-resolution optical images should not be neglected, and it is necessary to interactively fuse the global and local detail information to achieve better detection results. In this article, we propose a VMamba-based multiscale feature extraction network for remote sensing image CD (simplified as MF-VMamba). The method utilizes an encoder constructed by visual mamba for global feature extraction of bitemporal image pairs, which is conducive to reducing computational complexity and memory occupation. Meanwhile, a feature enhancement module (FEM) is introduced to extract the difference information between images, so as to obtain the local information effectively. In addition, in order to enable interaction between global and local features, they are simultaneously fed into a multilevel attention decoder (MAD) based on a feature pyramid structure as a way to facilitate cross-branch feature fusion, and finally produce a change prediction map by a classifier.

The main contributions of our work are summarized as follows.

1) A VMamba-based CD framework (MF-VMamba) is proposed to introduce Mamba into the CD task, which is able to better extract global and local information in bitemporal images and capture information in both spatial and spectral dimensions, contributing to identifying change regions more accurately while reducing computational costs.

2) A MAD based on feature pyramid structure is designed, which incorporates the large kernel convolution (LKC) and is capable of fusing the long-range dependencies of global and local characteristics of each scale branch to achieve the interaction between spatial and channel information, effectively reducing the edge blurring problem.

3) Extensive experiments are conducted based on four public datasets, confirming the effectiveness and efficiency of the proposed approach on information interaction through the fusion strategy of Mamba and CNN.

II. Related Works

A. Change Detection

CD has become an important research direction in the field of remote sensing. Traditional CD methods include pixel-based and object-based methods, such as regression analysis [30], image ratio [31], image difference [32], change vector analysis (CVA) [33], independent component analysis [34], and principal component analysis (PCA) [35], [36]. Although these methods can handle some specific tasks, they are relatively low in terms of efficiency and detection accuracy. In recent years, deep learning has been rapidly developing in the field of remote sensing, and was initially applied to tasks such as semantic segmentation, target detection, and image classification, and then was widely applied to CD tasks. In the early stage, Daudt et al. [37] initially introduced FCN to the CD field and proposed three models: FC-EF, FC-Siam-conc, and FC-Siam-diff. To enhance detection accuracy and extract finer features, Papadomanolaki et al. [38] proposed a fusion of U-Net [39] and long short-term memory (LSTM) [40] for extracting both spatial and temporal characteristics from the image. Cheng et al. [41] introduced a separable deep learning network (ISNet) consisting of spatial attention module (SAM) and channel attention module (CAM) to emphasize both semantic and spatial information. However, data distribution in CD often suffers from imbalance, and some networks may struggle to effectively differentiate changed from unchanged areas. To solve this problem, Hou et al. [42] used HRNet as the backbone network, taking the difference map of the bitemporal image and the original image as input. This enables them to learn features from bitemporal images and capture temporal information that reflects changes in the image over time. Chen et al. [15] presented the dual attention fully convolutional Siamese network (DASNet), which enhanced model perception of change information through the introduction of dual attention mechanisms while addressing sample imbalance issues. Chen and Shi [14] improved remote sensing image CD performance by introducing a Siamese network and self-attention mechanism (STANet) along with a multiscale partitioning strategy to better understand the relationships between pixels at different positions and times.

Despite significant advances in CNN-based CD techniques, there are still limitations in modeling long-range dependencies. With the emergence of transformers, the CD task has taken a step forward. Chen et al. [24] proposed a BIT capable of efficiently modeling the context in the spatiotemporal domain. Bandra and Patel [43] introduced a Siamese network based on transformer architecture (ChangeFormer), combining transformer encoders and multilayer perceptron (MLP) decoders to handle remote sensing image CD effectively. This architecture can capture multiscale features and long-range dependencies. Li et al. [44] proposed a hybrid transformer model (TransUNetCD) that combines the advantages of transformer and U-Net to reduce redundant information. Although the transformer has achieved better results in terms of accuracy and is able to model global context information, it is computationally intensive and has a quadratic growth in resource usage. Recently, the emerging Mamba has become an alternative to transformers. It is not only less computationally intensive but also capable of modeling global dependencies, opening up new possibilities for the CD task.

B. State Space Model

SSM was originally inspired by modern control system theory and was widely applied with the emergence of the S4 [25] model. The combination of SSM and deep learning improves the ability of the model to capture long-range dependencies with linear scaling properties. Gu and Dao [26] proposed a variant called Mamba by merging the selection mechanism into the SSM, which was superior to transformer in large-scale data processing, and then rapidly applied to image classification [21], [45], object detection [22], medical segmentation, etc. Currently, research areas based on Mamba for remote sensing images include image classification (RS-Mamba [29], SpectralMamba [46], and S2Mamba [47]), semantic segmentation (Samba [48], RS3Mamba [49], and CM-UNet [50]), hyperspectral denoising (HSIDMamba [51] and SSUMamba [52]), CD (ChangeMamba [28] and CDMamba [53]), and object detection. Although there are many related research fields now, the research on Mamba is still at a low level, and the research on multitemporal image remote sensing images needs to be further explored.

C. Large Kernel Convolution

LKC is a type of convolution operation in CNNs, which is characterized by the use of a larger size kernel for feature extraction. It is different from the traditional attention of a small kernel; the use of a large kernel means that it can encompass a larger number of input elements at the same time and thus capture a larger range of contextual information. In some cases, the use of LKC can reduce the number of network layers required, thereby reducing computational complexity and training time.

The proposal of LKC was inspired by VGGNet [54] by increasing the receptive field layer by layer; Szegedy et al. [55] proposed the inception module, which introduced the concept of LKC with its multibranch structure. Its multibranching structure introduced the concept of large sum convolution. Yu and Koltun [56] proposed dilated convolution, which expands the receptive field without increasing the number of parameters by introducing a dilation between convolution kernels. Chen et al. [57] proposed atrous spatial pyramid pooling (ASPP) module, which can significantly increase the receptive field and feature representation by capturing features at different dilation rates through multiscale dilation convolution (i.e., cavity convolution). Ding et al. [58] use 31×31 LKC, which can significantly increase the receptive field and feature expression by capturing the global background and enhance feature representation. Liu et al. [59] use kernel decomposition and sparse techniques to extend the kernel size to 51 × 51.

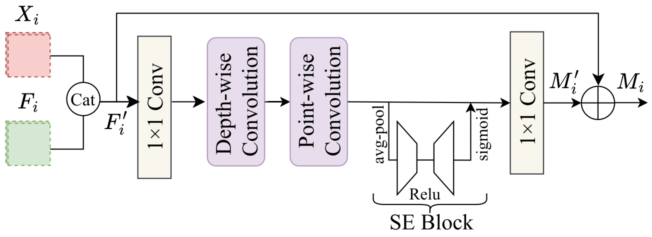


Fig. 1. LKC block. X_i denotes the output features of each stage and F_i denotes the enhanced features after the FEM, as shown in Fig. 2.

In this article, we use depth-wise separable convolution (DSC) block [60] (as shown in Fig. 1), which expands the receptive field by using dilation convolution, and decomposes the standard convolution into depth-wise convolution (DWConv) and point-wise convolution, which greatly reduces the amount of computation and the number of parameters. This block is able to capture more global contextual information and inter-dependencies between features and enhance the feature representation while maintaining computational efficiency.

III. Methods

A. Preliminaries: SSM

SSM is a model for describing the state representation of a sequence at each time step and predicting its next state based on the inputs. Specifically, the system maps a 1-D input function or sequence $x(t) \in \mathbb{R}$ to an output sequence $y(t) \in \mathbb{R}$ via a hidden state representation $h(t) \in \mathbb{R}^L$. This system is usually defined by linear ordinary differential equations (ODEs) [26]

$$\begin{cases} h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) = \mathbf{C}h(t) \end{cases} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$, and $\mathbf{B}$, $\mathbf{C} \in \mathbb{R}^{N\times L}$ are the system matrices.

In deep learning continuous time systems, the continuous equations have to be discretized in order to improve the computational efficiency. The discretization of SSM requires the conversion of the continuous time parameters ($\mathbf{A}$ and $\mathbf{B}$) into their discrete counterparts ($\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$) on a specified time scale (parameter $\triangle$). This discretization is called a zero-order hold (ZOH) approach [26]

$$\begin{cases} \overline{\mathbf{A}} = e^{\triangle \mathbf{A}} \\ \overline{\mathbf{B}} = (\triangle \mathbf{A})^{-1}\left(e^{\triangle \mathbf{A}} - \mathbf{I}\right) \cdot \triangle \mathbf{B} \end{cases} \tag{2}$$

$$\begin{cases} h'(t) = \overline{\mathbf{A}}h(t) + \overline{\mathbf{B}}x(t) \\ y(t) = \mathbf{C}h(t). \end{cases} \tag{3}$$

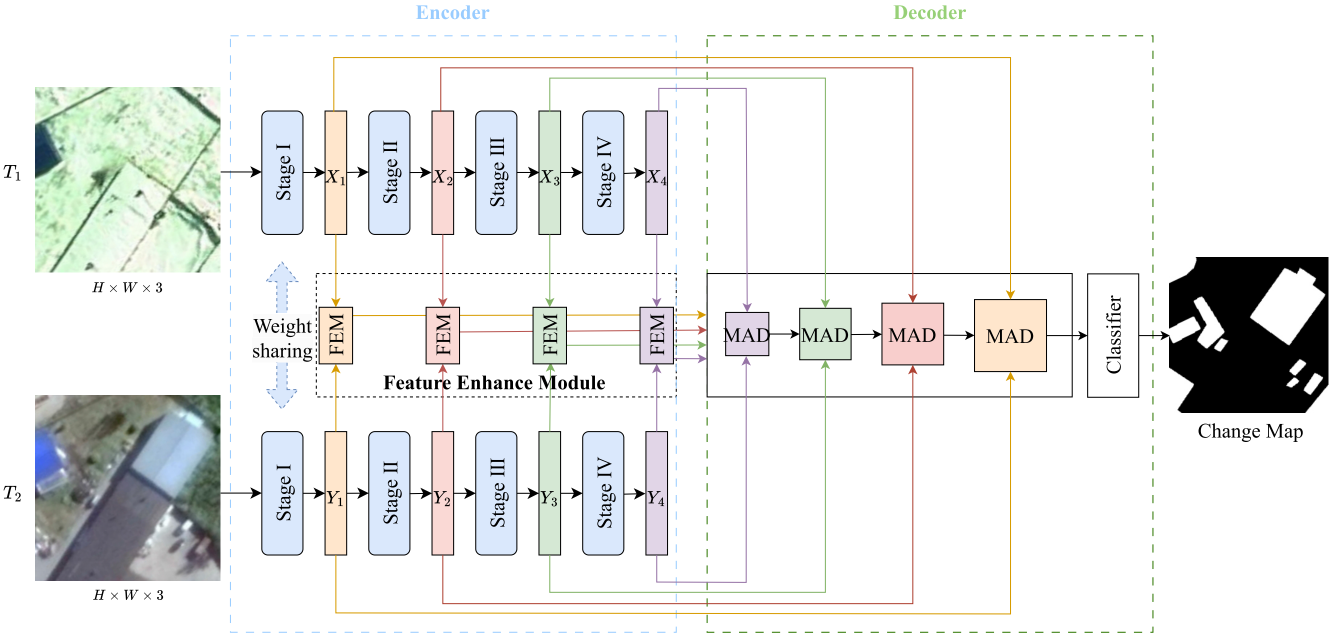


Fig. 2. Overview of the proposed MF-VMamba. X_i and Y_i denote the output features of each stage.

The Mamba architecture proposes a selective scanning mechanism based on S4, which is able to dynamically adjust the internal state of the system based on the input sequences, for example, filtering out irrelevant information for long sequence modeling. This can improve computational efficiency and less memory overhead.

B. Overview

In this section, we introduce the implementation details of MF-VMamba. MF-VMamba consists of three components, that is, VSS encoder, FEM, and MAD. Specifically, the MF-VMamba model extracts features through a weight sharing encoder based on the VMamba architecture, while the FEM extracts the difference information of each pair of images at different scale branches, and the features generated by the VSS encoder and the FEM are sent to the MAD at the corresponding scales of the decoder to perform the feature fusion. After feature extraction and feature fusion at four scales, the prediction map is generated. The overall structure diagram of the network is shown in Fig. 2.

C. Encoder Based on Visual SSM

Our proposed encoder is constructed based on VMamba [27], and the specific structure of the network is shown in Fig. 3(a). The core block of VMamba is the VSS block, in which SS2D is the core computational unit of the VSS block. VMamba is able to perform selective scanning in both horizontal and vertical directions, and is able to obtain spatial context information from different directions. The VSS encoder part is divided into four stages; stage I first passes the input data through a stem module, and then uses the VSS block to model the global context information. The following three stages all first downsample the input data, and then extract features using the VSS block. The output features of each stage are denoted by X_i and Y_i.

The output of VSS block needs to be obtained by summing the output of two information streams. One stream goes through a linear layer, a 3 × 3 DWConv, a SiLU activation, the 2D-selective scan module (SS2D), and the LN function. Finally, the outputs obtained from the two streams are combined by product, followed by a linear layer that sums the original inputs to obtain the final output X_i or Y_i.

D. Feature Enhancement Module

Since the encoder mainly focuses on extracting global features, some local detail information may be lost. To address this issue, we introduce an FEM to capture the difference information of each pair of images across different scale branches. The detailed structure is shown in Fig. 4. This module applies difference and summation operations at each layer of features to enhance feature extraction. The FEM is applied to each pair of features obtained from the four stages of the encoder. For clarity, we use the FEM in the first stage as an example. Let X_1 and Y_1 represent the features extracted by the encoder in the first stage from the two input images, respectively. The FEM first computes the summation and difference between X_1 and Y_1. The summation integrates fine local and global information, while the difference captures the local variations between the two images. Next, the resulting summed and differenced features pass through a sequence of operations: a 1 × 1 convolutional layer, followed by a batch normalization (BN), a rectified linear unit (ReLU) activation, a max pooling layer, an average pooling layer, and finally a fully connected layer. These operations produce two feature maps F_{diff}^1 and F_{sum}^1. Finally, the two enhanced feature maps are concatenated to form the output feature map F_1. This process not only preserves both local and global information but

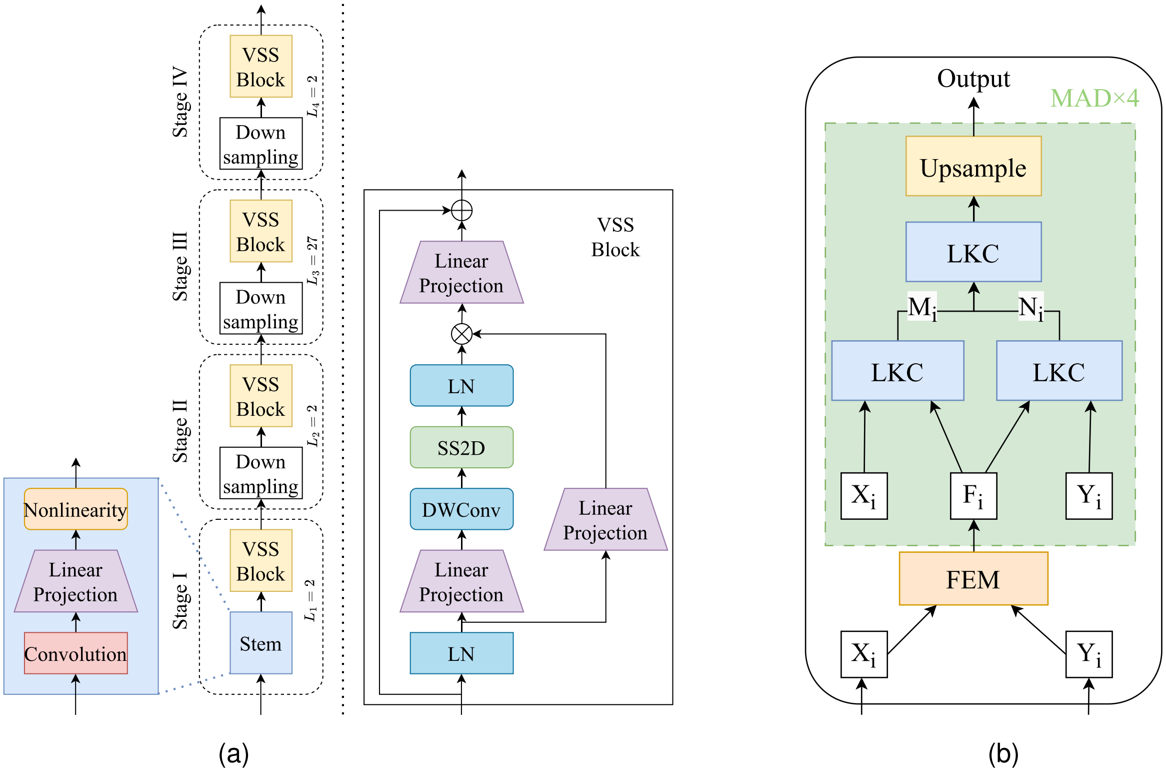


Fig. 3. Encoder and decoder part of the overall network structure diagram. The encoder part is mainly based on VMamba, divided into four branches, using VSS block. The decoder part uses LKC and FEM. (a) Encoder network. (b) Decoder network.

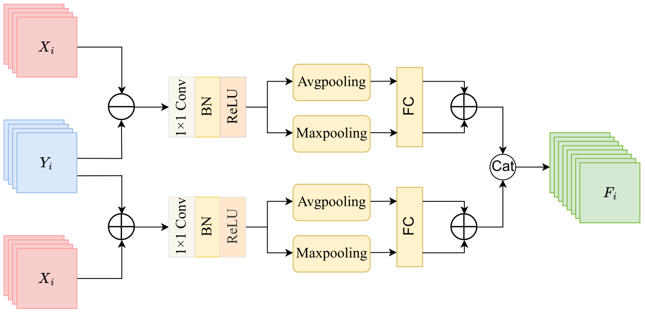


Fig. 4. Specific structure of the FEM. X_i and Y_i denote the output features of each stage.

also strengthens the model's ability to differentiate fine-grained details from global structures. The enhanced features from the FEM can be represented as follows:

$$\begin{cases} \hat{F}^i_{\text{diff}} = \text{Relu}(\text{BN}(\text{Conv}(X_i - Y_i))) \\ \hat{F}^i_{\text{sum}} = \text{Relu}(\text{BN}(\text{Conv}(X_i + Y_i))) \\ F^i_{\text{diff}} = \text{FC}\left(\text{Maxpool}\left(\hat{F}^i_{\text{diff}}\right) + \text{Avepool}\left(\hat{F}^i_{\text{diff}}\right)\right) \\ F^i_{\text{sum}} = \text{FC}\left(\text{Maxpool}\left(\hat{F}^i_{\text{sum}}\right) + \text{Avepool}\left(\hat{F}^i_{\text{sum}}\right)\right) \end{cases} \quad (4)$$

$$F_i = \text{Concat}\left(F^i_{\text{diff}}, F^i_{\text{sum}}\right) \quad (5)$$

where X_i and Y_i denote the hierarchical features extracted from T_1 image and T_2 image through the encoder, Concat(·) indicates the concatenation operation at the channel-wise, and Maxpool(·) and Avepool(·) mean the max and average pooling layer, respectively, F^i_{diff} and F^i_{sum} denote the features that are subjected to the difference and sum operation for each level of features, and F_i denotes the fusion features obtained through FEM. By applying the FEM at each stage of the encoder, the model achieves a more comprehensive feature representation, capturing both global and local information across multiple scales. This improves the network's ability to detect subtle changes and maintain overall context.

E. Multilevel Attention Decoder

We utilize a feature pyramid structure with DSC [60] block, as shown in Fig. 3(b), to capture spatial and channel position information, which is used to obtain a wider range of spatial information and a better understanding of correlation and contextual information between bitemporal images. Initially, the features X_i and Y_i obtained from each stage of the encoder are paired with the multistage enhancement features F_i, generated by the FEM, to serve as the inputs to the LKC module. After processing through the LKC, output features M_i and N_i are produced. These features M_i and N_i are then further processed through another LKC module, followed by an upsampling operation, to generate the final output features. LKC is a large kernel convolution containing DSC block (as shown in Fig. 1). First, we concatenate the two input features; subsequently, we perform a 1 × 1 convolution operation on the concatenated feature map F_i'. After that, we introduce dilated-wise convolution (DWC) to expand the receptive field to perform a convolution operation on each input channel to capture a larger range of contextual information, and use DSC to linearly combine each output obtained by DWC in the channel dimension. After squeeze-and-excitation (SE) block, 1 × 1 convolution to obtain M_i', finally, the processed feature

TABLE I

STATISTICS OF THE FOUR DATASETS, INCLUDING THE NUMBER OF CHANGED PIXELS AND THE NUMBER OF UNCHANGED PIXELS

Dataset		Spatial resolution	Size/image	Number of pixels			Number of images
				Changed pixel	Unchanged pixel	Imbalanced ratio	
CDD	train	0.03-1m/pixel	256×256	83,981,014	571,378,986	1:6.80	10,000
	val			24,676,497	171,800,431	1:6.96	3,000
	test			25,411,239	171,196,761	1:6.74	3,000
	total			134,068,750	914,376,178	1:6.82	16,000
LEVIR-CD	train	0.5m/pixel	256×256	21,412,341	445,203,349	1:20.79	7120
	val			28,162,258	64,292,596	1:22.83	2048
	test			6837,335	127,380,324	1:18.63	1024
	total			31,065,934	636,876,269	1:20.50	10,192
DSIFN	train	2m/pixel	256×256	75,030,941	142,655,301	1:1.90	3600
	val			6571,017	14,276,213	1:2.17	48
	test			499,903	2482,332	1:5.52	340
	total			82,051,861	159,413,846	1:1.94	3988
SYSU-CD	train	0.5m/pixel	256×256	167,833,309	618,598,691	1:3.68	12,000
	val			56,437,798	205,706,202	1:3.64	4000
	test			61,820,917	200,323,083	1:3.24	4000
	total			286,092,024	1024,627,976	1:3.58	20,000

map M_i' is fused with the original feature map in a fusion operation to obtain M_i.

By using DSC block, combined with dilated convolution and SE block, it promotes the interaction and integration of inter-channel information, facilitates the extraction of spatial features, and significantly improves the semantic information between images. At the same time, through the interactive fusion of different scales, the subtle features of the image can be captured in the local range, which improves the model's ability to perceive the local details, and is conducive to the reduction of the edge blurring problem.

F. Loss Function

In CD tasks involving remote sensing images, a common challenge is posed by class imbalance, where the quantity of changed samples and unchanged samples significantly differs, and the spatial arrangement of some data is not concentrated. Under these circumstances, the traditional contrastive loss function may suffer because it does not specifically take class imbalance into account. To deal with this problem, we design a hybrid loss function. It combines the weighted binary cross-entropy loss and the dice loss. The hybrid loss can be expressed as follows:

$$\text{Loss} = L_{\text{WBCE}} + L_{\text{dice}} \tag{6}$$

$$L_{\text{WBCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[w_0 y_i \log\left(\hat{y}_i\right) + w_1(1-y_i)\log\left(1-\hat{y}_i\right)\right] \tag{7}$$

$$L_{\text{dice}} = 1 - \frac{2\cdot Y \cdot \text{softmax}\left(\hat{Y}\right)}{Y + \text{softmax}\left(\hat{Y}\right)} \tag{8}$$

where L_{WBCE} represents the weighted binary cross-entropy loss, L_{dice} represents the dice loss. N is the number of samples, $\hat{y}_i$ represents the model's prediction for the ith sample, y_i denotes the true label for the ith sample, and w_i presents the weight parameter. In this weighted binary cross-entropy loss L_{WBCE}, the weights w_0 and w_1 are dynamically calculated based on the imbalance between changed and unchanged pixels in the dataset. Specifically, w_0 is set to the ratio of total pixels to changed pixels, and w_1 is the ratio of total pixels to unchanged pixels. This helps to balance the contribution of each class in the loss function. Table I counts the number of changing and unchanging image elements for the four different datasets. Y indicates the ground truth and $\hat{Y}$ represents the change map. The hybrid loss we mentioned can balance sample pairs of different categories and help prevent the network from being overly biased in learning unchanged samples.

IV. EXPERIMENT AND RESULTS

A. Datasets

In this study, four publicly available datasets with uneven sample distribution were selected, namely, the CDD dataset [61], the DSIFN-CD dataset [62], the LEVIR-CD dataset [14], and the SYSU-CD dataset [63]. Statistical information about datasets can be viewed in Table I.

1) The CDD dataset [61] is a publicly accessible dataset. The dataset consists of 11 multitemporal image pairs, of which 7 seasonal variation image sets have a resolution of up to 4750×2700 pixels and the other 4 image sets have a resolution of up to 1900 × 1000 pixels. The resolution of the multitemporal image pairs ranges from 3 to 100 cm/pixel, thus taking into account the effects of seasonal variations. To create this dataset, Ji et al. [64] used random cropping and partial rotation to transform the images to 256 × 256 pixels. It contains a training set of 10 000 pairs, a test set of 3000 pairs, and a validation set of 3000 pairs.

2) The DSIFN-CD dataset [62] is a publicly accessible binary CD dataset, consisting of 3988 remote sensing image pairs. These images are from six major cities in China: Xi'an, Chongqing, Beijing, Chengdu, Wuhan, and Shenzhen. The dataset covers a variety of land cover changes, including

water bodies, cultivated land, buildings, and roads. We use the default clipping samples with a size of 256 × 256, containing 3600 training pairs, 340 test pairs, and 48 validation pairs.

3) The LEVIR-CD dataset [14] is a publicly available large-scale RS dataset consisting of 637 pairs of VHR images with a size of 1024 × 1024 pixels and a spatial resolution of 0.5 m/pixel. We crop the default images into non-overlapping patches of size 256 × 256, and the sample set consists of 7120 pairs of training set, 1024 pairs of test set, and 2048 pairs of validation set.

4) The SYSU-CD dataset [63] is a large-scale remote sensing dataset. It contains 20 000 pairs of bitemporal images with a spatial resolution of 0.5 m, captured at different times, and includes a variety of objects such as buildings, roads, ships, and vegetation. The dataset is divided into 12 000 image pairs for training, 4000 for validation, and 4000 for testing, providing a comprehensive resource for evaluating CD models.

B. Experimental Setup

Experimental details: All of our comparison experiments, including the MF-VMamba and ablation experiments, are realized using PyTorch on a single NVIDIA GeForce RTX 2080 with 11 GB memory workstation. All four datasets we used were cropped to 256 × 256 pixels, and the data enhancement methods used include flipping, transposing, and swapping bitemporal images. During training, we use an Adam optimizer with an initial learning rate of 0.001, a weight decay of $5e^{-3}$, a batch size set to 4, and the number of training iterations set to 240 000. Validation was performed every 500 iterations to select the best model weights. Finally, the model's accuracy was evaluated on the test dataset. All comparison methods were implemented in the same environment, with hyperparameters set to the recommended values from GitHub. Also, our code can be accessed through https://github.com/121zzy/MF-Mamba.git.

Evaluation metrics: In order to evaluate the performance of the proposed model, we apply five key evaluation metrics: recall (Rec), overall accuracy (OA), precision (Pre), intersection over union (IoU), and $F1$ score ($F1$). The definitions of the above metrics are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

$$\text{OA} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{13}$$

where TP, TN, FP, and FN represent the counts of true positives, true negatives, false positives, and false negatives, respectively.

C. Comparison Methods

In order to demonstrate the effectiveness of our proposed method, we have selected some existing CD algorithms for comparison, including CNN-based methods (FC-EF [37], FC-Siam-diff [37], FC-Siam-conc [37], DTCDSCN [16], SNUNet [17], ISNet [41], SARAS-Net [17]), transformer-based methods (BIT [24] and ChangeFormer [43]), and the latest Mamba-based methods (ChangeMamba [28], RS-Mamba [49], and CDMamba [53]). We use MF-VMamba to detect changes in the four datasets: CDD, DSIFN-CD, LEVIR-CD, and SYSU-CD. The quantitative comparison results of the four datasets are shown in Tables II and III. The visualization comparison results of the four datasets are shown in Figs. 5–8. Different colors are used in the figure to indicate the detection results: white represents TP (i.e., changed), black corresponds to TN (i.e., unchanged), red indicates FP, and green indicates FN.

Quantitative results: As can be seen from the comparisons in Tables II and III, our proposed method, MF-VMamba, achieves excellent performance on key metrics (e.g., $F1$, OA, and IoU) on most datasets. Specifically, MF-VMamba consistently outperforms traditional CNN-based methods. For example, on the CDD dataset, the OA of MF-VMamba is 0.07%, 0.08%, and 0.62% higher than that of SNUNet, DTCDSCN, and ISNet, respectively. This demonstrates the advantages of Mamba-based architectures in handling large-scale CD by integrating spatial and contextual information more efficiently. While MF-VMamba performs well on most datasets, its improvement on the DSIFN-CD dataset is not significant. On this dataset, MF-VMamba achieves the OA of 96.01%, which is only 0.42% higher than DTCDSCN. In addition, ChangeMamba outperforms MF-VMamba in terms of $F1$ score, achieving 90.21%, which is 2.16% higher than MF-VMamba. This suggests that, while MF-VMamba provides good performance on high-resolution and diverse datasets, it may not perform well on datasets with lower spatial resolution or limited samples. This is because deep learning models typically rely on large amounts of data for training to capture complex patterns and improve generalization [23]. Datasets with smaller or poorer quality data may not provide enough information for the model, resulting in limited model performance. It performs well on datasets such as LEVIR-CD, obtaining the highest $F1$ score (90.64%) and OA (99.07%), outperforming all other models. Using the Mamba-based mechanism and the FEM, MF-VMamba is able to capture long-range dependencies and complex spatial patterns, which is crucial for accurate CD in large-scale scenarios such as LEVIR-CD dataset. On the SYSU-CD dataset, MF-VMamba has an IoU of 77.59% and an OA of 90.75%. Although it lags behind CDMamba in terms of $F1$ scores, it still outperforms the other models in a comprehensive comparison.

Qualitative results: Figs. 5–8 show the visualization comparison results of some of the methods on the CDD, DSIFN-CD, LEVIR-CD, and SYSU-CD datasets. It is evident that methods based on U-Net exhibit more pronounced occurrences of false negatives and false positives. The introduction of attention mechanisms significantly mitigates this issue. Although the methods can generally locate all objects that have changed, they still struggle with detecting small area samples and exhibit some challenges in recognizing changes. This may be due to irrelevant variations caused by seasonal differences

TABLE II

Comparison of Our Model With Other Methods on CDD, DSIFN-CD, and LEVIR-CD Datasets. The Top Two Results Are Highlighted in Red and Green. All Results Are Described as Percentages (%)

Type	Method	CDD					DSIFN-CD					LEVIR-CD				
		Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA
CNN-based	FC-EF$_{18}$	79.66	55.44	65.38	48.57	93.07	62.67	66.18	64.38	47.47	87.55	79.16	83.26	81.16	68.29	98.12
	FC-Sima-diff$_{18}$	91.36	57.22	70.37	54.29	94.31	68.22	56.87	62.03	44.96	88.17	89.31	78.88	83.77	72.08	98.44
	FC-Sima-conc$_{18}$	87.19	62.26	72.65	57.05	94.47	62.18	69.01	65.42	48.61	87.60	86.21	86.31	86.26	75.85	98.60
	DTCDSCN$_{21}$	94.93	93.35	94.14	88.93	98.63	88.32	86.13	87.21	76.58	95.59	92.32	87.32	89.75	81.41	98.98
	SNUNet$_{22}$	95.54	95.15	95.34	91.11	98.90	84.88	75.01	79.64	66.17	93.48	92.05	88.80	90.40	82.48	99.04
	ISNet$_{22}$	95.99	95.27	95.63	91.84	98.35	75.63	82.68	78.42	67.01	88.69	95.19	94.22	94.70	90.32	98.98
	SARAS-Net$_{23}$	93.67	92.72	93.19	87.33	98.26	67.65	67.51	67.58	51.04	89.01	90.89	85.98	88.37	79.16	98.84
Transformer-based	BIT$_{22}$	95.19	93.56	94.37	89.34	98.68	75.01	68.30	71.49	55.64	90.74	92.98	85.76	89.23	80.55	98.94
	ChangeFormer$_{22}$	94.20	94.22	94.21	89.06	98.63	88.78	85.95	87.34	77.53	95.76	91.66	89.29	90.46	82.58	99.04
Mamba-based	ChangeMamba$_{24}$	96.44	93.43	94.91	90.32	98.81	91.23	89.21	90.21	82.17	94.39	91.01	89.36	90.18	82.07	99.01
	RS-Mamba$_{24}$	93.61	93.58	93.59	87.96	98.49	88.48	84.94	86.67	76.48	95.56	91.11	88.08	89.56	81.06	98.96
	CDMamba$_{24}$	94.78	94.90	94.84	90.20	98.83	87.98	83.74	85.81	75.13	95.17	91.42	89.42	90.41	82.25	99.02
	MF-VMamba(ours)	95.46	95.93	95.69	91.58	98.97	90.08	86.11	88.05	78.73	96.01	92.49	88.87	90.64	82.89	99.07

TABLE III

Comparison of Our Model With Other Methods on the SYSU-CD Dataset. The Top Two Results Are Highlighted in Red and Green. All Results Are Described as Percentages (%)

Type	Method	SYSU-CD				
		Pre	Rec	F1	IoU	OA
CNN-based	FC-EF$_{18}$	83.77	80.73	82.08	70.83	87.64
	FC-Sima-diff$_{18}$	84.16	69.97	73.47	61.00	84.49
	FC-Sima-conc$_{18}$	86.82	81.97	84.01	73.50	89.20
	DTCDSCN$_{21}$	87.16	84.38	85.64	75.73	90.03
	SNUNet$_{22}$	77.60	80.02	78.79	65.00	89.84
	ISNet$_{22}$	87.21	84.07	85.48	75.51	89.96
	SARAS-Net$_{23}$	74.70	63.98	68.92	52.58	86.39
Transformer-based	BIT$_{22}$	86.16	83.67	84.81	74.55	89.42
	ChangeFormer$_{22}$	87.67	84.78	86.09	76.38	90.35
Mamba-based	ChangeMamba$_{24}$	82.17	77.03	79.52	66.00	90.64
	RS-Mamba$_{24}$	77.60	80.02	78.79	65.00	89.84
	CDMamba$_{24}$	86.04	83.70	87.78	74.50	89.38
	MF-VMamba(ours)	87.07	86.83	86.95	77.59	90.75

TABLE IV

Ablation Experiments on the Different Modules on the LEVIR-CD Dataset. All Results Are Described as Percentages (%)

FEM	MAD	BASE	F1	OA	IoU
×	×	✓	76.88	98.21	62.42
✓	×	✓	77.46	98.49	63.27
×	✓	✓	77.61	98.52	63.48
✓	✓	✓	**90.64**	**99.07**	**82.89**

and lighting conditions. Additionally, because the sample data (such as vehicles) are relatively small, the detected change areas remain incomplete and inaccurate. MF-VMamba is able to partially solve the problem of pseudo-changes, and the boundaries of the detected change objects are clearer, with better inter-object semantic consistency. The results from the CDD and LEVIR-CD datasets demonstrate that MF-VMamba yields better visual outcomes compared to other networks. The proposed method effectively reduces the occurrence of false negatives by learning inter-image feature representations through global context modeling. However, there are still shortcomings regarding small sample data. For example, in the third pair of images in Fig. 6, which features a house in the upper right corner, our method can only detect a small portion of the area, while other methods can also identify some parts but still fail to detect the entire region. Furthermore, as seen in the visualizations in Fig. 8, our method requires improvement in the areas with blurred edges.

MF-VMamba excels in multiscale feature extraction while effectively fusing global context and local detail information. CNN-based models (such as SNUNet and DTCDSCN), although powerful in feature extraction, typically struggle with long-range dependencies and require deeper networks to achieve comparable performance. This can lead to issues such as vanishing gradients or overfitting in complex datasets. In contrast, transformer-based models (such as BIT and ChangeFormer) are better suited for capturing global context but may lack the finely tuned multiscale processing present in MF-VMamba. This multiscale adjustment is crucial for datasets with varying resolutions and change scales, such as CDD and LEVIR-CD. While MF-VMamba performs well on most datasets, it faces challenges on datasets such as DSIFN-CD, where imbalanced pixel distribution and smaller change regions require more precise detection.

D. Ablation Study

To verify the effectiveness of each module of MF-VMamba, we conducted four ablation experiments as shown in Table IV, and the visualization results corresponding to each module are also shown in comparison, as shown in Fig. 9. Our network consists of three main components, including the VMamba-based encoder, the FEM, and the MAD. The first row indicates that it contains only the VMamba-based encoder, and the decoder that performs simple upsampling, which we call base. The second row adds the FEM on top of the first row, which is used to perform detailed feature extraction. The third row indicates replacing the simple upsampling decoder with MAD on top of the first row. The fourth row indicates combining the encoder, FEM, and MAD. As shown in Table IV, the experimental results outperform the baseline for both adding

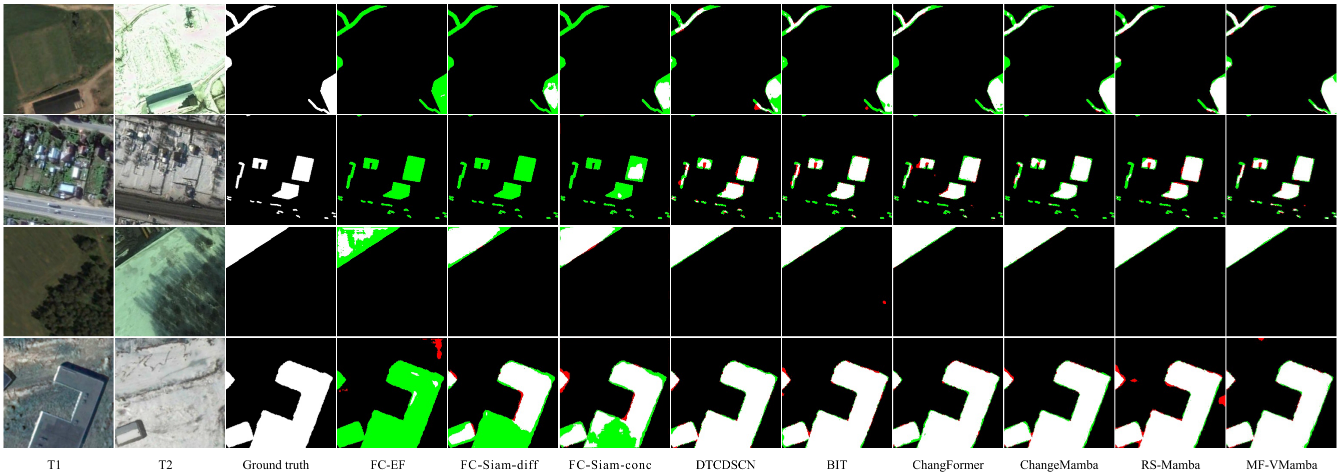


Fig. 5. Qualitative comparisons on the CDD dataset. From (left) to (right): T1 instance, T2 instance, ground truth, FC-EF [37], FC-Siam-diff [37], FC-Siam-conc [37], DTCDSCN [16], BIT [24], ChangeFormer [43], ChangeMamba [28], RS-Mamba [49], and MF-VMamba (ours). Color scheme: white represents TP (i.e., "changed"), black corresponds to TN (i.e., "unchanged"), red indicates FP, and green denotes FN.

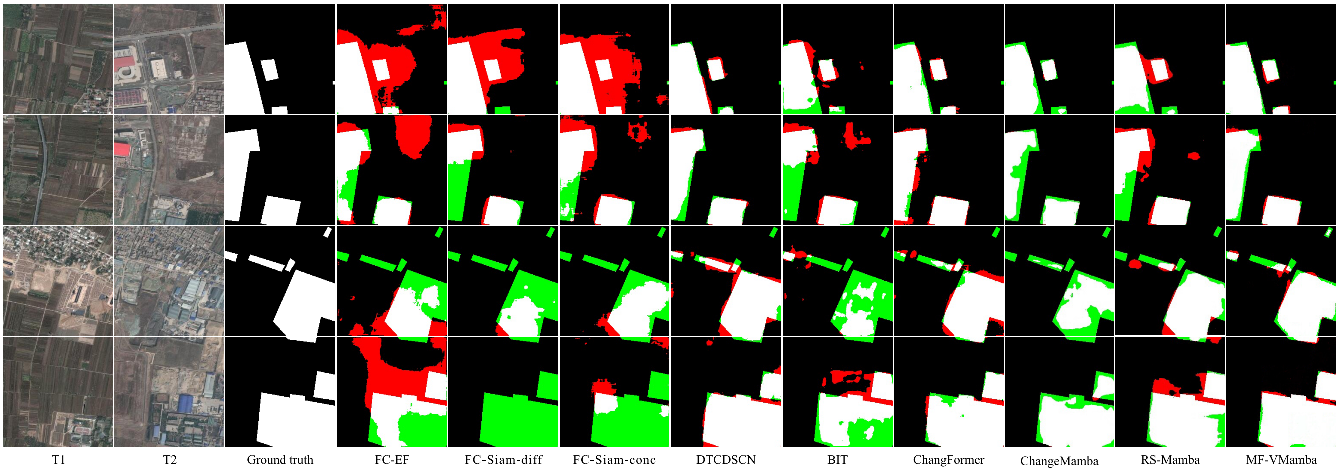


Fig. 6. Qualitative comparisons on the DSIFN-CD dataset. From (left) to (right): T1 instance, T2 instance, ground truth, FC-EF [37], FC-Siam-diff [37], FC-Siam-conc [37], DTCDSCN [16], BIT [24], ChangeFormer [43], ChangeMamba [28], RS-Mamba [49], and MF-VMamba (ours). Color scheme: white represents TP (i.e., "changed"), black corresponds to TN (i.e., "unchanged"), red indicates FP, and green denotes FN.

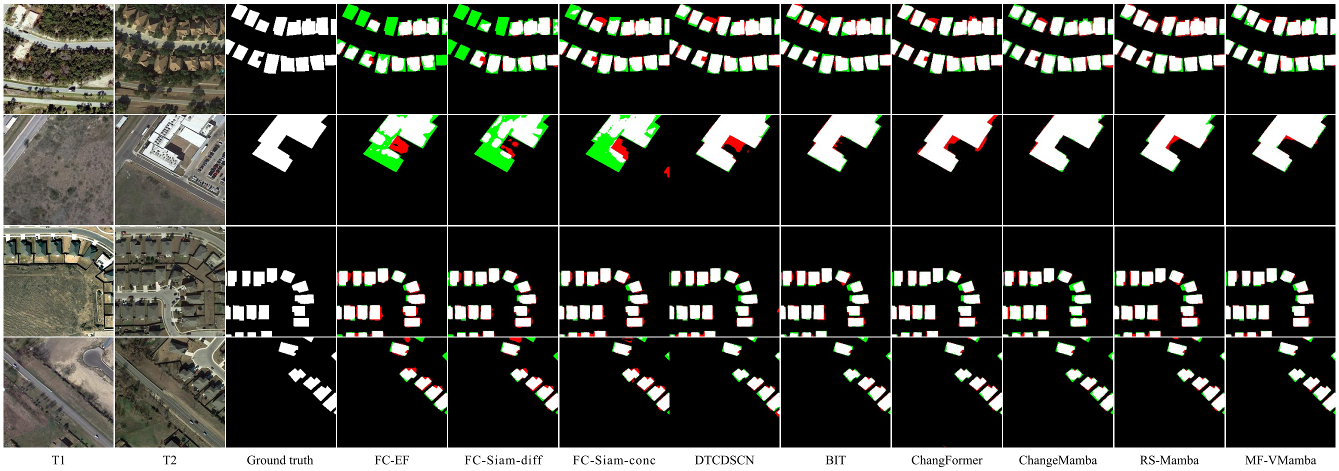


Fig. 7. Qualitative comparisons on the LEVIR-CD dataset. From (left) to (right): T1 instance, T2 instance, ground truth, FC-EF [37], FC-Siam-diff [37], FC-Siam-conc [37], DTCDSCN [16], BIT [24], ChangeFormer [43], ChangeMamba [28], RS-Mamba [49], and MF-VMamba (ours). Color scheme: white represents TP (i.e., "changed"), black corresponds to TN (i.e., "unchanged"), red indicates FP, and green denotes FN.

a module alone and in combination. For adding only FEM and MAD, respectively, both $F1$ and IoU are improved by 0.47%/0.73% and 0.85%/1.06%, respectively. When FEM and MAD are added in combination, $F1$ and IoU improved

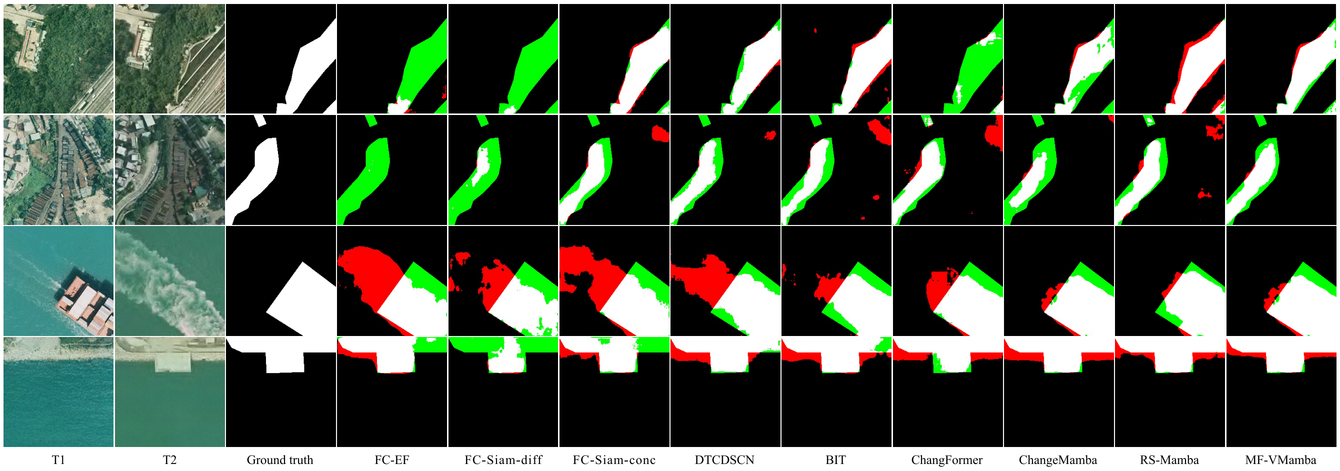


Fig. 8. Qualitative comparisons on the SYSU-CD dataset. From (left) to (right): T1 instance, T2 instance, ground truth, FC-EF [37], FC-Siam-diff [37], FC-Siam-conc [37], DTCDSCN [16], BIT [24], ChangeFormer [43], ChangeMamba [28], RS-Mamba [49], and MF-VMamba (ours). Color scheme: white represents TP (i.e., "changed"), black corresponds to TN (i.e., "unchanged"), red indicates FP, and green denotes FN.

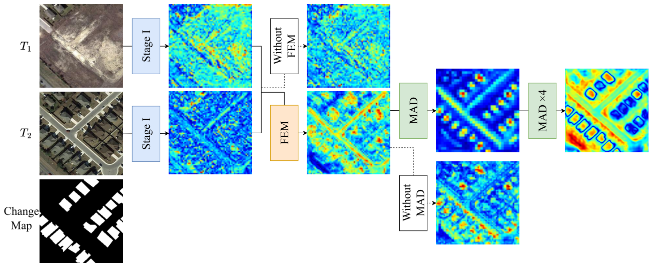


Fig. 9. Visualization of features in different modules. FEM denotes feature enhancement module. MAD denotes multilevel attention decoder. T_1 and T_2 denote a pair of example images selected on the LEVIR-CD dataset.

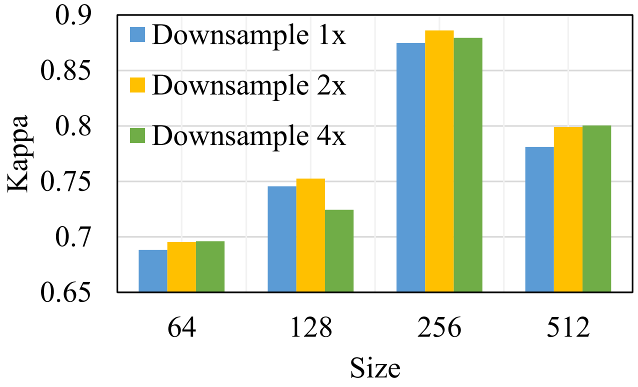


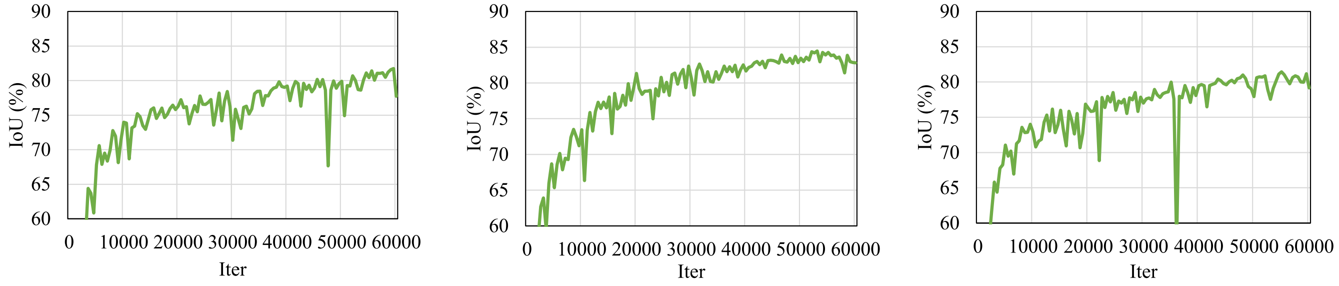
Fig. 10. Performance of the model on different image sizes and downsampling ratios.

by 13.76% and 20.47%, respectively. Meanwhile, from the perspective of feature visualization (see Fig. 9), when neither the FEM nor MAD modules are introduced, the extracted feature contours are relatively blurry and fail to accurately reflect the changes in the image. When the FEM is introduced alone, although the extracted features roughly show the contours of the change regions, they still lack sufficient detail. Furthermore, when only the FEM is used without the MAD module, the feature information is improved, but, compared to the combined use of the FEM and MAD modules, the feature localization accuracy is lower, and it includes more irrelevant information. When four MAD modules are introduced, the detected change regions are significantly enhanced: 1) the feature boundaries are clearer; 2) irrelevant information is effectively suppressed; and 3) the accuracy of CD is improved. This significant improvement proves that effective integration of global and local features facilitates the CD task. With the addition of the FEM, it is able to better capture local detail information. With the addition of the MAD module, it is able to better integrate global and local information, making clearer structures and edges.

E. Impact of Image Size and Spatial Resolution

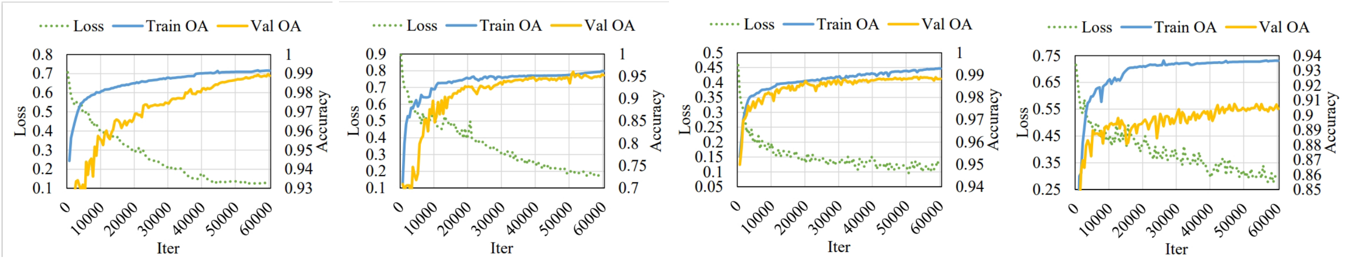
The image size determines the coverage and level of detail of the observed area, and larger image sizes can provide information about a wider area and help detect changes over a large area. However, excessively large image size may increase computational complexity and processing time. On the contrary, overly small image size results in limited coverage and may lose detail information and contextual information. Spatial resolution affects the detail clarity and accuracy of an image. High-resolution images can capture more subtle changes and are suitable for fine-grained CD, such as urban sprawl or vegetation changes. However, high-resolution images also bring higher data volumes and computational requirements that may make data processing more difficult. Low-resolution images, while reducing the burden of data processing, may not capture important details and small changes.

We conducted experiments on the LEVIR-CD dataset using the kappa coefficient as an evaluation metric. Due to the limitations of the experimental equipment, we tested combinations of image sizes (64, 128, 256, and 512) and downsampling ratios (1, 2, and 4). Fig. 10 shows the performance of the model on different image sizes and downsampling multiples, and it is found that the model performance is optimal when the image size is 256 × 256 and the downsampling ratio is 2. This may be due to the optimal balance of image resolution,



(a) (b) (c)

Fig. 11. Using different downsampling ratios on the LEVIR-CD dataset of the same size 256 × 256. (a) Downsample 1×. (b) Downsample 2×. (c) Downsample 3×.



(a) (b) (c) (d)

Fig. 12. Changes in network metrics and loss values for training and validation on (a) CDD, (b) DSIFN-CD, (c) LEVIR-CD, and (d) SYSU-CD datasets.

contextual information, and spatial feature capture in this configuration.

Specifically, with a downsampling ratio of 2, the image resolution is sufficient to capture small changes and details while avoiding the noise and computational burden associated with high resolution. The image size of 256 × 256 provides enough coverage and contextual information to help the model accurately understand the changes in the image. In addition, the moderate resolution and size make spatial features (e.g., edges, shapes, and textures) clearly visible, which improves the accuracy of CD. Also from Fig. 11, it can be seen that the IoU curve changes more gently during the training process when the downsampling magnification is 2 at the same size of 256.

On the contrary, a downsampling ratio of 1 increases the computational complexity and data processing burden, which may lead to noise and overfitting, although the image resolution is higher. A ratio of 4 is too low a resolution, resulting in insufficient detail and contextual information, blurred spatial features, and difficulty for the model to accurately recognize changes. In conclusion, the configuration with an image size of 256 × 256 and a downsampling ratios of 2 finds the optimal balance between resolution, contextual information, and spatial features, resulting in the highest model performance.

V. Discussion

A. Overall Comparison

Fig. 12 illustrates the changes in loss values and OA during the training and validation processes of our network across four datasets: CDD, DSIFN-CD, LEVIR-CD, and SYSU-CD. Overall, as the number of iterations increases, both training and validation OA show an upward trend, with the training accuracy slightly higher than the validation accuracy. The loss values decrease and tend to stabilize over time.

Specifically, in Fig. 12(a), the CDD dataset shows a rapid increase in validation OA, approaching 0.99 and stabilizing, indicating strong model performance on this dataset with minimal fluctuation in the validation curve. In Fig. 12(b), the model performs relatively worse on the DSIFN-CD dataset, particularly in the early stages, where both training and validation OA curves exhibit significant fluctuation. This may be attributed to the smaller size of the dataset, which leads to insufficient training. However, as iterations increase, the curves tend to stabilize, with the validation OA leveling off around 0.96, and the loss value also showing a downward trend toward equilibrium. Fig. 12(c) reveals that, on the LEVIR-CD dataset, the OA curve stabilizes and rises steadily after 40 000 iterations, with the validation OA settling at approximately 0.991. The loss value fluctuates around 0.12, indicating stable performance of the model on this dataset. In Fig. 12(d), for the SYSU-CD dataset, the larger and more diverse sample size results in slower improvement in validation OA, with significant fluctuations observed. Nevertheless, the overall trend is upward, with the validation OA eventually approaching 0.91, and the loss value gradually decreasing and stabilizing. In summary, the experimental results across these four datasets demonstrate that our proposed network exhibits good convergence behavior, with loss values and OA curves stabilizing in the later stages of training. This indicates that the model has strong generalization ability and robustness. These results validate the effectiveness of the network for remote sensing image CD tasks across a range of dataset types, particularly in handling more complex and diverse data.

B. Parameter Comparison

To further validate the computational efficiency of the model, Table V and Fig. 13 provide a comprehensive comparison of model parameters and computational costs for various methods applied to remote sensing image CD on the

TABLE V
COMPARISON OF MODEL PARAMETERS AND COMPUTATIONAL COSTS ON THE LEVIR-CD DATASET

Methos	Params	GFLOPs
FC-EF	1.35	14.31
FC-EF-conc	1.54	21.32
FC-EF-diff	1.35	18.91
DTCDSCN	31.25	52.89
SNUNet	12.03	21.33
ISNet	14.36	23.58
SARAS-Net	56.89	139.9
BIT	3.49	42.53
ChangeFormer	41.03	211.15
ChangeMamba	85.53	179.32
RS-Mamba	52.73	95.74
CDMamba	12.71	151.23
ours	57.84	94.06

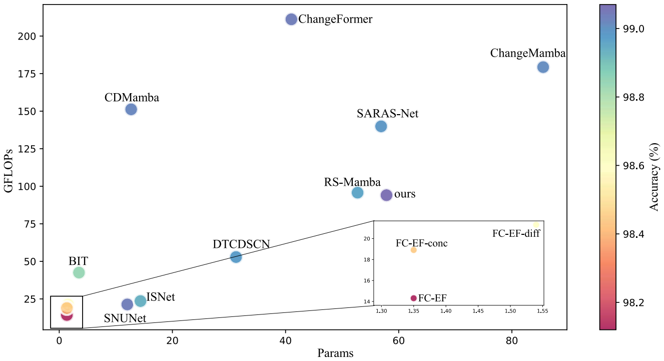


Fig. 13. Comparison of model parameters and computational costs on the LEVIR-CD dataset. The position of the bubbles is determined by the parameters and GFLOPs, and the depth of the color is determined by accuracy.

LEVIR-CD dataset. In Table IV, we observe a clear distinction in both the number of parameters (in millions) and the GFLOPs (giga floating-point operations per second) across different models. Notably, models such as FC-EF, FC-EF-conc, ISNet, and SNUNet are lightweight in terms of parameters and computational complexity, with the FC-EF family exhibiting the lowest parameter count and GFLOPs (1.35M parameters and 14.31 GFLOPs for FC-EF). Conversely, models such as ChangeFormer and ChangeMamba significantly increase both parameters (up to 85.53M for ChangeMamba) and GFLOPs (179.32 GFLOPs for ChangeMamba), likely due to their more sophisticated architectures aimed at capturing complex temporal and spatial changes.

Our model, highlighted with 57.84M parameters and 94.06 GFLOPs, strikes a balance between computational cost and accuracy, offering competitive performance without the prohibitive resource demands of the largest models. The bubble chart (see Fig. 13) reinforces these findings, illustrating the trade-offs between model complexity and efficiency, with our model positioned closer to RS-Mamba and SARAS-Net, which also achieve good accuracy at moderate computational costs. Meanwhile, methods such as ChangeFormer and ChangeMamba are computationally intensive but may justify their resource requirements with superior accuracy, as indicated by their higher color gradient on the chart. This comparison demonstrates that, while lightweight models are more efficient, they might sacrifice accuracy, whereas larger models, including ours, aim to optimize this trade-off, offering a viable solution for large-scale remote sensing CD tasks.

VI. CONCLUSION

In this article, we propose a multilevel feature extraction network based on VMamba for remote sensing image CD. The core idea of MF-VMamba is to establish long-range dependencies by fusing global and local information of branches at each scale to achieve information interaction in space and channel dimensions, thereby improving the accuracy of CD. The advantages of MF-VMamba can be attributed to several key factors.

Architectural Innovation: By employing VMamba, we effectively reduce computational complexity and memory usage, which is essential for processing high-resolution remote sensing images. The incorporation of LKC further allows for efficient extraction of spatial and channel information, leading to a decrease in model parameters without compromising performance.

Feature Fusion Mechanism: The VMamba-based encoder is specifically designed to extract multiscale global features, while the decoder facilitates the fusion of these features for enhanced detection of change objects in images. This dual approach is critical for achieving high performance across varied datasets.

Implications for Future Research: The improvements observed in our model not only signify advancements within CD methodologies but also serve as a foundation for future explorations into the Mamba architecture. We encourage researchers to build upon our findings, particularly in the areas of multiscale feature fusion and the integration of attention mechanisms, to address challenges associated with complex remote sensing imagery.

Through extensive experiments conducted on four public datasets, our model demonstrates superior performance compared to state-of-the-art CD methods. These results affirm the effectiveness of MF-VMamba in tackling the inherent complexities of remote sensing data.

REFERENCES

[1] M. Rußwurm and M. Körner, “Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1496–1504.

[2] Q. Zhu et al., “Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery,” *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 63–78, Feb. 2022.

[3] X. Wang, X. Fan, Q. Xu, and P. Du, “Change detection-based co-seismic landslide mapping through extended morphological profiles and ensemble strategy,” *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 225–239, May 2022.

[4] X. Wang et al., “Long-term landslide evolution and restoration after the wenchuan earthquake revealed by time-series remote sensing images,” *Geophys. Res. Lett.*, vol. 51, no. 2, Jan. 2024, Art. no. e2023GL106422.

[5] Y. Chen, X. He, J. Wang, and R. Xiao, “The influence of polarimetric parameters and an object-based approach on land cover classification in coastal wetlands,” *Remote Sens.*, vol. 6, no. 12, pp. 12575–12592, Dec. 2014.

[6] D. Wen et al., "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 68–101, Dec. 2021.

[7] F. Deng, W. Luo, Y. Ni, X. Wang, Y. Wang, and G. Zhang, "UMiT-Net: A U-shaped mix-transformer network for extracting precise roads using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5801513.

[8] Z. Lv, M. Zhang, W. Sun, J. A. Benediktsson, T. Lei, and N. Falco, "Spatial-contextual information utilization framework for land cover change detection with hyperspectral remote sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4411911.

[9] Z. Lv, Z. Lei, L. Xie, N. Falco, C. Shi, and Z. You, "Novel distribution distance based on inconsistent adaptive region for change detection using hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4404912.

[10] X. Wu, P. Gamba, J. Feng, R. Shang, X. Zhang, and L. Jiao, "A multitask framework for hyperspectral change detection and band reweighting with unbalanced contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5530213.

[11] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[12] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[13] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

[14] 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.

[15] J. Chen et al., "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2020.

[16] 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.

[17] 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, pp. 1–5, 2021.

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

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[20] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Cision Cattern Recognit.*, Jun. 2021, pp. 6881–6890.

[21] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 323–339.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[23] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[24] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[25] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2021, *arXiv:2111.00396*.

[26] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

[27] Y. Liu et al., "VMamba: Visual state space model," 2024, *arXiv:2401.10166*.

[28] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, "ChangeMamba: Remote sensing change detection with spatiotemporal state space model," 2024, *arXiv:2404.03425*.

[29] K. Chen, B. Chen, C. Liu, W. Li, Z. Zou, and Z. Shi, "RSMamba: Remote sensing image classification with state space model," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[30] J.-F. Mas, "Monitoring land-cover changes: A comparison of change detection techniques," *Int. J. Remote Sens.*, vol. 20, no. 1, pp. 139–152, Jan. 1999.

[31] Y. Ban and O. Yousif, "Change detection techniques: A review," in *Multitemporal Remote Sensing: Methods and Applications*. Cham, Switzerland: Springer, 2016, pp. 19–43.

[32] D. Lu, P. Mausel, E. Brondízio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.

[33] J. Chen, P. Gong, C. He, R. Pu, and P. Shi, "Land-use/land-cover change detection using improved change-vector analysis," *Photogramm. Eng. Remote Sens.*, vol. 69, no. 4, pp. 369–379, 2003.

[34] A. Lefebvre, T. Corpetti, and L. Hubert-Moy, "Object-oriented approach and texture analysis for change detection in very high resolution images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jun. 2008, pp. IV-663–IV-666.

[35] B. Desclée, P. Bogaert, and P. Defourny, "Forest change detection by statistical object-based method," *Remote Sens. Environ.*, vol. 102, nos. 1–2, pp. 1–11, May 2006.

[36] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

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

[38] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 214–217.

[39] 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. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[40] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 802–810.

[41] G. Cheng, G. Wang, and J. Han, "ISNet: Towards improving separability for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623811.

[42] X. Hou, Y. Bai, Y. Li, C. Shang, and Q. Shen, "High-resolution triplet network with dynamic multiscale feature for change detection on satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 103–115, Jul. 2021.

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

[44] 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.

[45] B. Wu et al., "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*.

[46] J. Yao, D. Hong, C. Li, and J. Chanussot, "SpectralMamba: Efficient mamba for hyperspectral image classification," 2024, *arXiv:2404.08489*.

[47] G. Wang, X. Zhang, Z. Peng, T. Zhang, and L. Jiao, "S^{2}Mamba: A spatial–spectral state space model for hyperspectral image classification," 2024, *arXiv:2404.18213*.

[48] Q. Zhu et al., "Samba: Semantic segmentation of remotely sensed images with state space model," 2024, *arXiv:2404.01705*.

[49] X. Ma, X. Zhang, and M.-O. Pun, "RS3Mamba: Visual state space model for remote sensing image semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[50] M. Liu, J. Dan, Z. Lu, Y. Yu, Y. Li, and X. Li, "CM-UNet: Hybrid CNN-Mamba UNet for remote sensing image semantic segmentation," 2024, *arXiv:2405.10530*.

[51] Y. Liu, J. Xiao, Y. Guo, P. Jiang, H. Yang, and F. Wang, "HSIDMamba: Exploring bidirectional state-space models for hyperspectral denoising," 2024, *arXiv:2404.09697*.

[52] G. Fu, F. Xiong, J. Lu, and J. Zhou, "SSUMamba: Spatial–spectral selective state space model for hyperspectral image denoising," 2024, *arXiv:2405.01726*.

[53] H. Zhang, K. Chen, C. Liu, H. Chen, Z. Zou, and Z. Shi, "CDMamba: Remote sensing image change detection with Mamba," 2024, *arXiv:2406.04207*.

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

[55] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-V4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 4278–4284.

[56] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[57] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[58] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11963–11975.

[59] S. Liu et al., "More ConvNets in the 2020s: Scaling up kernels beyond 51×51 using sparsity," 2022, *arXiv:2207.03620*.

[60] G. Lu, W. Zhang, and Z. Wang, "Optimizing depthwise separable convolution operations on GPUs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 1, pp. 70–87, Jan. 2022.

[61] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 565–571, May 2018.

[62] 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, Aug. 2020.

[63] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021.

[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 Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2018.

Zhongyu Zhang received the B.S. degree in geographic information science from Hebei GEO University, Hebei, China, in 2022. She is currently pursuing the Ph.D. degree with the State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu, Sichuan, China.

She is primarily engaging in theoretical research in machine learning and research on change detection in optical remote sensing images.

Xuanmei Fan received the Ph.D. degree in engineering geology from the Faculty of Geo-Information Science and Earth Observations (ITC), University of Twente, Enschede, The Netherlands, in 2013.

She is the Director of the State Key Laboratory of Geohazard Prevention and Geoenvironment Protection (SKLGP), Chengdu University of Technology, Chengdu, China. Following her studies, she was appointed as a Disaster Risk Reduction Specialist by the United Nations Institute for Training and Research (UNITAR). In 2015, she became a Leading Professor of geohazard risk assessment and prevention at SKLGP. She has published more than 140 ISI papers in *Nature Geoscience*, *Reviews of Geophysics*, *GRL*, and *JGR*. Her research interests include earthquake and climate change-induced chains of geological hazards.

Dr. Fan received the First Prize for National Science and Technology Achievements in 2014, the Richard Wolters Prize by the International Association for Engineering Geology and the Environment (IAEG) in 2016, the Top Ten Female Geologists in China, in 2017, the Outstanding Young Scholar Grant NSFC in 2021, the First Prize of Sichuan Natural Science Award in 2021, the Scientific Exploration Award in 2022, and the Chinese Young Female Scientist Award in 2023. She has served as the Principal Investigator for many national and international projects, including a U.K.–China collaboration project, European projects, and several National Science Foundation of China (NSFC) projects.

Xin Wang received the joint Ph.D. degree from Queen's University, Kingston, ON, Canada, in 2020, and the Ph.D. degree in geography from Nanjing University, Nanjing, China, in 2021.

He is currently an Associate Professor with the State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu, China. He has published over 40 peer-reviewed papers including GRL, ISPRS P&RS, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), and has reviewed for more than 20 journals such as *Nature Communications*, ISPRS P&RS, and IEEE TGRS. He is broadly interested in remote sensing techniques and applications, with particular emphasis on land cover change and time series analysis, and currently focuses on geological hazard mapping and susceptibility modeling driven by earthquakes and climate change.

Yingxiang Qin received the B.S. degree in geographic information science from Hebei GEO University, Hebei, China, in 2022. He is currently pursuing the M.S. degree with the College of Geography and Planning, Chengdu University of Technology, Chengdu, Sichuan, China.

His research interests include urban air pollution and GIS development.

Junshi Xia (Senior Member, IEEE) received the B.S. degree in geographic information systems and the Ph.D. degree in photogrammetry and remote sensing from China University of Mining and Technology, Xuzhou, China, in 2008 and 2013, respectively, and the Ph.D. degree in image processing from Grenoble Images Speech Signals and Automatics Laboratory, Grenoble Institute of Technology, Grenoble, France, in 2014.

From 2014 to 2015, he was a Visiting Scientist at the Department of Geographic Information Sciences, Nanjing University, Nanjing, China. From 2015 to 2016, he was a Post-Doctoral Research Fellow with the University of Bordeaux, Bordeaux, France. From 2016 to 2018, he was the Japan Society for the Promotion of Science (JSPS) Post-Doctoral Overseas Research Fellow at The University of Tokyo, Kashiwa, Japan. Since 2018, he has been with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan, where he is currently a Senior Research Scientist. His research interests include multiple classifier systems in remote sensing, hyperspectral remote sensing image processing, and deep learning in remote sensing applications.

Dr. Xia was a recipient of the First Place Prize in the IEEE Geoscience and Remote Sensing Society Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee in 2017. Since 2019, he has been an Associate Editor of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL), *Remote Sensing*, and *Frontiers in Remote Sensing*, and a Guest Editor of *Remote Sensing* and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS).